Genetic Epidemiology 29: 234–292 (2005)

ABSTRACTS FROM THE

FOURTEENTH ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

Park City, Utah
October 23–24, 2005

1

Polymorphism 5 – of the Leptin Gene Results in Sex-Specific Trends in Birth Weight and Cord Leptin Levels

R.M. Adkins(1), C.K. Klauser(2), J. Fain(3), E.F. Magann(4), J.C. Morrison(2)
(1) Dept. Ped., Univ. TN HSC, USA, (2) Obstet. & Gynecol., Univ. MS Medical Center, USA, (3) Dept. Mol. Sci., Univ. TN HSC, USA, (4) Obstet. & Gynecol., Naval Med. Cent. Portsmouth, USA

Leptin levels are positively correlated with adult BMI and birth weight, with female serum levels nearly twice those of males. Postnatally, SNPs in the leptin gene have been shown to exhibit sex-specific associations with BMI and circulating leptin levels. The purpose of this study is to test the association between variation in the leptin gene and both birth weight and cord leptin levels. 366 Caucasian and African-American subjects were recruited at three different sites in both case-control and cross-sectional designs. Genotype at 5 SNPs in the leptin gene was determined for all newborns. Using either birth weight Z score or log(leptin) as dependent variables, ANOVA was performed using a model composed of potential correlates of birth weight or leptin levels (gestational age, gender, size for gestational age, SNP genotype) with an interaction term included to detect a differential association between the dependent variable and genotype, depending on newborn gender. Leptin levels were about twice as high in females as in males (14.8 vs. 7.0 ng/ml; $p<0.001$), and leptin levels positively correlated with birth weight ($r2=0.33$; $p<0.001$). Relative to homozygotes for the major allele at a SNP upstream of the leptin gene, heterozygous males exhibited an increase in birth weight and cord leptin levels, while females exhibited a decrease in both variables ($F=4.7$, $p=0.01$ for gender x genotype interaction).

2

Robust estimation and testing of haplotype effects in case-control studies

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

Case-control genetic association studies are popular mechanisms for examining the genetic influences of complex disease. Of particular interest is estimating and testing the effect of haplotypes, as such analyses can provide critical information regarding the function of a gene and may be more powerful than single loci methods. When haplotype phase is uncertain, likelihood methods are forced to model the nuisance distribution of haplotypes and can, if this distribution is misspecified, lead to substantial bias in parameter estimates even when complete genotype information is available. We use a geometric approach to estimation in the presence of nuisance parameters, and develop locally-efficient estimators of the effect of haplotypes on disease that are robust to incorrect estimates of haplotype frequencies. The methods are demonstrated with a simulation study of a case-control design.

3

Linkage analysis of DNA repair genes in high-risk Utah breast cancer families

K. Allen-Brady & N.J. Camp
Genetic Epidemiology, Dept of Medical Informatics, University of Utah, Salt Lake City, UT

The protein products of ATM, MRE11, NBS1, RAD50, and XRCC4 genes play an integral role in the repair of DNA double-strand breaks and thus represent good candidate genes for breast cancer susceptibility. We have performed a linkage analysis using haplotype tagging-SNPs (tSNPs) identified in these five DNA repair genes in 139 high-risk, Utah non-BRCA1/2 breast cancer pedigrees. Given that tSNPs are chosen to eliminate redundancy, the linkage disequilibrium between them is minimal. The number of tSNPs selected for each gene were: ATM=1, MRE11=4, NBS1=2, RAD50=3, and XRCC4=4. Three inheritance models (dominant, recessive, and codominant) for affected-only subjects (n=478) were used for this analysis and based on a segregation model by Cui et al. that also excluded BRCA1/2 mutation carriers [Am J Hum Genet, 2001. 68: 420–31]. Using the robust multipoint linkage statistical program MCLINK, a program that uses a Markov chain Monte Carlo approach to reconstruct haplotypes across markers, multipoint theta LOD (TLOD) scores were computed. Correcting for multiple tests, a TLOD threshold of 1.6 can be considered significant (5 genes $\times$ 3 models=15 tests). Only the analysis of NBS1 reached this level of significance under both a codominant and recessive model (TLOD=1.75, p=0.002 and TLOD=1.71, p=0.0025, respectively). The next step will be to identify linked pedigrees and breast cancer cases carrying segregating haplotypes to establish whether NBS1 explains the breast cancer occurrence seen in those families.

Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20103

4

Genome wide linkage scan of 781 affected sibling pair families with rheumatoid arthritis using the Illumina Linkage IV set of ~5,600 SNP markers

C. Amos, W. Chen, AT Lee, MF Seldin, LA Criswell, DA Kastner, EF Remmers, W Li, M Kern, and PK Gregersen

The North American Rheumatoid Arthritis Consortium (NARAC) has assembled a large collection of affected sibling pairs with rheumatoid arthritis (www.naracdata.org). In addition to a strong signal in the HLA region, previous microsatellite genome screens on 512 of these families have revealed modest evidence linkage (LOD 1.5–2.5) at several other chromosomal locations. We have now performed a complete genome wide linkage scan using ~5,600 informative SNP markers on 758 NARAC families containing 931 affected sibling pairs. Linkage analysis was carried using a call to MERLIN from SNPLINKer. Allowance for linkage disequilibrium between SNP markers (D–<0.7) required the removal of approximately 17% of the genotypes, and eliminated an apparently spurious linkage peak at the extreme telomeric portion of chromosome 6q. The average LOD score decreased by 0.15 after dropping markers in LD. The data confirm a very strong and broad linkage peak on chromosome 6p, centered on the HLA region (max NPL score=18.9 at 32.9 cM), but extending from 13 cM on 6p out to 80cM on 6q with LOD scores >3. This 67 cM linkage interval suggests the likely involvement of multiple loci on this chromosome in RA susceptibility. In addition, previously identified regions with very modest evidence linkage on chromosomes 4, 5 and 11 now show more compelling evidence of linkage (LOD scores >2.8, >2.8 and >3.5 respectively). Finally, we have been able to narrow a previously observed linkage peak on chromosome 18q to a region around 52 cM.

5

***XRCC1* haplotypes are associated with smoking-related CHD: the Atherosclerosis Risk in Communities (ARIC) study**

CL Avery(1), D Zeng(2), DY Lin(2), G Heiss(1), D Couper(2), AF Olshan(1), MS Bray(3), KE North(1)
Depts. of (1) Epid and, (2) Biostat, Univ. of North Carolina at Chapel Hill, USA, (3) Dept. of Pediatrics, Baylor College of Medicine, USA

While evidence of increased risk for coronary heart disease (CHD) associated with smoking is well established, the mechanisms that link smoking to CHD are poorly understood. Cigarette smoke metabolism is a multi-step process and variations in toxicological response have been attributed to heritable DNA repair differences. We previously examined eight single SNP associations in the *XRCC1* DNA repair gene and wanted to further evaluate the gene. Thus, we calculated maximum likelihood estimates of hazard ratios (HR) for eight *XRCC1* haplotypes from five haplotype tagging SNPs that were measured by a case-cohort design.
All incident CHD cases (n=1000) and a cohort representative sample (n=980) were selected from the biracial ARIC study. Analyses were stratified by race and adjusted for age, sex, and study center, assuming an additive model. Three of the five tagging SNPs were intronic, one was located in the UTR, and the other was a non-synonymous (trp194arg) variant. Interaction (P=0.1) was observed in the Caucasian sample between ever-smoking and a particular haplotype (7% frequency). The HR for the joint effect was 1.30 (95% CI: 1.03, 1.64), compared to the main effects of ever-smoking (HR=1.72, 95% CI: 1.47, 2.01) or the haplotype (HR=1.13, 95% CI: 0.70, 1.82). Among ever-smokers, the HR for the haplotype was 0.76 (95% CI: 0.58, 0.98). Our results suggest that *XRCC1* may be a good candidate for further study.

6

Incorporating environmental components into heritability analyses of PTSD symptoms, depression and anxiety in a family study of survivors from an Armenian earthquake

J.N. Bailey, D.P. Walling, E.P. Nobel, T. Ritchie, H.A. Goenjian, A.K. Goenjian

Post Traumatic Stress Disorder (PTSD) has both a genetic and environmental component. One of the challenges in genetic epidemiology studies of PTSD is that the disorder by definition requires an environmental exposure to trauma which rarely happens to family members concurrently. Natural disasters offer opportunity to ascertain subjects for PTSD genetic epidemiology studies. A natural trauma occurred in 1988 when the Spitak Earthquake struck Armenia. Gumri was one of the most devastated cities; nearly all the inhabitants experienced direct life-threat. Destruction was uniform throughout the city. This study examined 202 subjects from 13 families randomly ascertained from Gumri, Armenia; survivors of the Spitak earthquake. All family members were assessed for quantitative measures of PTSD, depression and anxiety, using the UCLA Posttraumatic Stress Disorder Reaction Index, (PTSD-RI), Becks Depression and Anxiety Inventories. An earthquake exposure scale was also administered, to assess environmental variables that may influence the traits. Heritability analyses were performed using variance component quantitative genetic methods (in SOLAR), which incorporates and tests for significance of environmental covariates. Some variables (sex, age, whether an individual had seen death) affected scores of all three traits. Other variables were more specific for the trait they influenced. Adjusting for covariates decreased the heritability estimates for all three traits, but they remained significant indicating strong genetic components. These families offer a unique opportunity to study the genetic epidemiology of not just PTSD, but also anxiety and depression.

7

Sequence-level population genomic simulations

Balding DJ, Clark T, Hoggart CJ
Centre for Biostatistics, Imperial College London, UK

The field of genetic epidemiology is sometimes marred by inadequate simulations used to test new methodology. Often population genetic features that are crucial to the phenomenon under study are not incorporated into the simulation method. The recent popularity of coalescent-based simulation software, such as Hudson's MS, has improved the situation. Coalescent methods work backwards in time, which is computationally efficient but is limited in the amount of recombination that can be incorporated, as well as the flexibility to include important features such as gene conversion or selection. With increased capacity of computers, it is now feasible to implement more flexible, forwards-in-time simulation strategies. We have developed software for forwards-in-time, whole-population simulation of DNA sequences over large genomic intervals that can incorporate different demographic models, as well as recombination, both cross-overs and gene conversions, at highly variable rates. As illustrations of its potential uses, we (1) examine the performance of algorithms for identifying recombination hotspots, and (2) find approximations for genome-wide significance levels of genetic association studies using high marker density or resequence data, under different assumptions about demography and gene conversion.

8

A method to detect regions of strong differential selection between two distinct populations

M.J. Barber, A.C. Lam, H.J. Cordell, J.A. Todd
JDRF/WT Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, UK

Evidence of selection could increase the prior probability that a candidate region harbors genes for susceptibility to common immune mediated diseases. Using population-specific allele frequencies from genome-wide SNP (single nucleotide polymorphism) data from the International HapMap Project, we investigate the analytical problem of detecting regions with evidence of strong differential selection between two distinct populations, where differential selection has occurred at some point since their common ancestry. Within a test region containing a number of SNPs, the goal of our method is to meta-analyze the single SNP p_excess (or delta) metrics. A single SNP p_excess metric estimates the haplotype frequency shift associated with the single SNP, and has the desirable property of being independent of the overall SNP allele frequency. The advantage of such a regional meta-analysis method is to enable the detection of regions with a highly unusual pattern of consistent and extreme divergence, which could be due to rapid positive selection of a single (unknown) haplotype, in contrast to the relatively commonplace extreme divergence at a single SNP. Practical application has been made to data available from the International HapMap Project specifically comparing the CEPH (Utah residents with ancestry from northern and western Europe) samples with the combined samples of the Japanese (in Tokyo, Japan) and of the Han Chinese (in Beijing, China), which has highlighted several potential regions.

9

Identification of disease susceptibility loci using a co-evolution measure and haplotype phylogenies

C Bardel, P Darlu and E Génin
Inserm U535, Villejuif, France

The localization of susceptibility loci involved in the determinism of complex diseases is a central question in genetic epidemiology. Templeton suggests that the reconstruction of the evolutionary history of case and control haplotypes through phylogenetic trees could be a solution to this problem. He developed a cladistic method consisting in looking for mutations defining groups containing significantly more haplotypes carried by case individual than controls. Applications on real data suggest that the method is promising but its efficiency has never been assessed through simulations. In this work, we propose a new phylogeny-based method to identify susceptibility loci. It consists in building a phylogeny of the different haplotypes, in defining a new character "S" corresponding to the disease status for each haplotype and in looking for the site whose evolution is the most correlated to the character "S". We have performed extensive simulations to determine the efficiency of this method for different genetic models. We have found that it is particularly efficient when the disease susceptibility allele is relatively frequent (around 0.4). We have also applied our method to a data set concerning the DRD2 gene and schizophrenia. We were able to successfully identify one variant that have previously been shown to be involved in the disease susceptibility.

10

Open-Source Multifactor Dimensionality Reduction (MDR) Software for Detecting and Interpreting Gene-Gene Interactions

N. Barney(1), W. Holden(1), B.C. White(1), L.W. Hahn(2), W. Bush(2), M.D. Ritchie(2), J.H. Moore(1)
(1) Dept. of Genetics, Dartmouth Medical School, USA, (2) Dept. of Molecular Physiology and Biophysics, Vanderbilt University, USA

Multifactor dimensionality reduction (MDR) is a data mining approach to detecting gene-gene interactions. MDR has been successfully applied to detecting interactions in studies of atrial fibrillation, bladder cancer, breast cancer, hypertension, myocardial infarction, prostate cancer, schizophrenia, and diabetes, for example. We describe here a comprehensive MDR software package with a user-friendly GUI that is programmed in Java and provided for free as open-source. MDR is distributed in three modules for 1) analysis, 2) permutation testing, and 3) data processing. The MDR analysis module carries out a combinatorial search for the best interaction model as determined by n-fold cross-validation. Output includes statistics such as accuracy, sensitivity, specificity, odds ratios, IF-THEN rules, and publication-quality graphical output of the multilocus models. The fitness landscape of all models evaluated is provided in the form of line graphs and histograms that can be explored interactively. The

software can be run from the command line for scripting and will automatically run in parallel on a multiprocessor computer with threading. The permutation testing module provides empirical p-values while the data processing module facilitates imputation of missing genotypes and resampling. Tools for genome-wide analysis will be included in future versions. MDR can be downloaded from www.epistasis.org/mdr.html.

11

Coverage and Efficiency of Whole Genome Association Studies

J.C. Barrett(1), I. Pe'er(2), M.J. Daly(2), L.R. Cardon(1)
(1) Wellcome Trust Centre for Human Genetics, Oxford, UK, (2) Broad Institute of MIT and Harvard, Cambridge, MA, USA

Falling SNP genotype costs and the recent availability of genome-wide surveys of genetic variation have made whole genome association (WGA) studies imminent. Despite such advances these studies are still expensive and present significant analytical hurdles. It is important, therefore, to consider what the goal of a WGA study is, how well newly developed SNP panels address that goal and how downstream analysis choices influence the interplay of those two questions.
One crucial choice is deciding upon a set of ''target variants'' that we wish to test for association to disease. Alternative strategies have proposed focusing on coding variants, SNPs in genes or conserved regions, or all known variants. Regardless of approach we must also consider which variants are within the accessible allele frequency spectrum. Further complicating these issues is the relationship between the set of genotyped SNPs and subsequent association tests and analysis.
Recent data from the HapMap project and Perlegen have made it possible to examine how well current and soon-to-be-available SNP genotyping products can capture ''target sets'', and to explore the consequences of initial marker selection on subsequent analyses. Using the available data, we compare different gene-based SNP panels with LD-derived and randomly selected panels of 100,000, 250,000 and 500,000 SNPs. We show how the initial genotyping strategy strongly influences later analysis choices and the eventual chances for identifying specific trait loci. Our results emphasize the importance of considering the entire study from SNP selection to final analysis in order to maximize efficiency.

12

Defining the Relation Between a Categorical Trait and a Quantitative Endophenotype at a Linked Locus

Christopher W. Bartlett and Veronica J. Vieland
Center for Statistical Genetics Research, University of Iowa, USA

Linkage studies of categorically defined traits (CT) often include consideration of quantitative endophenotypes (QT). We describe a novel way to assess the relationship between the CT and the QT at a linked locus, using the posterior probability of linkage (PPL) framework. We have previously developed a CT PPL and a QT-PPL, in which unknown trait parameters are integrated out of the likelihood. Here we add a mixed CT/QT PPL (C/Q-PPL), which uses the CT for ''affected'' individuals and QT values for all other persons, and is based on an underlying threshold model. Note that all three forms of the PPL are on the probability scale (0.1) and use a 2% prior probability of linkage. We simulated families with one CT and one QT assuming (1) only the CT is linked, (2) only the QT is linked, (3) the QT is linked and the CT is defined via a threshold on the QT scale. Under (1), the PPL outperforms the QT-PPL, which gives evidence against linkage ($<2\%$) and the C/Q-PPL yields intermediate values (mean PPL=66%, QT-PPL=1.4%, C/Q-PPL=27%). Under (2), the QT-PPL is the highest and again the C/Q-PPL is intermediate, with the PPL giving evidence against linkage (mean PPL=1.5%, QT-PPL=95%, C/Q-PPL =49%). Finally, under (3), all three methods show evidence for linkage, in descending order (QT-PPL=90%, C/Q-PPL=89%, PPL=22%). This pattern of results can therefore be used to establish whether the CT and QT are genetically related at the locus in question.

13

A comparison of methods for evaluating the relative effects of neighbouring polymorphisms associated with a disease

J.M. Biernacka, H.J. Cordell
Department of Medical Genetics, University of Cambridge, UK

Several methods have been proposed to assist in distinguishing potentially causal polymorphisms from those that show association due to LD. Given a postulated effect at a primary variant, we may ask if any other variants in the region appear to further contribute to the trait, indicating that the additional variant is causal or is in LD with another causal locus. Various methods of approaching this problem using case-parent trio data have been proposed, among those the stepwise conditional logistic method described by Cordell and Clayton (2002), and a permutation-based method proposed by Spijker et al. (2005). We compared these methods and other related approaches by simulation. Under the conditions investigated, no advantage was found using the permutation procedure over asymptotic conditional logistic regression. Because the procedure described by Spijker et al. (2005) and unconditional logistic regression rely on AFBAC (affected family-based) controls, they are prone to bias and therefore increased type 1 errors when haplotypes cannot be inferred for all families, as illustrated in our simulations. This is a greater problem when SNPs rather than microsarellite markers are analyzed. We propose an alternative to the permutation method of Spijker et al. (2005), which does not rely on haplotyping, and thus avoids the bias inherent in the procedures that use AFBAC controls. Our results show that this alternative procedure leads to good type 1 errors and power.

14

Methods for identification of polymorphisms responsible for a linkage signal

J.M. Biernacka, H.J. Cordell
Department of Medical Genetics, University of Cambridge, UK

Recently, several methods have been proposed to aid in the identification of disease associated polymorphisms that may explain an observed linkage signal. Li et al. (2005) proposed an approach to quantify the degree of LD between a candidate SNP and a putative disease locus using affected sib pair (ASP) genotype data, and developed tests to characterize the relationship between the candidate SNP and the disease locus. In the Li approach, no use is made of parental data in addition to ASP genotypes. We propose a modified approach in which parental data at the candidate SNP and at a series of linked markers is also used, which has the advantage that allele frequency estimation is not required, and that identity by descent information is increased. We apply our method to type 1 diabetes data for ASPs and their parents at candidate SNP and microsatellite marker loci near the INS region. Similarly to Li et al. (2005), we consider the case of a SNP closely linked to a causal locus, with no other causal loci linked to either of these, and ask whether the observed linkage signal can be explained by association with the given SNP of interest. We estimate 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 causal variant in the region. We study the properties of our method by simulation and compare it to the method of Li et al. (2005). Furthermore, we extend our method to test for LD between a disease locus and haplotypes composed of two tightly linked candidate SNPs.

15

Bringing into shape complex phenotypes: Methods to exploit intertwining between monogenic and complex genetic contributions for a single phenotype

S. Boehringer, T. Vollmar, C. Tasse, R.P. Wuertz, D. Wieczorek, B. Horsthemke

Complex traits are thought to have complex genetic contributions. Traits themselves, however, can be markedly simple, e.g., dichotomous or univariate. Many traits might be considered in a multivariate setting, e.g., analyzing hypertension jointly with metabolic disorders. In the wake of multivariate analyses like these we consider a roughly 4000 dimensional phenotype representing the face. Investigations are under way to conduct association and linkage analyses with respect to the phenotype to answer fundamental questions about the developement of the face. We here focus on the phenotype itself and present the steps needed to transform the data set into a form adequate for further analysis aiming at reduction of dimensionality and noise in the data set. We present a framework of three components and demonstrate its success on a population-based sample of 570 individuals: 1. Use of a different data set of 120 faces related to monogenic disorders to characterize variation that represents stable characteristics in human faces rather than noise. 2. Model selection procedures in conjunction with other dimensionality reduction methods like principal component analysis. 3. Heritability analysis for individual dimensions. This procedure yields a low (20 to 40) dimensional representation that captures relevant information of the data set. The parameters are the amount of variation captured, classification rates, and heritabilities. They allow further analysis in standard association or linkage settings. The knowledge gleaned from monogenic disorders substantially improved the characterization of the population-based data set – a result that we expect to hold for other phenotypes as well.

16

Robustness of the conditional logistic regression to the biallelic modeling

Bourgey M., Clerget-Darpoux F.
INSERM U535, Villejuif, France

Most human diseases involved numerous factors. For detecting them, testing association between the disease and intragenic markers is a widely used approach. It often happens that candidate genes are clustered and that it exist LD between their alleles. In that case, association may be observed with several genes. Methods have been developed to detect if one or several genes are involved in the disease susceptibility while taking into account the LD. In such methods, assumptions are often made on the gene effect modeling. In particular the gene effect is very often considered and coded as biallelic. In this study, we study the possible impact of such an assumption on the conclusion of the conditional logistic regression (Cordell and Clayton, 2002) if the correct modeling is tri-allelic. To do this, we used an extension of the MASC method able to generate the segregation of two genes in a trio data set. The first one (A) is tri-allelic and has an effect on the disease status; the second one (B) is biallelic, in LD with the alleles of A and has no effect on the disease. We apply the conditional logistic regression on this data set with A coded as biallelic. We show that such a coding leads to a wrong evidence for an effect of B additional to the one of A. Thus conditional logistic regression is not robust to the gene effect modeling and requires a cautious interpretation of its results.

Cordell HJ, Clayton DG. (2002). Am J Hum Genet 70: 124–141

17

Multivariate from the Univariate: a New Screening Strategy for Linkage Analysis

A. Buil(1), L. Almasy(2), J.M. Soria(1), J. Blangero(2)
(1) Institut de Recerca del Hospital de Sant Pau. Barcelona, Spain, (2) Southwest Foundation for Biomedical Research San Antonio, TX

Multivariate linkage analysis for quantitative traits in extended pedigrees is more powerful than the univariate

counterpart. However, with more than 2 traits, multivariate analysis requires a huge amount of computation time. We present here a rapid multivariate test useful for any number of traits. The test is a modification of the strategy described in de Andrade *et al.* (BMC Genetics, in press). The basic idea is that given a genetic locus, we can build a multivariate test combining the univariate linkage tests at that locus for the different traits.
Our proposed test is as follows: given n traits, we perform the standard univariate linkage analysis for each trait. Let λ_{ij} be the value of the loglikelihood ratio test for the trait **i** at genomic position **j**. Then, the multivariate test for a given position **j** will be the sum of λ_{ij} for **i** from 1 to n, and will be distributed as a complex mixture of chi squares with n terms and a point mass at zero.
We simulated 5 correlated traits, a quantitative trait locus (QTL) that explained part of the variance of each trait and a polymorphic marker linked to the QTL. Then we performed the new test for different subsets of traits. We repeated the process 10000 times and compared the power of the multivariate test with the power of performing the separate univariate analyses.
We observed that the power of the new test increased remarkably compared to the univariate. At the same time the type I error was near the expected values.

18

Efficiency of sampling designs within a cohort to estimate interaction effects between genetic and environmental risk factors

A. Bureau(1,2), M.S. Diallo(1,3)
(1) Centre de recherche Université Laval Robert-Giffard, (2) Department of social et preventive medicine and (3) Department of mathematics and statistics, Université Laval, Canada

Understanding the role that interactions between genes and environmental exposures play in complex diseases is a current challenge. To study such interactions, genotyping of candidate variants is increasingly undertaken on large prospective cohorts originally assembled to study environmental risk factors. Given the cost of genotyping the large number of subjects in these cohorts, being able to select a sub-sample for genotyping that contains most of the information of the entire cohort would lead to substantial savings. We consider nested case-control and sub-cohort sampling designs with or without stratification and compare their efficiency relative to the entire cohort for estimating the effects of genetic and environmental risk factors and their interactions. Asymptotic calculations show that for a range of scenarios for the relationships between genes, environmental exposures and disease status, a nested case-control sample balanced with respect to a risk factor preserves a high proportion of the information of the cohort. Using the interaction between Apolipoprotein E and smoking on the risk of coronary heart disease as application, we confirmed these results by simulation of the sampling designs within the Framingham Offspring Study. With 24% of the subjects in a nested case-control sample frequency matched on smoking status the relative efficiency to estimate the interaction effect on an additive risk scale reaches 73%.

19

Investigating Interactions Between Variants in Multiple Candidates Genes and Association with Breast Cancer

N.J. Camp(1), L. Hulme(2), M.W. Reed(3), A. Cox(4)
(1) Div of Genetic Epidemiology, University of Utah School of Medicine, USA, (2) Dept. of Molecular Biology and Biotechnology, (3) Academic Unit of Surgical Oncology, (4) Institute for Cancer Studies, University of Sheffield, UK

We investigated interactions between variants in multiple candidate genes and association with breast cancer using classification tree analysis. Data on 988 breast cancer cases and 996 controls were available from individuals ascertained from the North of England. Genes included in the analysis were XRCC2 (R188H) XRCC3 (A4541G, A17893G and T241M), BRCA2 (N372H) ATM (D1853N) RAD51 (G135C) and LIG4 (D501D). When analyzed independently, only XRCC3(A17893G) was significant.
An exhaustive CHAID analysis was performed on all individuals with valid data at the 8 loci (N=1,591). In the analysis, each locus was allowed to be considered as a dominant or recessive grouping of genotypes with respect to the rarer allele (2), that is, **11** vs **12,22** (dominant), or **11,12** vs **22** (recessive).
The resulting tree had 3 levels and included 4 loci: all three SNPs in XRCC3 and RAD51(G135C). As expected, the root node was split initially by XRCC3(A17893G) for a dominant grouping of genotypes, with individuals carrying the rarer allele at increased risk (p=0.008). The node with those carrying allele 2 was further split by XRCC3(T241M) for a recessive grouping with rare homozygotes at decreased risk (p=0.003). The node containing the homozygous 11 at XRCC3(A17893G) was split by RAD51(G135C) for a dominant grouping, for which carriage of the rare allele decreased risk (p=0.011). This was followed by the node for RAD51 **11,12** being split by XRCC3(A4541G) for a recessive grouping, for which rare homozygotes were at decreased risk (p=0.012). The inclusion of all three variants in XRCC3 is in accordance with other analyses we have performed investigating haplotypes in that gene. These results suggest a potential interaction of XRCC3 and RAD51 in breast cancer susceptibility.

20

Patterns of Linkage Disequilibrium for XRCC3 in Breast Cancer Cases and Controls and Haplotypes Associated with Breast Cancer

N.J. Camp(1), L. Hulme(2), M.W. Reed(3), A. Cox(4)
(1) Div. of Genetic Epidemiology, University of Utah School of Medicine, USA, (2) Dept. of Molecular Biology and Biotechnology, (3) Academic Unit of Surgical Oncology, (4) Institute for Cancer Studies, University of Sheffield, UK

We investigated associations between three SNPs in XRCC3 (A4541G, A17893G and T241M) and the occur-

rence of breast cancer in a sample of cases (N=988) and controls (N=996) ascertained from the North of England. Independently, XRCC3(A4541G) and XRCC3(T241M) were not significant. A significant association was found for XRCC3(A17893G) (p=0.002), however, this result was driven by an over-representation of heterozygotes in cases and a reduction of both homozygotes. Using SNPHAP, we estimated haplotypes and their frequencies in cases and controls separately, and characterized the linkage disequilibrium (LD) pattern for both groups using the PCA method of Horne and Camp (2004). In controls, two LD groups were extracted. The first contained SNPs at positions A17892G and T241M in strong negative LD (haplotypes 1-1-2 and 1-2-1 were the most common with frequencies 0.34 and 0.27, respectively); the second contained all three SNPs with A17892G and T241M in positive LD, and both in negative LD with A4541G (1-2-2 and 2-1-1 with frequencies 0.19 and 0.02). In cases, however, the major driving force was the negative LD between A17893G and T241M. Thus haplotypes 1-1-2, 1-2-1, 2-1-2 and 2-2-1 were seen more frequently in the cases and 1-2-2 and 2-1-1 less. Global tests for all haplotypes or all composite genotypes are both strongly significant ($p=9.2\times10^{-12}$ and 2.0×10^{-8} respectively). The interplay of a variety of haplotypes suggests that it may be important to characterize the LD structure and haplotypes in cases and controls separately before analysis, and that multiple variants in XRCC3 may be important in breast cancer susceptibility.

21

A sequential association test in family-based analysis with parental phenotypes

H.-S. Chen, Q. Sha, S. Zhang
Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931

Family-based association designs are not affected by population stratification. However, studies based on these tests almost always ignore the parental phenotypes, primarily due to the lack of suitable analytic framework. Recently, statistical methods have been proposed to include parental phenotypes. It has been shown that the incorporation of parental phenotypes in association tests can considerably increase power. In this article, we propose a sequential test incorporating the parental phenotypes. The proposed test combines a test for population stratification and a test for association. The test inherits the advantage of the family-based design that can control for population stratification and also has similar power of population-based association design when no population stratification is present.

22

Power of Variance Component Linkage Analysis in Large Pedigrees

W-M. Chen, G.R. Abecasis
Department of Biostatistics, University of Michigan, Ann Arbor, MI

Variance component linkage analysis is commonly used to map quantitative trait loci (QTLs) for quantitative traits in general pedigrees. The use of large pedigrees is especially attractive for these studies because they provide much greater power per genotyped individual than small pedigrees. However, when pedigrees are large, previously proposed analytical approaches to assess the power of the variance component linkage analysis are not feasible. We propose accurate and practical methods to calculate the analytical power of variance component linkage analysis that can accommodate large pedigrees. Our analytical power computation involves the approximation of the noncentrality parameter for the likelihood-ratio test by its Taylor expansions. Efficient algorithms to compute the second and third moments of the IBD sharing distribution enable rapid computation of the noncentrality parameter. The algorithms take advantage of natural symmetries in pedigrees and can accurately analyze many large pedigrees in a few seconds. The accuracy of our power calculation is verified via simulations of pedigrees with 2–6 generations and 2–8 siblings per sibship. We apply this proposed analytical power calculation to a recently collected data set with $>6{,}000$ phenotyped individuals where the largest pedigree includes >600 phenotyped individuals. Simulations based on 8 representative traits show that the difference between our analytical calculation of the expected LOD score and the average of simulated LOD scores is less than 0.05.

23

Examination of variance-level effects on QTL's and a framework for their examination in Mx

J. Corbett

In linkage analysis, covariate effects are usually only examined at the mean level, for example through the adjustment of phenotypes for race, sex, age, or other factors. The direct examination of these sorts of effects on the covariance structure is much rarer and, in general, has only been applied in certain special cases, such as in the case of gene-by-sex interaction effects. Even in this limited case, most analyses in which sex is thought to play a factor lose significant power by being carried out by masking the phenotypic information on one sex at a time.
This analysis shows how to examine these effects using the structural equation modeling package Mx and is accompanied by selected simulation studies to examine under what circumstances accounting for these effects increase power as well as showing that doing so does not, in general, lead to an inflation of Type I Error.

24

Consanguinity may lower the age of onset of nasopharyngeal carcinoma in North African populations

M. Corbex, W. Ben Ayoub, M. Khyatti, S. Dahmoul, M. Ayad, F. Maachi, W. Bedadra, M. Abdoun, S. Mesli, M. Hamdi-cherif, K. Boualga, N. Bouaouina, L. Chouchane, A. Benider, F. Ben Ayed & D.E. Goldgar

Nasopharyngeal carcinoma (NPC) is rare in most parts of the world, while in some populations of China, South Asia, Alaska and North Africa, it occurs in an endemic form with an incidence between 5/100 000 and 30/100 000. In these high-risk populations, NPC incidence tends to rise monotonically to reach a peaks at 45–54 years, and thereafter declines. Contrasting with this typical pattern, the North African population is characterised by a bimodal age distribution, with one peak occurring in the teens and the other at 45–50. A multicentrique case-controls study carried out in Morocco, Algeria and Tunisia allowed us to study the determinants of this North-African characteristic. Demographic and environmental information as well as DNA were obtained for 664 NPC cases and 625 controls. A young age of onset (age <30) was associated with having blood-related father and mother (OR=1.76 $p<0.02$). Marriages inside the family are common in Arabic countries (25% of marriages in Tunisia, 15% in Algeria and 23% in Morocco). A higher consanguinity in young cases may indicate the presence of one or more recessive susceptibility loci for NPC. Besides, the absence of familial concentration of NPC in the North African population advocate for the presence of more than one gene and epitasis. This result underline the interest of specific population features for genetic studies and confirm the presence of low risk recessive genes for NPC as suggested by previous segregation analyses.

25

Missing data in association studies: a multiple imputation approach applied to case/control and family data

P. Croiseau(1), E. Genin(1), H.J. Cordell(2)
(1) Genetic Epidemiology and Structure of Human Populations, INSERM U535, France, (2) Dept of Med Genetics, Univ of Cambridge, United Kingdom

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 (such as unrelated cases and controls or related family members) be genotyped at the full set of markers and that phase can be reconstructed. Individuals with missing data, either in the form of missing genotypes or unknown phase, are excluded from the analysis. This can result in an important decrease in sample size and a loss of information. A possible solution to this problem is to use multiple imputation. 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 replicate imputed data sets via one of a number of possible data augmentation algorithms.
We performed simulations to test the efficiency of this approach for case/control and case/parent trio data under several different genetic models. We found that the multiple imputation procedures generally gave unbiased parameter estimation with correct type 1 error and confidence interval coverage. Multiple imputation had some advantages over missing data likelihood methods with regards to ease of use and model flexibility. Multiple imputation methods represent promising tools in the search for disease susceptibility variants among a large set of linked markers.

26

A demonstration of oligogenic simultaneous segregation and linkage analysis to map modifier loci: two new hypertrophic cardiomyopathy loci

E.W. Daw(1), S. Shete(1), Y. Lu(1), J. Ma(1), S.N. Chen(2), G. Czernuszewicz(2), R. Roberts(3), A.J. Marian(2)
(1) MD Anderson Cancer Center, Houston TX, (2) Baylor College Medicine, Houston TX, (3) Ottawa Heart Institute, Ottawa ON, Canada

In many complex genetic disorders with an identified causative mutation, evidence for additional genetic variation has been found, leading to the search for "modifier" loci. Such loci provide further insight into disease mechanisms and treatment targets. We examined oligogenic simultaneous segregation and linkage analysis as a tool to identify modifier loci. Such analysis includes known gene effects while searching for modifier loci. We applied these methods to hypertrophic cardiomyopathy (HCM) in a 244-member pedigree with 79 affected people. The family–s causal mutation is in myosin binding protein C on ch 11. In 100 individuals, we typed 811 microsatellites.
Using demographic data and mutation status, we searched for modifier loci of left ventricular mass. We produced a Bayesian linkage score, an "L-score". In a previous simulation study based on this HCM data, the L-score cut-off for an empirical chromosome-wide p-value of 95% was 7.3; 99% was 18.9. In this analysis, two regions had L-scores >7.3, both >18.9: ch 8q12 (L-score: 21.3), and ch 10p13 (L-score: 19.3). These results suggest that oligogenic simultaneous segregation and linkage analysis is a useful in mapping modifier loci.

27

Variance Component Diagnostics using the S-Plus/R Multic library

M. de Andrade(1), E. Lunde(1), E. Atkinson(1), J. Chen(2), C.I. Amos(2)
(1) Division of Biostatistics, Mayo Clinic, Rochester, MN, USA, (2) Department of Epidemiology, UT M.D. Anderson Cancer Center, Houston, TX, USA

In our newly implemented S-Plus/R library called "multic", we have linked the analysis methods of the Multic program from the ACT software with the flexibility of S-Plus. Linkage analysis is available for single, multivariate and longitudinal traits. A screening test for determining the benefit of multivariate analysis is also included. We have added new graphical diagnostic tools to check for influential individuals and families and for checking normality assumptions. We have also incorporated bootstrapping and jackknife tools to help confirm results at a given locus. The software will be publicly available by the time of the meeting, with user-friendly documentation.

28

Effects of body height and weight in genetic studies of bone geometry

S. Demissie(1), S. Menn(2), L.A. Cupples(1), D. Karasik(2), D.P. Kiel(2)
(1) Biostatistics, Boston Univ. School of Public Health, (2) Hebrew SeniorLife & Harvard Medical School, Boston, MA

Femoral geometry contributes to bone strength and predicts hip fracture risk. There is high correlation between height (HT), weight (WT) and bone geometry indices. It is unknown if these correlations are due to effects of HT or WT and the appropriate control for these variables is unclear. The goal of this study was to evaluate the effect of adjustments for HT and WT on heritability (h^2) estimates and linkage signals of femoral geometric measures. We studied endosteal diameter (END), subperiosteal width (WID), cross-sectional area (CSA), and section modulus (Z) in the proximal hip narrowest neck (NN), intertrochanteric (IT), and shaft (S) regions. h^2 and linkage were evaluated using maximum likelihood variance components in 1227 Framingham Heart Study subjects from 322 pedigrees, ages 40-80. We used cohort, sex, age and physical activity-adjusted normalized residuals (M1) with further adjustment for WT (M2), HT (M3), and WT and HT (M4). h^2 estimates ranged from 44%–70% for M1; and 34%–51% for M4. Significant LOD scores, with notable variability among different models, were observed for NN-WID & -Z; IT-END & -WID; and S-CSA (e.g., for NN-WID on Chr 12 at $\sim$15 cM, LOD for M1=3.7, M2=3.6, M3=1.3, M4=1.4). Stronger linkage signals from HT- (WT) unadjusted models may suggest QTLs that control both HT (WT) and bone geometry, while those from adjusted models point to QTLs for bone geometry independent of HT (WT). In genetic analyses of bone geometry investigation of multiple adjustment models may be warranted.

29

Semiparametric Variance-Component Methods for Mapping Quantitative Trait Loci With Censored Data

G. Diao, D.Y. Lin
Department of Biostatistics, University of North Carolina, USA

Variance-component (VC) methods are widely used in the linkage and association analysis of quantitative traits in general human pedigrees. The standard VC methods assume that the trait values within a family follow a multivariate normal distribution and are fully observed. These assumptions are violated if the trait data contain censored observations. When the trait pertains to the age at onset of a disease, censoring is inevitable because of loss to follow-up and limited study duration. Censoring also arises if the assay cannot detect values smaller (or larger) than some threshold. Applying the standard VC methods to such censored trait data would result in inflated type I error and reduced power. We develop valid and powerful VC methods for censored trait data based on a novel class of semiparametric linear transformation models. Under the proposed models, the latent trait values follow a specific distribution, such as the normal distribution, after a completely unknown transformation.

We construct appropriate likelihood functions for the observed data, which may contain left or right censored observations, and devise efficient algorithms to implement the corresponding estimation and testing procedures. Our methods can be used for both linkage and association analysis. Extensive simulation studies demonstrate that the proposed methods outperform the existing methods in practical situations. We illustrate the usefulness of the new methods through the age-at-onset of alcohol dependence data from the Collaborative Study on the Genetics of Alcoholism.

30

Mitochondrial Genetic Effects $\times$ Age Interaction in a Marker of Oxidative Stress in the San Antonio Family Heart Study

V.P. Diego(1), D.L. Rainwater(1), D. Winnier(1), X.-L. Wang(2), L. Almasy(1), J.T. Williams(1), J. Blangero(1), J.W. MacCluer(1), M.C. Mahaney(1)
(1) Southwest Foundation for Biomedical Research, (2) Baylor College of Medicine, USA

Oxidative stress plays an important role in cardiovascular disease and mitochondria are main contributors to oxidative stress. Because mitochondrial effects are age-dependent, we developed a model that can detect mitochondrial genetic effects along the age continuum. Using this mitochondrial genetic effects $\times$ age interaction model, we analyzed two markers of oxidative stress, namely platelet activating factor acetylhydrolase (PAF-AH) and paraoxonase-1 (PON1), and an indicator of redox homeostasis, namely plasma total antioxidant status (TAS), in Mexican American families participating in the San Antonio Family Heart Study (SAFHS). For three clinical visits for which we have data, we report heritabilities for PAF-AH ranging from 0.46 to 0.55, for PON1 ranging from 0.81 to 0.88, and for TAS ranging from 0.25 to 0.54. PAF-AH for the second clinical visit exhibited evidence of significant mitochondrial genetic effects $\times$ age interaction (p=0.00982). Our model allows for two sources of interaction: 1) variance heterogeneity in the mitochondrial genetic effects and 2) heteroplasmy—heterogeneous mitochondrial genome arising from mutation and drift—which drives an age-related decay in the correlation structure of mitochondrial genetic effects. We found that the mitochondrial genetic effects $\times$ age interaction for PAF-AH was due only to variance heterogeneity. Specifically, we found that the mitochondrial genetic variance significantly increased with age (p=0.00382). To our knowledge, this is the first report of mitochondrial genetic effects $\times$ age interaction in a marker of oxidative stress. Our model is novel in that it jointly accounts for heterogeneity in the mitochondrial genetic variance and heteroplasmy.

31

Estimating two-locus disease model parameters in a lod score analysis with Genehunter-Twolocus

J Dietter(1,2), K Lenzen(3), T Sander(3), K Strauch(1,2)
(1) Institute for Med. Biometry, Informatics, and Epidemiology, Univ. of Bonn, Germany, (2) New address: Institute for Med. Biometry and Epidemiology, Univ. of Marburg, Germany, (3) Gene Mapping Center, Max-Delbrück-Centrum, Berlin, Germany

The choice of an appropriate disease model is particularly difficult for a parametric linkage study with two disease loci. For this reason, the two-locus lod score is sometimes maximized with respect to the disease model parameters. In order to avoid the incompleteness involved with a maximization by hand we have implemented a mod-score functionality into Genehunter-Twolocus. For optimization we have chosen a simulated annealing procedure which can be smoothly tuned from a stochastic search to a deterministic search. The resulting model parameters can be checked to some degree by looking directly at the dependency of the maximum lodscore on the disease model parameters. Here, up to three disease model parameters are chosen while leaving the other parameters fixed. The graphical representation of this dependency is done by means of a color coded three dimensional plot. We show the results of an application of these new methods to a data set of idiopathic generalized epilepsy.

32

Genetic Models for Thyroid Cancer Based on the Swedish Family Cancer Database

A. Ding(1), A. Antoniou(1), K. Czene(2), K. Hemminki(2,3), D.F. Easton(1)
(1) Cancer Research UK, Dept. of Public Health and Primary Care, University of Cambridge, (2) Karolinska Institute, Stockholm, (3) German Cancer Research Center (DKFZ), Heidelberg

To evaluate genetic models for thyroid cancer (TC), we conducted segregation analyses using data obtained from the Swedish Family-Cancer Database. 7465 pedigrees ascertained through a proband with TC, including 201 with a family history of the disease, were included. Likelihoods were computed using routines implemented in MENDEL. Survival analysis methods were adapted to model age- and sex-specific hazard rates, constraining the overall incidence to agree with population rates. A generalized estimating equation approach was used to account for multiple ascertainment of probands. Our results suggest that major gene and pure polygenic models fit poorly. The best fitting model was a mixed model, including a rare dominant major locus (q=0.00003) conferring a relative risk (RR) of 6800 (890-51940) to carriers and an additional additive polygenic component with standard deviation 2.4 (1.2-3.7). Although the RR due to the major locus appeared to decrease with age, the trend test was not significant. There was no significant difference in the RR between males and females. Cumulative risks, however, were lower in males (59% vs 87% by 80 years, averaged over polygenotypes). The predicted familial risks to offspring of affected individuals were comparable to those observed. These findings suggest that genes conferring high risks of TC may be identifiable, but that lower risk susceptibility alleles should also be taken into account in counselling.

33

Indication of interaction between genetic susceptibility to asthma and passive exposure to tobacco smoke by a genome-wide screen in 110 French EGEA families

M.H. Dizier(1), M. Guilloud-Bataille(1), V. Siroux(2), E. Bouzigon(3), F. Demenais(3)
(1) INSERM U535, Villejuif, France, (2) INSERM U578, La Tronche, France, (3) INSERM EMI 0006, Evry, France

Asthma is a multifactorial disease resulting from many genetic and environmental factors. Passive exposure to tobacco smoke (ETS) in early life is one of the risk factors of asthma. A genome scan for asthma, recently carried out in 110 French EGEA families with at least two asthmatic sibs, showed indication of linkage of asthma to the 1p34 region (p=0.005 at 75 cM from pter). Our present aim was to investigate interactions between genetic susceptibility to asthma and passive ETS in early life by performing the genome scan in smoke-exposed and non-exposed sib pairs. The predivided sample test (PST) was used to compare the IBD (identity by descent) distribution estimated by the Maximum Likelihood Score (MLS) method at all 378 autosomal markers of the genome scan between these two subsets. A gene-ETS interaction was detected (p=0.002) in the 1q32 region (PST statistic being maximal at 212 cM from pter). Linkage was only apparent in smoke-exposed sib-pairs (p=0.002), while there was no indication of linkage in non-exposed pairs (p=0.29). This result indicates that passive smoke exposure in early life modulates the effect of gene(s) in 1q32 region on asthma susceptibility. Further linkage analyses will be conducted for bronchial hyper-responsiveness, an asthma-related phenotype which is also associated with passive ETS.

34

Determining Optimal Ratios of Affected to Unaffecteds in a Propensity Score Calculation to Maximize Power Gains for the Identification of the Lung Cancer 6q23-25 Linkage Peak

BQ Doan(1), D Behnemann(1), JE Bailey-Wilson(1) and the GELCC(3)
(1) IDRB/NHGRI/NIH, (2) Gen. Epi. Lung Cancer Consortium

Using a propensity score (PS) for affection (predicted probability of being affected given covariates, which is estimated from a logistic regression) may increase power in nonparametric covariate-based linkage analyses. This power gain may be attributed to the inclusion of unaffected individuals in the PS. However the optimal ratio of unaffected individuals (UI) to affected individuals (AI) in the PS calculation is unknown, and power gains

were observed in datasets with equivalent or more AI than UI. For late onset diseases like some cancers, more UI than AI are typically ascertained to account for missing parents. We explore means to optimize construction of the PS, by varying the ratio of AI to UI (2:1, 1:1, 1:2, 1:3) in the PS calculation, for lung cancer where a linkage peak on 6q was recently identified. A ratio of 2:1 yielded the max LOD score of 1.7 for the PS (based on pack-year (PY) and PY^2) compared to a LOD of 1.0 without covariates. We also compared linkage results using a PS covariate (based on PY) or a PY covariate to see if power gain is from UI, and found that direct use of PY resulted in a higher LOD score, cautioning against the blanket use of a PS. To help provide guidelines for constructing the PS, we further explore these questions in a simulation study with known genetic and covariate effects, and use these guidelines to re-examine the lung cancer dataset.

35

Relationship among the Correlations of Liabilities and Binary Traits, and Odds Ratios

Y. Dong, Y. Luo, R.C. Elston
Dept. of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve Univ., USA

Binary Traits have been analyzed via liability, a continuous random variable, for a long time in the studies of statistical genetics. However, the results from the analyses sometimes are not reliable due to the nature of discrete disease status. So, it is desirable to formulate and compare the statistical characteristics of both liability models and corresponding binary traits. The authors developed a numerical method to formulate the relationship between the correlation coefficient of the liabilities of two relatives and the correlation coefficient of the two relatives–binary traits. The related error analysis is attached. Furthermore, a mathematical expression is derived for the relationship between the correlation coefficient of two relatives– binary traits and the odds ratio which is one of the measures on how likely a relative is affected given the other being affected. Comparing to the work in *Hopper J: Encyclopedia of Biostatistics, Edited by P. Armitage and T. Colton, John Wiley & Sons Ltd., 1998; 4626–4629.*, the assumption of common prevalence is not needed.

36

Simple multipoint statistic for linkage analysis based on sib pair

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

Multipoint linkage analysis is a powerful tool to localize susceptibility genes for complex diseases. Liang et al. (2001) proposed a new identity-by-descent (IBD) based procedure to estimate the location of an unobserved susceptibility gene within a chromosomal region framed by multiple markers. The method is robust in that no assumption about the genetic mechanism is required. Analyzing IBD status of several markers together based on Liang et al. (2001) study, I found a specific IBD pattern on a chromosome that can be used to reveal candidate chromosomal intervals for genes associated with the disease. Implementing this idea, I found the estimate of the location of the susceptibility genes and derived a regression-based simple multipoint statistic for genome-wide linkage scanning. The empiric distribution function (Monte Carlo simulations) is used to obtain the exact genome-wide significance level. Bonferroni correction is necessary not for the number of markers but for the number of candidate regions. The new method have been applied to the simulated data from the 11th Genetic Analysis Workshop, Problem 2 (Greenberg et al., 1999) and to the data on Alzheimer disease collected by the National Institute of Mental Health Genetics Initiative (Blacker et al., 1997). The new method of linkage analysis is used to reconfirm known genes, to find new candidate genes, and (in the future) to study genes interaction that results in manifestation of common diseases. Examples of application on both simulated and real data will be presented.

37

A Physiologically-based Pharmacokinetic Modeling Platform for Genetic and Exposure Effects in Metabolic Pathways

L.Du, D.V. Conti, D.C. Thomas
Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

Metabolic pathways have provided fresh insights into the etiology of complex diseases. However, sophisticated methodology needs to be developed to interpret observational data aimed at identifying key components of the pathways. We present a hierarchical modeling platform employing enzymatic reaction kinetics to characterize the joint effects of genes on the metabolic activation and detoxification of carcinogenic exposures. We simulate datasets containing exposure, genotype, and disease information for individuals under a known metabolic model. Individual-specific metabolic rates are described by Michaelis-Menten enzymatic parameters Vmax and Km, which are regressed on genotypes at the relevant loci. If available, proteomic and metabolomic data can be incorporated into the platform by treating them as flawed measurements of an individual–s long-term metabolic rates and metabolite concentrations, respectively. Posterior estimates of the parameters are obtained for simulated datasets using Markov Chain Monte Carlo methods. The simulation and estimation platform allows investigating the precision of the parameter posterior estimates in relation to the magnitude of the simulated effects and individual variability in metabolic rates, and comparing alternative designs incorporating proteomic or metabolomic data. We also examine the robustness to model specification and compare results with simpler descriptive models. The approach appears to have great utility for characterizing genetic and exposure effects and dissecting complex pathways.

38

Identification of Novel QTL for Intraocular Pressure, which co-localizes with Blood Pressure Loci

P. Duggal(1), A.P. Klein(2), K.E. Lee(3), R. Klein(3), J.E. Bailey-Wilson(1), B.E. Klein(3)
(1) NIH/NHGRI/IDRB, Baltimore, MD (2) Dept of Oncology & Pathology, JHMI, Baltimore, MD (3) Dept. Ophtm & Visual Sciences, UW-Madison Medical School

Elevated intraocular pressure (IOP) is a principal risk factor for primary open-angle glaucoma (POAG). We performed a non-parametric genome-wide scan (GWS) of 1019 sibling pairs with measurements for IOP. The sibpairs were ascertained through a population-based cohort, the Beaver Dam Eye Study, Wisconsin. The higher IOP measurement between eyes was used as a continuous trait, and treatment with drops, systolic blood pressure, sex and age were covariates adjusted prior to analysis. We performed singlepoint and multipoint linkage analysis using the modified Haseman-Elston regression models in SIBPAL(SAGEv4.5). P values were obtained using the asymptotic distribution of the likelihood ratio test statistics and we obtained empirical p-values using Monte Carlo permutations. We also performed VC linkage analysis on nuclear families in Merlin. The GWS identified 3 loci as regions of interest, of which 1 reached suggestive linkage according to Lander & Krugylak thresholds for sibs and half-sibs in genome wide scans. This novel linkage region had empirical p-values of 2×10^{-4} (singlepoint) and 5.1×10^{-5} (multipoint), respectively. These 3 regions are especially interesting since each has been identified as potential linkage regions in GWS for blood pressure. The results of this GWS provide evidence that a QTL may influence elevated IOP and may co-localize with blood pressure loci. These loci may control systemic pressure reflected in the eye and blood.

39

Structured Incorporation of Prior Information in Identification Problems

T Egeland(1) and NA Sheehan(2)
(1) Dept. of Medical Genetics, Ullevaal University Hospital 0407 Oslo, Norway, (2) Dept. of Health Sciences and Dept. of Genetics, University of Leicester, UK

Determining the relationships between specified individuals is frequently of interest. Applications are diverse in behaviour, evolution and conservation research and include forensic problems ranging from standard paternity cases to inheritance claims, immigration cases and identification of remains following disasters. For human applications, standard forensic DNA markers, methods and software often suffice but important problems requiring less routine approaches can arise. We consider cases where DNA data alone is inconclusive and additional non-DNA information required. For instance, two individuals can be classified as parent and child from DNA marker data but determining which is not possible without other information, such as age. Extra information is often used in practice but in a generally ad hoc manner. We will follow along the lines proposed by Egeland et al. (2000) and implemented in the freeware program familias (www.nr.no/familias), by placing a prior distribution over the set of possible relationship structures and considering posterior probability ratios, rather than classic likelihood ratios, to compare any two alternatives. We note that some issues pertaining to relationship identification are relevant to error checking of pedigree data. Other Bayesian models have been proposed previously for error checking, e.g. Göring and Ott, 1997.

Egeland T, Mostad PF, Mevåg B and M Stenersen. Forensic Sci Int 2000;110:47–59
Göring HH and Ott J. Eur J Hum Genet. 1997 Mar–Apr;5(2):69–7

40

A penalized likelihood approach for linkage analysis of quantitative traits under locus heterogeneity

C.T. Ekstrøm
Dept. of Natural Sciences, Royal Veterinary and Agricultural Univeristy, Denmark

Genetic heterogeneity is one of the major problems in linkage analysis of quantitative traits. The presence of genetic heterogeneity decreases the power of a linkage study and may result in biased estimates of the locations of the quantitative trait loci.
Mixtures of multivariate Gaussian distributions can model locus heterogeneity for linkage analysis of quantitative traits but the asymptotic distribution of the test statistics for Gaussian mixture models is undefined. We examine the use of a penalized likelihood suggested by Chen and Chen (2000) and discuss the use of a modified test statistic for genome-wide linkage analysis of quantitative traits in the presence of locus heterogeneity. A simulation study is used to illustrate the usefulness of the method.

41

Estrogen receptors and CYP450 enzymes involved in estrogen metabolism in French-Canadian women: Associations with bone measures and susceptibility to osteoporosis?

L. ELfassihi(1,2), S.Giroux(1), S. Dodin(1), K. Morgan(3), N. Laflamme(1,2,4), and François Rousseau(1,2,5)
(1) Centre de Recherche de l Hôpital St-Francois d Assise du Centre Hospitalier Universitaire de Quebec, Quebec City, Quebec, Canada, (2) Department of Medical Biology, Faculty of Medicine, Laval University, Quebec City, Quebec, Canada, (3) Department of Human Genetics, McGill University, Montreal, Quebec, Canada, (4) Institut National de Santé Publique du Quebec, Ste-Foy, Quebec, Canada, (5) Centre for the Development, Evaluation and Rational Implementation of New Diagnostic Tools in Medicine (CEDERINDT)/Consortium Interdisciplinaire d Evaluation des Technologies Diagnostiques du Quebec (CETDEQ), Quebec City, Quebec, Canada

Bone mineral density (BMD) has a significant heritability. Since estrogen is required for attainment of peak

bone mass, we have investigated associations between five bone measures and six single nucleotide polymorphisms (SNPs) within five genes: estrogen receptor α (ESR1) intron 1 (T/C (PVUII)) and (A/G (XbaI)), estrogen receptor α (ESR2) exon 5 (Val328Val), CYP1A1 exon 7 (Ile462Val), CYP1B1 Exon 2 (Ala119Ser) and CYP17 promoter region (T/C), in a cross-sectional study of 2886 postmenopausal and 1335 premenopausal healty French-canadian women. **MATERIALS AND METHODS**: Three Heel bone parameters (BUA,SOS,STIFF) were measured by right calcaneal QUS in 1335 premenopausal and 2886 postmenopausal women. Half (690) premenopausal and 1093 postmenopausal had also their BMD evaluated at two sites: femoral neck (FN) and lumbar spine (L2–L4) by DXA. All bone measures were tested separately for association with each SNP genotypes by analysis of covariance for premenopausal and postmenopausal groups.
RESULTS: All SNPs were in Hardy-Weinberg equilibrium. Two markers within ESR1 gene were in complete linkage disequilibrium ($|r|=0.79$, $|D-|=1$). Only three of the four possible haplotypes were observed in the population under study. A statistically significant association between PVUII-XbaI haplotype and femoral neck BMD was observed in premenopausal group: women carrying two TA haplotypes had a 3.4% (0.03 g/cm^2) lower femoral neck BMD than those carrying the other haplotypes ($p=0.0033$), independently of the other risk factors adjusted for. The interaction between smoking and ESR2 genotype was associated with BUA measure in postmenopausal group ($p=0.0051$) as already reported in a smaller group (1189); among women carrying GA+AA genotype, smokers had lower BUA than no-smokers (the Bonferroni adjusted p-value is 0.0013). However, no association was observed between CYP450 polymorphisms and the five bone measures in the two groups analyzed.
CONCLUSION: Estrogen receptors are significantly associated with bone measures in both premenopausal and postmenopausal groups but not influenced by CYP450 polymorphisms studied.

42

Physical activity modifies the genetic effect of vitamin D on insulin secretion in Hispanic and African Americans: the IRAS Family Study

CE Engelman(1), TE Fingerlin(1), CD Langefeld(2), SS Rich(2), RN Bergman(3), DW Bowden(2), JM Norris(1)
(1) U of Colorado at Denver and Health Sciences Center, (2) Wake Forest U School of Medicine, (3) U of Southern California

Polymorphisms in the vitamin D receptor (VDR) gene have been implicated in explaining the variation in insulin secretion (IS). Physical activity (PA) is also associated with IS. Fourteen SNPs in the VDR gene (average spacing of 4.5 kb) were typed in 90 Hispanic American (HA) families (1424 subjects) and 42 African American (AA) families (604 subjects) from the IRAS Family Study. We used the family based association test (FBAT) software to investigate the association of VDR SNPs and IS, as measured by acute insulin response to glucose (AIRg). The association was explored over all levels of PA and in the upper and lower 50th percentiles of 1-year PA. In AAs, the CGC haplotype (formed by rs9729, rs1544410 and rs2239185) in the VDR gene was not associated with AIRg overall ($p=0.12$) or in active individuals ($p=0.31$), but significantly associated ($p=0.02$) with lower AIRg in inactive individuals. In HAs, a single SNP in the VDR gene (rs10735810) was not associated with AIRg overall ($p=0.12$) or in active individuals ($p=0.97$), but significantly ($p=0.02$) associated with AIRg in inactive individuals. These results suggest an important role of PA in modifying the effect of the VDR gene on IS in HAs and AAs.

43

Improved Association Analyses of Disease Subtypes in Case-Parent Triads

M.P. Epstein(1), I.D. Waldman(2), G.A. Satten(3)
(1) Dept. of Human Genetics, Emory Univ., USA, (2) Dept. of Psychology, Emory Univ., USA, (3) Centers for Disease Control and Prevention, USA

The sampling of case-parent triads is an appealing strategy for conducting association analyses of complex diseases. In certain situations, one may have interest in using the triads to identify genetic variants that are associated with a specific subtype of disease, perhaps related to severity of symptoms or sensitivity to medication. A straightforward strategy for conducting such a subtype analysis would be to analyze only those triads with the subtype of interest. While such a strategy is valid, we show that triads without the subtype of interest provide additional genetic information that increases power to detect association with the subtype of interest. We incorporate this additional information using a likelihood-based framework that permits flexible modeling and estimation of allelic effects on disease subtypes and also allows for missing parental data. Using simulated data under a variety of genetic models, we show that our proposed association test consistently outperforms association tests that only analyze triads with the subtype of interest. We also apply our method to a triad study of attention-deficit hyperactivity disorder and identify a genetic variant in the dopamine transporter gene that is associated with a subtype characterized by extreme levels of hyperactive-impulsive symptoms.

44

Power Implications of the Dichotomized Continuous Trait FBAT

D. Fardo, C. Lange
Dept. of Biostat, Harvard School of Public Health, USA

Family-based association tests (FBATs) are used to test associations between a genetic marker and a disease susceptibility locus for many different scenarios. Here we examine the effects of using a binary affectation status based on a continuous phenotype. Based on simulation studies and analytical considerations, we propose a rule for this dichotomization that minimizes power loss under fairly general circumstances. The proposed approach is illustrated using a genetic asthma study.

45

A QTL on 15q21 influences HDL-Cholesterol: NHLB-FHS

M.F. Feitosa(1), M.A. Province(1), G. Heiss(2), D.K. Arnett(3), R.H. Myers(4), J.S. Pankow(5), P.N. Hopkins(6), I.B. Borecki(1)
(1) Washington Univ, (2) Univ North Carolina, (3) Univ Alabama, (4) Boston Univ, (5) Univ Minnesota, (6) Univ Utah, USA

Low plasma high-density lipoprotein cholesterol (HDL-C) is a potent risk factor for coronary heart disease susceptibility and is heritable within families, probably being influenced by multiple genes, environmental factors and their interactions. We sought to identify loci influencing HDL-C by combining two groups of families using a variance components based linkage approach. Genotypes for 988 White subjects in 267 sibships were obtained for 243 STR markers typed by the Utah Molecular Genetics Laboratory, and 2,666 White subjects distributed among 397 largest pedigrees were typed for 402 markers by the Mammalian Genotyping Service. African-Americans will be investigated once genotyping has completed. Suggestive evidence of a QTL (Empirical LOD=1.6, p=0.003154) influencing the variation of age-sex- adjusted HDL-C was found on chromosome 15q21. Adjusting HDL-C for age, sex, body mass index (BMI), smoking (SK) and alcohol (ALC) consumption significantly enhanced the evidence for linkage (Empirical LOD=3.6, p=0.000023). Exclusion of diabetic subjects from the sample corroborated linkage at the same chromosome 15 location for age-sex-BMI-SK-ALC adjusted HDL-C (Empirical LOD=3.3, p=0.000048, despite a ∼36% reduction in the number of sib-pairs). This study suggests that the search for genes influencing HDL-C levels can be compromised by phenotypic heterogeneity and the power to detect this QTL on 15q21 increased by excluding individuals without perturbations in glucose metabolism.

46

Multipoint linkage analysis when linkage equilibrium present among tightly linked markers

Z. Feng(1), H. Zhao(1), W. Wong(2)
(1) Dept. of Epi & Public Health, Yale Univ., USA, (2) School of Computer Science, Univ. of Waterloo, Canada

In complex disease gene mapping, linkage tests based on allele sharing among the affected relatives (e.g., sib pairs, parent-offspring pairs and first cousins) are commonly used. These approaches have the advantage of not having to specify the inheritance model. In multipoint linkage analysis, all the methods currently used assume linkage equilibrium among markers including GENEHUNTER (Kruglyak et al., 1996), Allegro (Gudbjartsson et al., 2000) and Merlin (Abecasis et al., 2002). When the phase information is unknown, based on the assumption of linkage equilibrium, equal probability are assigned to all possible phases that are compatible with the observed genotype data (O–Connell and Weeks, 1995, Kruglyak et al., 1996). Nowadays, dense set markers are gaining more popular use. Although much information can be obtained with these markers, there is often strong association among them. In this scenario, when phase information is unknown, linkage analysis based on all current programs may lead to inflated false positive results (Huang et al., 2004). Motivated by this problem, we propose to develop a new statistical framework to infer the allele sharing status at candidate sites among the affected relatives. We extend the currently used Hidden Markov model to incorporate linkage disequilibrium. We conduct simulation study to demonstrate the improvement of our method by comparing to GENEHUNTER. To illustrate the use of our new method, it is applied to linkage analysis of real data (GAW14).

47

Searching for SNP combinations associated with disease susceptibility

B.-J. Feng(1), D.Goldgar(2), C.M. Van Duijn(1)
(1) Dept of Epidemiology & Biostatistics, Erasmus Medical Center, Netherlands, (2) University of Utah, Salt Lake City

We have adapted a guided search algorithm, the Genetic Algorithm (GA) for application to large-scale population-based association studies. This algorithm can search for combinations of risk factors associated with disease susceptibility, considering both interactions and heterogeneity. Significance of the findings is tested by permutation and likelihood ratio test. To test this algorithm we simulated a model with 8 causal mutations located on separate chromosomes. The relationships between these causal mutations in the etiology of the disease included multiplicative interactions, threshold effect interactions and genetic heterogeneity. Their Main Effects ranged from 1.18 to 2.0. Genotypes of 1025 random noise SNPs and these 8 causal mutations were simulated on 500 cases and 500 unrelated controls. Random missing genotypes are set as 2%. The results of the simulation studies demonstrated that the GA algorithm is more powerful than other approaches, including univariate analysis with FDR or bonferroni correction, logistic regression, CART, and MDR.

48

Can we estimate the local false discovery rate in large scale association studies?

C. Fischer(1), L. Beckmann(2), M. Obreiter(2), A. Kindler-Röhrborn(3), J. Chang-Claude(2)
(1) Human Genetics, Univ. of Heidelberg, Germany, (2) Unit of Genetic Epidemiology, DKFZ, Heidelberg, Germany, (3) Dept. of Neuropathology, Univ. of Bonn, Germany

We applied three methods to explore the utility of the local false discovery rate (FDR) in a genomewide association study. As an alternative to p-values the FDE for a hypothesis H_i is the probability that H_i is true conditioned on all observed p-values. As an alternative to p-values the FDR has been proposed to identify a certain percentage of interesting hypotheses for further investigation and it is commonly used in microarray experiments.
Methods: 1. Twilight (Scheid and Spang, 2005 Bioinf 21(12):2921–2); 2. LocFDR (Efron and Tibshirani, 2002

Genet Epidemiol 23(1):70–86); 3. PermFDR (Ge et al., 2003 Test 12(1):1–77). Data: F2 intercross rats from a highly susceptible and a completely resistant inbred strain were analyzed with respect to malignant schwannoma development.
Conclusions: Twilight requires no specification of the a-priori probability π_0 of true null hypotheses. Dependencies between markers lead to non uniformely distributed p-values and clustering patterns in regions of densely spaced markers. LocFDR can handle dependencies but local FDR values differ according to the choice of spline types for estimating the density under the global null. In PermFDR the density of p-values for true hypothesis is estimated by a resampling approach. The FDR estimates were displaced depending on a constant corresponding to π_0. None of the methods can be recommended without further investigation of the observed bias.

49

Meta-analysis of association studies confirms an association between MDR1 and ulcerative colitis

S.A. Fisher, C. Onnie, C.M. Lewis, C.G. Mathew
Guy's, King's and St Thomas' School of Medicine, King's College London, UK

Several studies have reported association between polymorphisms in the multidrug resistance gene 1 (*MDR1*) and ulcerative colitis (UC). However, other studies have failed to replicate this finding. We have genotyped the C3435T polymorphism in a collection of 580 British UC cases and 280 controls. Results from this study were combined with six published association studies of this polymorphism in UC containing a total of eight independent populations (3 German, 2 British, 1 North American, 1 Jewish and 1 Slovenian) in a meta-analysis. The combined cohort across all studies comprised 1743 UC cases and 2931 controls. Pooled odds ratios were obtained under both fixed and random effects models.
In our own association study, the frequency of the 3435T allele in cases (54.5%) was not significantly different from controls (53.9%) (OR=1.02, 95% CI 0.84, 1.25). However, from meta-analysis of these data with eight previously published studies, the pooled odds ratio under a random effects model was significant (p=0.022) (OR=1.11, 95% CI 1.01, 1.23). No significant between-study heterogeneity was observed (Q-statistic, p=0.35). A similar risk was obtained under a fixed effects model (OR=1.11, 95% CI: 1.02, 1.22). In all individual studies, the 95% confidence interval for the OR was consistent with the OR obtained from meta-analysis. These results suggest that the MDR1 gene does play a role in UC susceptibility, but the disease risk associated with the C3435T polymorphism is low.

50

Fetal Alcohol Syndrome and 3-D Facial Imaging: A Preliminary Classification Study

L. Flury(1), J. Rogers(1), T. Foroud(1), L. Robinson(2), E. Moore(3)
(1) Indiana University SOM, Indianapolis, USA, (2) SUNY Buffalo, USA, (3) St. Vincent–s Hospital, Indianapolis, USA

Intro: The effects of prenatal alcohol exposure lie on a phenotypic continuum spanning structurally normal to dysmorphic facial features. Typically, the presence of specific dysmorphic features such as short palpebral fissures or a smooth philtrum are required for the diagnosis of fetal alcohol syndrome (FAS). The objective of this study was to ascertain if analyses of 3-dimensional (3-D) images can be utilized for more effective clinical diagnosis of FAS as well as the more broadly defined Fetal Alcohol Spectrum Disorders (FASD).
Methods: Laser images of the face of 19 subjects with FAS and 22 control subjects without prenatal alcohol exposure were obtained and 3-D facial images were created. For each image, 21 morphological measurements were calculated and used in a linear discriminant analysis, with age. Two models were employed: (1) a stepwise model and (2) principal component (PC) analyses. The jackknife method was used to compute misclassification error rates.
Results: The stepwise function performed better, correctly classifying 82% of controls and 84% of FAS subjects. The PC approach utilized 2 PCs: PC1 was a weighted average of most measures, and PC2 distinguished facial breadth variables in contrast with facial height. The PC model correctly classified 77% and 74% of controls and FAS subjects respectively.
Conclusions: Having established the value of 3-D images for FAS classification, we are now pursuing studies to evaluate if facial measurements from 3-D images can also discriminate individuals with prenatal alcohol exposure who do not have the full dysmorphic expression of FAS. Employing the moderating variable of the ADH1B*2 allele in future analyses can improve the validity of FASD diagnosis.
Support: 5U24AA014809.

51

The invalidity from stratification of affected sib pairs on covariates with possible genetic determinants: problem and solution strategies

C.E. Frangakis(1), F. Li(1), B.-Q. Doan(2)
(1) Dept. of Biostatistics, and (2) Dept. of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University

The affected sib/relative pair design often uses covariates in order to increase power for finding loci linked to the studied disease. Standard methods that stratify on covariates, implicitly assume the usual null distribution of IBD sharing is preserved. However, such ''covariates'' (e.g., comorbidities or addictions) can also have genetic determinants. As the studied markers approach those determinants, the similarity created from the stratification on such covariates can induce increased IBD sharing even under no linkage to the disease. We show that, as a consequence, standard methods for using covariates can declare linkage between the disease and marker loci even when no such linkage exists. We also provide strategies for addressing the problem.

52

The Powerful Sting of the WASP: the Weighted Affected Sib Pair Mean Test

D. Franke, A. Ziegler
Institute of Medical Biometry and Statistics, University at Lübeck, Germany

For the analysis of affected sib pairs (ASPs), a variety of test statistics can be applied in genome wide scans using microsatellite markers. Even in multipoint analyses, these statistics might not fully exploit the power of a given sample, as they do not account for incomplete informativity of an ASP. We propose a new class of test statistics based on the mean test for ASPs. Families are weighted inversely proportional to their marker informativity. The weighting scheme itself is based on the distribution of alleles identical by descent as represented in the de Finetti triangle. We derive the limiting distribution, demonstrate its validity in simple Monte-Carlo simulation studies, and show that the weighted mean test can be far more powerful than the classical mean test. In addition, we re-analyze two published data sets. In both applications, the weighted mean test outperforms the classical approaches substantially. An extension of the WASP to sibships >2 or affected relative pairs is required. Furthermore, the behavior of the test statistic in the presence of missing parental data and misspecified allele frequencies needs to be studied.

This work was supported by the Deutsche Forschungsgemeinschaft (ZI 591/12–1).

53

Bivariate Analysis of Body Weight and Energy Expenditure in Young Twins

PW Franks(1), AD Salbe(1), E Ravussin(2), WC Knowler(1), RL Hanson(1)
(1) NIDDK, Phoenix, AZ, USA, (2) PBRC, Baton Rouge, LA, USA

Body weight (WT) and total energy expenditure (TEE) are strongly correlated, but the extent to which the relationship reflects genetic or environmental factors is unknown. To examine this issue, bivariate genetic analyses were undertaken in 57 monozygotic and 36 dizygotic same-sex twin pairs aged 4–10yrs. TEE was determined by the doubly-labeled water method. The correlation between WT and TEE was 0.79, after adjustment for age, sex, ethnicity, birth year and season of study. Bivariate variance components methods were used to estimate the proportion of variance for each trait due to additive genetic (a2), shared environmental (c2) and unique environmental (e2) factors and to estimate the correlation among the genetic (rA), common (rC) and unique (rE) environmental determinants. Both TEE and WT were strongly determined by genetic factors (for TEE a2=0.70, 95% CI 0.60-0.83; for WT a2=0.88, 95% CI 0.82–0.91); however TEE, but not WT, also had a significant shared environmental component (for TEE, c2=0.20, 95% CI 0.02-0.40, for WT c2=0.00, 95% CI 0.00-0.13). There was nearly complete overlap between genetic determinants of WT and TEE (rA=0.96, 95% CI 0.88-1.00), while there was significant, but incomplete overlap between unique environmental determinants (rE=0.54, 95% CI 0.33-0.69). These analyses suggest that familial aggregation of TEE and of WT in these children reflects, to a large extent, the action of genetic factors and that the genetic determinants of WT and TEE are largely shared, perhaps due to the action of pleiotropic genes.

54

Missing Phenotype Data Imputation in Pedigree Data Analysis

B. Fridley(1), M. de Andrade(2)
(1) Dept. of Statistics, University of Wisconsin 〉 La Crosse, La Crosse, WI, USA, (2) Div. of Biostatistics, Mayo Clinic, Rochester, MN, USA

Methods to handle missing data have been a statistical area of research for many years. However, little research has been done in the area of missing phenotype information within a pedigree analysis. Recently, Fridley et al. (2003) proposed a data augmentation approach within a Markov chain Monte Carlo for imputing missing phenotype information for polygenic models using family data. The imputation for the missing data takes into account the familial relationships and uses the observed familial information. We propose extending the use of data augmentation as a means to impute values for the missing phenotype information, from which a pedigree data analysis using a major gene model could be fit. Using a polygenic model, data augmentation will produce a set of k complete-augmentation data from which a major gene model can be fit. By producing a set of k complete datasets, the total variance associated with an estimate can be partitioned into a within-imputation and a between-imputation component.

55

Sexual dimorphism and large QTL on chromosome 2p influencing serum uric acid levels in hypertensive families from a relative isolate population

F. Gagnon(1), J. Pintos(2), D. Gaudet(2), J. Tremblay(2), A.W. Cowley(3), P. Hamet(2)
(1) Dept. of Epi, Univ of Ottawa, (2) CHUM, Canada, (3) Medical College of Wisconsin, USA

Epidemiological and experimental data have long suggested a relationship between serum uric acid levels (SUA) and essential hypertension (HT). SUA is ignored in clinical practice because it is believed not to have a causal role in HT. Recently a biological mechanism by which SUA could cause HT in humans was identified. We completed a genome scan for SUA in 93 families (n=1116), from a French-Canadian relative isolate, ascertained through hypertensive dyslipidemic probands. Using covariated-adjusted (age and sex) joint linkage and segregation analysis based on Bayesian Markov chain Monte Carlo methods, we localized a quantitative trait locus (QTL) with

an individual contribution of ~13% of total SUA variance. We estimated a strong sexual dimorphism for SUA with an effect of ~22% of the total variance due to being female; age had an effect of ~4%. Multiple regression analysis supported an interaction of age and sex. The SUA QTL is located on chromosome (chr) 2p at the HT locus (OMIM 607329) identified by Angius et al. (2002). We re-analyzed chr 2 adding an interaction term for age and sex to the covariates, which increased by ~2 fold the linkage signal, with intensity ratio (IR) for linkage of ~13. The IR is the posterior acceptance rate of QTL positions, calculated here for 2 cM intervals, to the expected rate. We assessed the empirical significance for this QTL using the LOP score (log of the ratio of the posterior probabilities of linkage to the chr 2 QTL and ten simulated chr without QTL; Daw et al., 2003): LOP 3.13 with an estimated gene effect on the squared-root variance of ~31 μmol/L. Model-free and variance-component linkage analyses of chr 2 provided additional evidence for this QTL with Lod of 2.54 and 1.52, respectively. In view of the sexual dimorphism observed for SUA, including our report for higher uric acid heritability estimates in females (Am. J. Hum. Genet 76:815–832, 2005), a variance-component analysis modeling genotype-by-sex interaction is under way.

56

Emergence of Novel Genetic Effects on Left Ventricular Mass in Adolescence: Longitudinal Evidence from the Georgia Cardiovascular Twin Study

D. Ge, G.K. Kapuku, F.A. Treiber, H. Snieder
Georgia Prevention Institute, Department of Pediatrics, Medical College of Georgia, USA

A genetic contribution to left ventricular mass (LVM) is generally recognized, but whether and how this influence changes with age is unknown. To answer this question, we studied 472 white and black twin pairs (mean age: 14.7±3.0; 42% black) from the South-eastern US at visit 1 (1996 to 2000), and 431 white and black twin pairs (mean age: 17.6±3.3; 45% black) at follow-up on average 4.1 years later. LVM was measured using echocardiography and was indexed by height$^{2.7}$. Structural equation modeling (Mx software) was used to analyze the data. No significant ethnic and gender effects on heritability of LVM were detected. Apart from body mass index (BMI) as a common predictor for LVM at both visit 1 and 2, resting systolic blood pressure (SBP) and heart rate showed small but significantly independent effects only at visit 2. After the effects of BMI, SBP, and heart rate were removed, the best-fitting multivariate longitudinal model showed that the heritability of LVM was 0.45 (95% CI: 0.35-0.53) and 0.53 (95% CI: 0.43-0.61) at visit 1 and 2, respectively. The genetic correlation between LVM at visit 1 and 2 was only 0.62 (95% CI: 0.46-0.76) indicating the emergence of novel genetic influences in late adolescence, which were responsible for one third of the total LVM variance at visit 2. These findings suggest the emergence of novel genetic influences on LVM during late adolescence and warrant further developmental genetic studies.

57

Evidence for Heritability of Urinary Norepinephrine and Epinephrine Excretion Rates in Young African- and European-Americans

D. Ge, Y. Dong, G.A. Harshfield, J.S. Pollock, F.A. Treiber, H. Snieder
Georgia Prevention Institute, Department of Pediatrics, Medical College of Georgia, USA

Overnight urinary norepinephrine (NE) and epinephrine (EPI) excretion provide aggregate measures of the basal levels of sympathetic nervous system activity. To understand the genetic and environmental contributions to excretion rates of NE ($U_{NE}V$) and EPI ($U_{EPI}V$) and their association, bivariate genetic modeling was performed on data from 39 African-American (mean age: 17.3±2.7) and 44 European-American (mean age: 18.7±3.3) mono- and dizygotic twin pairs from the South-eastern US. Overnight urine collections were assayed by radioimmunoassay for NE and EPI, and the excretion rates were adjusted for body surface area to take body size into account. The ethnic and gender effects on heritability estimates of excretion rates were tested using Mx software and no significant effects were observed. $U_{NE}V$ (mean±SD: 1.9±1.3 μg/hr) and $U_{EPI}V$ (mean±SD: 0.2±0.2 μ956;g/hr) were highly correlated (r=0.81, P <0.001). The best-fitting model showed significant heritabilities for $U_{NE}V$ (0.69, 95% CI: 0.52-0.80) and $U_{EPI}V$ (0.74, 95% CI: 0.59-0.83). The genetic correlation was 0.86 (95% CI: 0.77-0.92) indicating large overlap in the genes influencing $U_{NE}V$ and $U_{EPI}V$. In summary, individual differences in $U_{NE}V$ and $U_{EPI}V$ and their association are substantially heritable and measurements of $U_{NE}V$ and $U_{EPI}V$ provide a viable method for the study of sympathetic tone in genetic epidemiological research.

58

LSMatchMaker: An integrated environment for association analysis

R.S. Gejman(1), J.S. Otto(1), J. Duan(1), J. O–Connell(2), M. Martinez(1,3)
(1) ENH & Feinberg School of Medicine, Northwestern Univ. Evanston, USA, (2) University of Maryland School of Medicine, Baltimore, USA, (3) INSERM EMI0006, France

Large-Scale MatchMaker (LSMatchMaker) is a software package in development to facilitate large-scale association studies (including genome-wide association scans) performed using high-density single nucleotide polymorphisms (SNPs). It integrates data management, quality control routines, association and simulation analyses, and generates graphical outputs of association scan results. It is suitable for case-control and family-based designs. Options for using different approaches for estimating pairwise LD between SNPs, haplotype frequencies and for testing association are being implemented (e.g., UNPHASE package, FBAT, TRANSMIT, EH). Furthermore, for case-control designs, LSMatchMaker derives pointwise, genome-wide and gene-wide empirical significance thresholds using permutation-based simulation. LSMatchMaker is written in Java and can be run natively under most operating systems. It will be available free of charge for non-commercial research institutions.

59

Dissection Of Complex Traits: Advantages Of Multivariate Phenotypes Over End-point Binary Traits

S. Ghosh, S. Bhattacharjee

A complex trait is usually a function of a multivariate phenotype comprising correlated quantitative variables. Since end-point traits are usually binary in nature (affected/unaffected) and hence contain minimal information on variation within trait genotypes, it may be statistically more powerful to use a correlated multivariate phenotype for identifying genes for the complex trait. Mapping a multivariate phenotype traditionally uses some function of quantitative values of sib-pairs or other sets of relatives as a response variable and marker IBD scores as explanatory variables. In these analyses, linkage inferences depend strongly on the assumed probability distributions of the quantitative variables, particularly for variance components approaches. We propose, along the lines of Sham et al. (2002), a linear regression formulation in which the response and explanatory variables are interchanged. Analyses do not require modeling the covariance structure of the multivariate phenotype vector or any data reduction technique such as principal components. It can simultaneously incorporate qualitative and quantitative traits and can use data on n siblings as (n−1) independent observations. Using simulations under different correlation structures and probability distributions of a multivariate phenotype and an associated binary trait, we find that the proposed method is more powerful than the Haseman-Elston regression and the reverse regression procedures based on (i) the first principal component of the correlated phenotypes and (ii) the end-point binary trait. An application of the method is illustrated using data on alcoholism related phenotypes from the COGA study, each of which has provided evidence of linkage on Chromosome 4 using univariate analyses.

60

On The Value Of Molecular Haplotypes In The Context Of A Family-Based Linkage Study

E. Gillanders(1), J. Pearson(2), A. Sorant(1), J.Trent(2), J. O–Connell(3), J. Bailey-Wilson(1)
(1) IDRB, NHGRI NIH, Bethesda MD, (2) TGen, Phoenix AZ, (3) University of Maryland, Baltimore MD

In the following simulation study we aimed to investigate the value of molecular haplotypes in the context of a family-based linkage study. In extended pedigrees we simulated a qualitative trait, and highly polymorphic microsatellite (STR) marker data. We then compared the relative power of analyzing these data assuming that various levels of experimentally derived haplotypes were available. Several conclusions can be drawn for an extended pedigree, fine-mapping linkage study of STR markers. When genetic homogeneity is expected or marker data is complete it is not efficient to generate molecular haplotyping information. However, with levels of heterogeneity and missing data patterns typical of complex traits there is at least a 12% increase in the power to detect HLODs greater than 3.0 when affected individuals are molecularly haplotyped. When four markers are included in multipoint analyses, power increases from 63% when only genotype information is available to nearly 75% when directly measured haplotypes are included for affected individuals. The power when everyone in the pedigree is haplotyped is only slightly higher. Additional simulations will examine the value of molecular haplotyping information in the context of: 1) a genome wide density linkage study of STR markers; 2) a SNP based genome wide density linkage study; 3) a linkage study using less informative family structures (i.e. nuclear families or sib-pairs); and 4) only a single affected individual being haplotyped per pedigree.

61

Personal and family history of autoimmune conditions and the risk of Hodgkin lymphoma

L.R. Goldin, O. Landgren, E.A. Engels, R.M. Pfeiffer
Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD

Abnormalities of the immune system (immunosuppression and autoimmunity) are known risk factors for developing lymphoma. Several studies have found that individuals with systemic autoimmune (AI) diseases such as rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE) are at increased risk for developing non-Hodgkin lymphoma. Hodgkin lymphoma (HL) has a strong familial component and may also be associated with autoimmunity. We tested whether personal and/or family history of autoimmune diseases is associated with an increased risk of HL using population-based linked registry data in Sweden and Denmark. Thirty-five separate autoimmune conditions in 7,476 HL cases, 18,573 matched controls, and 86,163 first degree relatives of cases and controls were scored from hospital discharge diagnoses. First, we assessed the association of each condition with HL in separate logistic regression models. We then also fit a hierarchical regression model that incorporated information on the 35 autoimmune conditions by classifying them into 3 broad categories (those with detectable autoantibodies with/without systemic involvement, and those without detectable autoantibodies) that were assumed to share overall group mean effects. Variations within an AI category were modeled by random effects. For personal history of systemic AI conditions, strong associations were found for RA, collagenosis, Sjogren–s syndrome, and SLE and for immune thrombocytopenic purpura and sarcoidosis using the conventional and hierarchical models. Family history of AI conditions was generally not associated with risk for HL but there was a marginally elevated risk for family history of sarcoidosis and ulcerative colitis. We conclude that certain defined autoimmune conditions are important risk factors for HL but genetic factors predisposing to HL are likely to be different from those predisposing to autoimmunity.

62

Association of MC1R variants and risk of melanoma (MM) in MM-prone families with CDKN2A mutations

A.M. Goldstein(1), M.T. Landi(1), S. Tsang(2), M.C. Fraser(1), D.J. Munroe(2), M.A. Tucker(1)

(1) Genetic Epidemiol Br, DCEG, NCI, NIH, DHHS, USA, (2) Lab Molecular Technol, NCI-Frederick, SAIC-Frederick, NIH, DHHS, USA

Major risk factors for MM include many nevi, especially dysplastic nevi, fair pigmentation, freckling, poor tanning ability, and germline mutations in the CDKN2A, CDK4 or MC1R genes. We evaluated the relationship between MC1R and MM risk in 395 subjects from 16 American MM-prone CDKN2A families with extensive clinical and epidemiologic data. MM risk factors were assessed by clinical exam or questionnaire; MC1R was sequenced. Odds ratios were estimated by logistic regression. We examined the distribution of MC1R variants and median ages at MM diagnosis in multiple primary melanoma (MPM) and single primary melanoma (SPM) patients. Presence of multiple MC1R variants was significantly associated with MM, even after adjustment for major MM risk factors. All 40 MPM patients had at least one MC1R variant; 65% of MPM patients versus only 17% of SPM patients had >=2 MC1R variants ($p<0.0001$). For all 69 MM patients combined as well as the 40 MPM patients, there was a significant decrease in median age at diagnosis as numbers of MC1R variants increased ($p=0.010$ and $p=0.008$, respectively). In contrast, no significant reduction in age at diagnosis was observed for SPM patients ($p=0.91$). The current study suggests that the presence of multiple MC1R variants is associated with the development of multiple MM tumors in patients with CDKN2A mutations. Additional studies are needed to confirm these findings and to explore the mechanisms that may contribute to this relationship.

63

Transforming Growth Factor-β_1 Leu10Pro variant and Breast Cancer Risk: A Case-Control study and Meta-Analysis

AM Gonzàlez-Zuloeta Ladd(1), A Arias Vàsquez(1), J Witteman(1), A Uiterlinden(2), JW Coebergh(1), A Hofman(1), BHCh Stricker(1), CM van Duijn(1)
(1) Epidemiology & Biostatistics Department, (2) Deparment of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands

TGF-β_1 has a dual role as a tumor suppressor in early stages and a tumor promoter in later stages of carcinogenesis. A Leu10Pro substitution in its gene leads to higher circulating levels of TGF-β_1. This variant has been studied in relationship to the risk for breast cancer with contradicting results. We aim to unravel this through an association study and meta-analysis. Women participating in the Rotterdam Study (N=3590) including 163 breast cancer patients were genotyped for this polymorphism. We carried out a logistic regression and then a survival analysis; finally, we performed a meta-analysis of 6 studies including 6282 patients and 12738 controls using a random effect model to compute pooled odd ratios. The logistic regression showed an increased risk for Pro allele carriers (OR=1.5;95% CI=1.1-2.2) against non-carriers, this was maintained among incident cases (OR=1.8;95% CI=1.1-2.8). No differences in risk among prevalent cases were seen. The survival analysis showed that Pro allele carriers had a HR of 1.8 (95% CI=1.2-2.7) compared to non-carriers. The meta-analysis showed the same increased risk of breast cancer for Pro allele carriers versus non-carriers (OR pooled =1.5, 95% CI=1.15-1.96, heterogeneity=91%). This was consistent in different ethnic groups. The findings in our population based study and meta-analysis suggest that the TGF-β_1 Leu10Pro polymorphism plays a role in breast cancer risk.

64

Regression-Based QTL analysis method incorporating parent-of-origin effect

O.Y. Gorlova, L. Lei, Y. Zhang, D. Zhu, S. Shete, W.D. Li, R.A. Price, C.I. Amos

We present an extension of Sham et al.–s (2002) regression-based quantitative-trait linkage analysis method to incorprate parent-of>origin effects. We separately regressed total, paternal, and maternal IBD sharing on traits–squared sums and differences. We also developed a test for imprinting that indicates whether there is any difference between paternal and maternal regression. We use a panel of statistics to detect imprinting, which includes an overall T statistic (a test for total linkage), both parental T statistics (tests for parental-specific linkages), and the D index (a test for imprinting). We performed an extensive simulation to examine the performance of the panel. We found that when using empirical percentiles the method is very powerful in detecting parent-specific linkage with correct type I error rate for the non-linked parental component. Missing parental genotypes increase the type I error rate of both the linkage and imprinting tests and decreases the power of the imprinting test. When the major gene has low heritability, the power of the method decreases dramatically but the panel still performs well. We also used a permutation algorithm, which can ensure the appropriate type I error rate. We applied the method to a data from a study of 6 body size related measures and 23 loci on chromosome 7 for 255 nuclear families. Multivariate identities-by-descent were obtained using a modification of SIMWALK program. A parent-of-origin effect consistent with maternal imprinting was suggested at 99.67–111.26 Mb for body mass index, bioelectrical impedance analysis, waist circumference, and leptin concentration.

65

Evaluation of Alternative Sequential Updating Procedures for Computing the Posterior Probability of Linkage (PPL) Across Clinically Defined Data Subsets

M. Govil, V.J. Vieland
Ctr. for Stat Genet Research, Univ. of Iowa, IA

The PPL, an approach to linkage analysis designed for complex traits, uses Bayesian sequential updating to accumulate evidence for or against linkage across hetero-

geneous datasets and/or clinically or epidemiologically based data subsets. The PPL converges to 1 (under linkage) and to 0 (under no linkage) with increasing sample size, regardless of how the updating is done. However, the optimal way to define subsets has yet to be systematically examined. We simulated data based on pedigrees/phenotypes from an ongoing study of cleft lip/palate, with families divided into small (2–3 generations, N=115) and large (>=4 generations, N=104) sets. Genotypes were simulated under a single-locus model (disease allele frequency=1%; penetrances=0.4, 0,0), and 100 replicates were generated under different levels of heterogeneity (1000 under ''no linkage''). Within each replicate, updating was performed across subsets defined randomly (R), based on true (T) status (linked/unlinked), by individual pedigrees (PED), or without any subsetting (NS). Under linkage, T yields larger PPLs compared to NS, R or PED; while under no linkage, R and PED yield PPLs close to NS, although NS (the ''true'' group under no linkage) yields stronger evidence against linkage. While R is neither helpful nor harmful, PED is less effective than NS. These results are consistent across small and large pedigrees. While T is not possible to achieve for real data, its superior performance underscores the utility of finding variables that can help delineate homogeneous data subsets.

66

Family-based association method for binary traits

C Gray-McGuire, RC Elston
Dept of Epi & Biostat, Case Western Reserve Univ, USA

Methods to estimate major gene, environmental, and polygenic effects are increasingly important in characterizing multifactorial disease.
We present here such a method; an extension to binary traits of the Elston-Stewart algorithm for assessing association to any number of covariates, including marker or candidate gene alleles, in the presence of residual familial correlations. This extension models a binary trait by logistic regression, conditional on the assumption of a multivariate normally distributed latent trait variable or ''liability'', in both random and ascertained samples. Our method allows for calculation of heritability and familial correlations of the liability through estimated polygenic, marital, sibling, and familial variance components.
We offer an evaluation of this method, through simulation, for nuclear and extended families, as well as several combinations of residual familial effects and sampling schemes. We establish that the true type I error rates are, on average, less than the nominal and, for the chosen sample sizes, power is greater than 90% in several circumstances. We show that the accuracy of the association parameter is not affected by family structure, sampling scheme or analysis model, but that the accuracy of the variance component estimates are highly dependent on the sampling scheme, and, generally, true population values are best estimated from a random sample but are reasonable for those samples corrected for ascertainment. Finally, we give an example of the successful application of this method to a real data set, identifying strong association and likely locus-specific mode of inheritance.

67

Local Score statistic: application to large-scale association studies

M.Guedj(1,2), G. Nuel(2)
(1) Serono Genetics Institute, (2) Statistic and Genome Laboratory

Genetic epidemiology aims at identifying biological mechanisms responsible for human diseases. When dealing with complex traits, classical linkage analyses are facing their limitations. Genome-wide association studies, made possible by recent improvements in genotyping, are now promisingly investigated.
In these studies, commonly used strategies focus on the marginal effect of markers on disease. Such approaches lead to multiple-testing and are unable to capture the possibly complex interplay between genetic factors. Association analyses need methodologies able to manage dimensionality and complexity to success. Hoh et al. (2001) have suggested using sums statistics to combine single-marker genetic marginal effects. Without assuming any interaction pattern this method selects a set of interesting markers for further analyses with rigid control of the genome wide significance level and presents encouraging results.
As part of this sums statistics family methods, we have adapted the local score statistic, already frequently used to analyse long DNA or protein sequences. A score is assigned to each marker in order to obtain a sequence of successive scores; then our strategy identifies subsequences presenting unexpected high cumulative scores, assuming that such features may be biologicaly relevant. Via sums statistics, this method captures local dependence between markers and combines it with possibly spaced information. Fast to compute, such an approach is dedicated to handle large-scale case-control data, and reduces at the same time the multiple-testing problem.
Hoh et al., 2001. Genome Research. 11: 2115–2119

68

Combined Haplotype Relative Risk (CHRR): A simplified genetic Association Test combining Triads and unrelated Subjects

C.Y. Guo(1,3), K.L. Lunetta(2), A.L. DeStefano(2), L.A. Cupples(2)
(1) Department of Mathematics and Statistics, Boston University, USA, (2) Department of Biostatistics, Boston University School of Public Health, USA, (3) National Heart, Lung and Blood Institute–s Framingham Heart Study, USA

In some genetic association studies, samples contain both parental and unrelated controls. Under such scenario, instead of analyzing only trios using family-based association tests or unrelated subjects using a case-control study

design, Epstein et al. (2005) proposed a method that implemented a likelihood ratio test to combine the two different types of data. Here, we put forward a new strategy of combining such data based on the Haplotype Relative Risk (HRR) proposed by Falk et al. (1987). The HRR compares parental marker alleles transmitted to an affected child to those not transmitted as a test for association, a strategy which is similar to the case-control study that compares allele frequencies in diseased cases to healthy controls. Thus, affected children can be pooled with diseased cases and the parental controls can be treated as the healthy controls if the affected children and diseased cases were both randomly selected from the same population. Therefore, unrelated subjects can be incorporated into the HRR intuitively and effortlessly. For trios without complete parental genotypes, we adopted the strategy proposed by Guo et al. (2005), which is an easier approach than the one proposed by Weinberg (1999).

69

Testing informative Missingness in family-based Linkage and/or Association Study

C.Y. Guo(1,3), L.A. Cupples(2), Q. Yang(2)
(1) Department of Mathematics and Statistics, Boston University, USA, (2) Department of Biostatistics, Boston University School of Public Health, USA, (3) National Heart, Lung and Blood Institute–s Framingham Heart Study, USA

In many family-based linkage and/or association studies, extended pedigrees were collected and analyzed. Because all pedigrees have different family structures, it is difficult to determine whether the missing genotype data are informative or not. Although most current methods assume that parental genotypes are missing at random, such as the widely used Family-Based Association Test (FBAT) by Rabinowitz and Laird (2000), this property is critical for transmission disequilibrium type procedures to assure an unbiased conclusion. To date, there is no test that examines the missing data pattern. In this research, we proposed a test based on randomly selected triads (each has exactly one affected child and both parents) from independent extended pedigrees. Derived from the distributions of ascertained triads (Table 1, Guo et al., 2005), we proved that a chi-square test with 2 degrees of freedom can be implemented to test the missing data pattern of parental genotypes. To account for potential confounding factors, a logistic regression approach was further proposed to compare the distributions of parental genotypes between complete and incomplete triads.

70

Heritability estimates for serologically assessed infections with common pathogens in Alaskan Eskimo participants in the GOCACAN Study

HHH Göring(1), J Zhu(2), S Laston(1), SA Cole(1), AG Comuzzie(1), Sven OE Ebbesson(2), BV Howard(2), JW MacCluer(1), M Davidson(2)
(1) Dept. of Genetics, SFBR, San Antonio, TX, USA, (2) MedStar Research Institute, Hyattsville, MD, USA

Chronic infections may be important contributors to risk of common diseases of old age such as atherosclerosis. The goal of this investigation is to determine whether genes play a role in resistance or susceptibility to infections with 5 common pathogens—*Chlamydia pneumoniae* (*Cp*), *Helicobacter pylori* (*Hp*), cytomegalovirus (CMV), and herpes simplex viruses 1 (HSV1) and 2 (HSV2). The infection status of ~500 Alaskan Eskimo family members was assessed serologically (mostly using commercial ELISA kits) at 2 time points ~15 years apart. The pathogens were found to be common, with seropositivity rates for *Hp*, CMV and HSV1 >75% in adults, and rising with age to ~60% for *Cp* and ~40% for HSV2. Seroreversion rates were between <1%–10% over ~15 years, suggesting that these infections are chronic and/or that the immune system has good long-term memory. The quantitative antibody titer and the discrete serostatus were found to be significantly heritable for *Cp*, *Hp*, CMV and HSV1 (with most variance components-based heritability estimates >20%) at both clinic visits.
However, no evidence for genetic regulation of infection risk with HSV2 was observed. We speculate that the differences in heritabilities are perhaps due to different means of transmission (sexual for HSV2 and oral/respiratory for the others) and the associated different exposure risks (limited for HSV2 and probably near universal for the others).

71

Analysis of Admixture in Puerto Rican women and association of Admixture with phenotypes

I. Halder(1), T. Frudakis(2), J. Fernandez(3), M.D. Shriver(1)
1) Pennsylvania State Univ., PA, 2) DNAPrint Genomics, FL, 3) Univ. of Alabama, Birmingham, AL

We have analyzed individual Admixture (ADM) in a sample of 64 Puerto Rican women from New York City using a panel of 177 highly informative Ancestry Informative Markers (AIMs). A previous analysis of the same individuals using 35 AIMs (Bonilla et al. Hum Genet. 2004;115:57–68) demonstrated significant association of ADM with skin pigmentation and Bone Mineral Density (BMD). The present marker panel was chosen to increase marker information content for ancestry and balance information between all ancestry axes. Using the present marker panel the estimated mean European (Eu) and African (Af) ADM was similar to those obtained previously however proportional Indigenous American (IA) ADM was significantly reduced in the sample. Tests for stratification confirmed genetic structure within the sample. Significant associations were observed between skin pigmentation and all three ancestry axes. Previous studies had not shown any association between IA ADM with skin pigmentation. Associations with BMD were not observed, instead suggestive association between Fat free mass and both Af and IA ADM was seen. Theoretically,

admixture estimates are more precise with increased number of markers. This study demonstrates that while smaller marker panels may provide useful estimates of ADM, larger and preferably balanced panels are required to estimate ADM reliably. Studies using small, unbalanced marker panels may detect associations caused by inflated variance in estimated ADM as an artifact of the information content of marker panels.

72

jPAP: Document-Driven Software for Genetic Analysis

S.J. Hasstedt
Department of Human Genetics, University of Utah, Salt Lake City, UT, USA

jPAP (Pedigree Analysis Package for Java) is a cross-platform software package for computing likelihoods or simulating phenotypes on pedigree data using genetic models. The software's analytical power is derived from its Fortran predecessor PAP, while its interface has been wholly recast along document-based lines, a design familiar from applications such as word processors and spreadsheets. jPAP includes a sophisticated graphical user interface (GUI), with editors for pedigree structure, phenotypic observations, variable definition, and genetic models. The editors present with common elements and behavior, and are tightly integrated. The advantages of a document's ''all-at-once'' view, and corresponding ability to dynamically track dependencies, are most evident in the genetic model editor, which makes more manageable the great variety of models that PAP supports.

While the user can enter and save all the information required for an analysis directly through the GUI, jPAP also allows import from sources such as the popular ''linkage'' format as well as legacy PAP formats. All of PAP's analysis options remain supported, including segregation analysis, variance components analysis, transmission disequilibrium testing, genetic model fitting, power analysis, and p-value estimation. Following the planned incorporation of Markov Chain/Monte Carlo methods, jPAP will also be capable of performing multipoint linkage analysis using either major locus or variance component models.

For more information on jPAP see http://hasstedt.genetics.utah.edu/.

73

Quantification and characterization of genotyping error in a large epidemiological sample with repeated measurements

IM Heid(1,2), C Lamina(1), G Fischer(1), C Gieger(1), C Huth(1), N Klopp(1), M Kolz(1), C Vollmert(1), S Wagner(1), B Paulweber(3), SC Hunt(4), A Peters(1), HE Wichmann(1,2), H Küchenhoff(2), F Kronenberg(1,5), T Illig(1)
(1) GSF-Inst. of Epidemiology, Germany, (2) Univ. of Munich, Germany, (3) Private Medical Univ. Salzburg, Austria, (4) University of Utah, USA, (5) Innsbruck Medical Univ., Austria

Genotyping error is often discussed as one source of bias and of inflated type II error in association studies, but precise estimates on error size and error model in an epidemiological setting are still lacking.

We have thus collected the data of all studies involving more than 1000 subjects each, which were genotyped at our institute since 2004 on a MALDI TOF-MS system (Sequenom) with at least 5% duplicates. Altogether 298 430 genotypes with 118 943 duplicates have been analyzed involving 230 SNPs and 11745 subjects. Among the 85446 pairs with both duplicates present, 664 were discordant (0.8%). Only 7 pairs showed discordance due to one homozygous wildtype and one homozygous for the minor allele. Interestingly, the missingness of either of the measures did not depend on the genotype of the other. We also considered call quality and the reason for duplication (routine duplicate on the same plate or second typing of full plate).

Given the observed genotype-specific discordance rates, we estimated the transition probabilities, pij=Prob(observed genotype=i | true genotype=j), i,j=0,1,2,3, as well as the true genotype frequencies using a likelihood based approach. Furthermore, we evaluate the impact on association estimates in a recent study.

74

Optimal choice of covariates in affected relative-pair linkage analysis

P. Holmans
Cardiff University

Studies of genetic linkage in complex traits are complicated by problems such as genetic heterogeneity, variable phenotype definition and gene-environment interaction. These reduce the power of standard linkage analyses, and make it difficult to replicate positive findings. A promising method for increasing the power of linkage analyses in the presence of potential confounding factors is to include the confounders as covariates in affected relative pair linkage analyses. However, this requires a single covariate to be defined for each relative pair, and the best way to obtain this from the individuals' quantitative trait values is unclear. This question was investigated by simulating a binary trait and a related quantitative trait in a sample of affected sib pairs using a pleiotropy model. A number of pairwise covariates were calculated from the individuals' quantitative trait values (sum, difference, cross-product) and included in linkage analyses of the binary trait using the logistic regression model of Rice (1997 Am J Med Genet 74:112–114). The power gained by including covariates is greatest when the binary trait model includes phenocopies, since the quantitative trait may help distinguish these. In general, including both the pairwise sum and difference of quantitative trait values gave the greatest power. The effect of dichotomising the quantitative trait was also investigated. This increased power when the heritability of the quantitative trait was high, its distribution followed a bimodal pattern, and its mode of inheritance (dominant or recessive) was the same as the binary trait.

75

Mitochondrial DNA mutation in noise-induced hearing loss

Y.S. Hong, M.J. Lee, J.Y. Kwak, Y.H. Lee, J.Y. Kim
(1) Department of Preventive Medicine, College of Medicine, Dong-A University, Medical Research Center for Cancer Molecular Therapy and The Research Society of Environmental Genetic Epidemiology, Korea, (2) Department of Public Health, Kobe University School of Medicine, Japan, (3) Medical Research Center for Cancer Molecular Therapy, College of Medicine, Dong-A University, Korea, (4) Department of Preventive Medicine, Kosin Medical College, Korea

Mitochondrial DNA mutations have been reported in recent years in association with sensorineural hearing loss. The purpose of this study is to identify the association between the noise-induced sensorineural hearing loss and the A to G mutation at nucleotide 3243, 1555, 7445 of mitochondrial DNA. Study subjects were established by history and chart review, and audiological and clinical data were obtained. Blood was sampled from 354 normal controls, 252 noise-induced hearing loss, and 58 sensorineural hearing loss. The DNA of these individuals were extracted, and mitochondrial DNA fragments were analyzed by polymerase chain reaction. Subsequently, the coding sequence of mitochondrial DNA 3243, 1555, 7445 were sequenced, and compared to the normal sequence, and all sequence variations were analyzed by restriction enzymes. Mitochondrial DNA mutations(3243, 1555, 7445) were not detected by polymerase chain reactions in any patients with noise-induced hearing loss, sensorineural hearing loss, and normal controls. The DNA sequencing of PCR products did not revealed an A to G substitution at nucleotide 3243,1555,7445 of mitochondrial DNA.

76

Powerful statistics for testing the null hypothesis of no genetic association in case-control studies

J.J. Houwing-Duistermaat, R. el Galta
Dept. of Medical Stat. and Bioinf., LUMC, Leiden, The Netherlands

In association studies, the data can be summarized in 2 by m tables with m the number of observed haplotypes. If one haplotype carries a mutation, this haplotype is over represented in the cases. For $m > 4$, the classical Pearson–s chi-square has low power. Two approaches can be followed to obtain powerful statistics, namely taking the maximum and averaging over the possible associated haplotypes.
For the maximum approach, we consider Zmax defined as the maximum of chi-square statistics of all 2 by 2 tables of one haplotype versus the rest and Zclump defined as the maximum of chi-square statistics of all 2 by 2 tables of combinations of haplotypes versus the rest. For the average approach, the likelihood is written as a weighted sum of conditional likelihoods given that a haplotype is associated. Here, the weights are the prior probabilities. Terwilliger (1995, Am J Hum Genet 56, 777–87) proposed to use allele frequencies. As alternative, we consider flat weights, i.e. 1/m. Two novel score statistics are proposed as test statistics.
We compared heuristically and empirically the performance of the 5 statistics under the null and various alternative hypotheses. Further we applied the tests to an association study on a disease and three candidate genes described by 5, 5 and 3 common haplotypes. The score statistics and Zmax gave significant p-values for all genes while the chi-square was only significant for the gene with 3 haplotypes. We conclude that Zmax has high power if only one haplotype is associated while the score statistics provide a good average power.

77

An Improved TREE Algorithm for Dissecting Complex Interacting Genetic Pathways

S. Howell, J.Wu, M.A. Province
Washington Univ Med School, St. Louis, MO, USA

Identifying genomic regions responsible for a complex trait is difficult, due to the intricacy of the causative pathways. Traits such as atherosclerosis, or inflammation have complex etiologies involving multiple genes, GxG and/or GxE interactions which are poorly understood. Recently, Province et al. (2001) proposed recursive partitioning (RP) techniques to help disentangle such complexity. TREE linkage/association separates a sample into increasingly similar subgroups, each with a simpler etiology, by automatically splitting the data according to combinations of measured covariates (other genes, exposures, etc.). But the original TREE algorithm of Shannon et al. (2001) used the traditional MSE as the improvement metric for splitting. Applied to Haseman-Elston (H-E) linkage, this is not specific enough. False-positive groups may be defined this way, since subgroup differences in the intercepts only of a H-E regression (but equal slopes) improve the MSE, but do not constitute different linkage groups. We present an improved TREE algorithm, which specifically focuses on subgroup slope differences to define RP splits. Simulations show that both algorithms have much greater power to detect genes when there are complex interactions, but the original too often identifies the wrong subgroups. The original can also behave poorly under the null when there are subgroup mean differences in the H-E phenotypic similarity, whereas the improved protects us. Additionally, the improved algorithm has better median sensitivity, specificity, impurity and kappa (1.00, 1.00, 0.00, 0.88) than the original (0.94, 0.94, 0.07, 0.84).

78

Linkage analysis in the presence of genotyping errors

Q. Huang and C.I. Amos
Johnson and Johnson, New Brunswick, NJ and U.T. M.D. Anderson Cancer Center, Houston, TX

It is well known that genotyping errors can have an enormous impact on linkage analysis. Such effects include

inflation of distance for multiple markers, decrease in power, and inaccuracy of the disease gene localization. Several statistical methods have been proposed to detect and remove genotyping errors to recover the linkage information. However, error detection rates can be poor, especially in the sib-pair design when parents are not available for genotyping and for less informative markers such as SNPs. We assessed the performance in a sib-pair design using tightly linked SNPs of a version of Merlin that uses maximum likelihood to jointly assess linkage while allowing for genotyping errors. We simulated 100 nuclear families comprising affected sib-pairs under different scenarios: both parents are available for genotyping, one of the parents missing, or both parents missing. We simulated 8 SNP markers with equal allele frequencies and an inter-marker distance of 0.5 cM with the disease locus between markers 4 and 5. We assumed a recessive disease model with disease allele frequency of 0.01 and used a nonparametric IBD-sharing analysis. When both parents were available for genotyping, first identifying and removing Mendelian errors recaptured most of the linkage information. However, when one or both parents were not available, using an error model in the linkage analysis greatly increased LOD scores. For sib-pair data, the maximum LOD score can be increased by as much as 20% for 0.01 error rate and by as much as 90% for 0.03 error rate. When there is no error, LOD scores remain the same.

79

A cluster-based SNP linkage mapping set based on genetic distances and on haplotype heterozygosity

F.C.L. Hyland, J. Ziegle, J. Day, C. Scafe, R. Koehler, N. Peyret, C. Larry, M. Rhodes, T. Woodage, X. You, L. Xu, E. Spier, F.M. De La Vega

A SNP linkage mapping set that combines the heterozygosity of microsatellite loci and the efficiency, accuracy, cost advantages and automation potential of SNP genotyping is described. To enable high-throughput and cost-effective genotyping, the platform for this human linkage mapping set is the SNPlex™ Genotyping System, a multiplexed high throughput genotyping platform based on the oligonucleotide ligation assay that utilizes capillary electrophoresis as the readout. The set is composed of SNP clusters, with spacing between clusters based on genetic rather than physical distance. A 3.9 cM resolution SNP map developed by The SNP Consortium (TSC) was used as the framework. First, the TSC markers with the most robust SNPlex System assays were chosen. Next, validated SNPs from the TaqMan® SNP genotyping assays were added to increase cluster density. To select new candidate SNPs, NCBI physical coordinates were transformed to genetic distances using the TSC genetic map. The transform was refined using the metric linkage disequilibrium map (expressed in LD Units). 10 candidate SNPs were selected for each cluster. Haplotypes were computed for every possible combination of 4 of these 10 SNPs, and the 4 SNPs that maximized haplotype heterozygosity across four populations (Caucasian, African American, Japanese, Chinese) were chosen and extensively validated to select only highly robust SNPs. A genetic map was calculated using genotypes of 574 samples from 43 CEPH families. Genetic distances proved to be highly correlated with LDU: the average correlation between LDU and observed cM was 0.99. The linkage mapping set is composed of clusters spaced 1.9 cM apart, with the mean distance between SNPs being 1.04 cM, and no gaps larger than 10 cM. The average information content is 0.95.

80

Patterns of linkage disequilibrium are conserved within and between ethnic groups: sampling vs. ancestry in the HapMap project

F.C.L. Hyland, C.R. Scafe, H. Isaac, F.M. De La Vega

The recent release of the HapMap data raises important questions about the extent to which similarities and differences observed in the patterns of sequence variation between populations are an effect of sampling rather than due to intrinsic differences; and whether results from the Yoruban African population are predictive for African Americans. We computed metric linkage disequilibrium (LD) maps (expressed in LD units, or LDUs) for 104824 SNPs on chromosomes 6, 21 and 22 in four HapMap populations. We compared these to 28155 SNPs previously genotyped using TaqMan® SNP Genotyping Assays from comparable CEPH, Chinese and Japanese populations, and from African Americans. 12206 SNPs were in common. The correlation of minor allele frequency (MAF) between HapMap and AB samples within a population is much higher than correlation between populations. In contrast, LDU patterns are highly consistent between and within populations. The mean correlation between LDU across populations is 0.9993 (SD 0.0006). The mean correlation of LDU within a population is 0.9992. This is consistent with the hypothesis that LD is largely driven by recombination rate; hence the fixed locations of 'steps' and 'plateaux' in the LD map. Correlation between LD across populations is 10-fold higher than the correlation between LDU and physical locations (mean 0.9935), and highly significant. The correlation between MAF in African Americans and Yorubans is similar to that within a population. The length and LDU positions of the African American and Yoruban LD maps are consistent. These results confirm that conclusions on MAF or LD between populations observed in the HapMap data reflect underlying patterns rather than sampling effects, and that the patterns of the LD maps are consistent across populations.

81

A multilocus association analysis method based on projection pursuit discriminant analysis

R. Jiang, J. Dong, S. Zhang, Q. Sha
Department of Mathematical Sciences, Michigan Technological University, USA

We propose an association analysis method based on projection pursuit discriminant analysis. The projection

pursuit is a dimension reduction method in statistics. Our goal is to localize disease susceptibility loci using case control data. We mainly focus on interactions of several susceptibility loci. For a complex disease, it is usually assumed that there are several disease genes, and marginal effect of each locus is small. If each locus is tested individually, it will be very hard to find the susceptibility loci. On the other hand, if one tries to test several loci jointly, the computational burden will soon exceed the capacity of computers. We propose a method of association mapping, which is based on projection pursuit discriminant analysis. The method can detect the interaction of several disease susceptibility loci with small marginal effects, and the computational burden of handling several hundreds of loci is within the capacity of a personal computer. We simulate six disease models with number of markers ranges from 20 to 500, and the preliminary results show that the power of tests based on this method is high.

82

Coverage and Power in Genome-wide Association Studies

E. Jorgenson, J. Witte

Large sets of SNPs are becoming available for genome-wide association studies. The extent to which these SNP sets provide information about unobserved SNPs (coverage) is currently unknown. We can use information from resequencing efforts focused on specific regions of the genome and groups of genes as a gold standard for determining the coverage that large genome-wide SNP sets provide. Since less than complete linkage disequilibrium between the causal locus and observed SNPs will result in a reduction in power, quantifying sequence coverage will provide a metric for the reduction in power to be expected when using a particular SNP set to perform genome-wide association studies. We compare average and cumulative distribution measures of sequence coverage, using real SNP data in multiple ethnic groups, to determine the reduction in power. Measures based on the cumulative distribution reveal that power is 8–17% lower than expected under commonly used average metrics. This has important implications for genome-wide association studies, as average metrics, including the average maximum r^2, can inflate power estimates and lead to a higher than expected false negative rate.

83

A complex genetic model for a locus on 1p36 for cortical cataracts in the Beaver Dam Eye Study

G. Jun(1), B.E.K. Klein(3), R. Klein(3), K.E. Lee(3), S.K. Iyengar(1,2)
(1) Dept. of Epi & Biostat, Case Western Reserve Univ., USA, (2) Dept. of Ophthal, Case Western Reserve Univ., USA, (3) Dept. of Ophthal & Vis Sci, Univ. of Wisconsin, USA

We conducted a genome-wide scan (GWS) with 2252 individuals (1009 sibpairs) from 487 pedigrees in the Beaver Dam Eye Study (BDES). We performed a linkage analysis using a quantitative trait for cortical cataract adjusted for covariates, both untransformed and transformed. For the most significant regions in the GWS, we tested heterogeneity by removing linked sib pairs. We assessed gene-gene and gene-environment interactions for the markers with the nominal $p<0.01$. The linkage signals on chromosome 1 were 1p36 (D1S3669; $p=8\times10^{-4}$), 1p21 (D1S1631; $p=7\times10^{-5}$), 1q31 (D1S1660; $p=1\times10^{-4}$), and 1q41 (D21S2141; p=0.009) in the GWS. Removal of 117 linked pairs on 1p36 revealed heterogeneity on 1q31 (D1S1677; before removal (BR):0.15, after removal (AR):9×10^{-4}) and on 1q41 (D1S2141; BR:0.0088, AR:5×10^{-7}). Linkage signals best explained by smoking were on 1p36 (main effect (ME):0.17, interaction (I):1×10^{-4}) and 1p21 (ME:0.33, I:0.008). Significant gene-gene interactions with the marker D1S3669 were observed on 2q11 (ME:0.003, I:6×10^{-4}), 3q13 (ME:0.99, I:7×10^{-7}), 6q16 (ME:0.19, I:6×10^{-5}), 8q13 (ME:0.065, I:4×10^{-4}), 14q22 (ME:0.023, I:3×10^{-6}), 14q14 (ME:0.003, I:5×10^{-5}), and 21q22 (ME:0.65; I:3×10^{-4}). We recognize that multiple testing is an issue and permutation tests are being conducted. The 1p36 region for cortical cataract is involved in linkage heterogeneity, as well as in gene-gene and gene-environment interaction.

84

Development and Validation of a Risk Prediction model for Pancreatic Cancer: PANCPRO

Klein AP(1,2,3), Wang W(4), Chen, S(4,5) Parmigiani G(1,2,4)
(1) Oncology, (2) Pathology, (3) Epidemiology, (4) Biostatistics, (5) Environmental Health Sciences, Johns Hopkins University

Pancreatic cancer is the 4th leading cause of cancer death in the US due to its rapid progression and profound resistance to treatment. Familial aggregation of pancreatic cancer is well documented and segregation models support major gene inheritance. Given the genetic factors responsible for the majority of the familial aggregation have yet to be identified; the development of methods to predict susceptibility is a critical step to identify individuals who may benefit from the screening modalities currently under investigation. Therefore, we developed PANCPRO, a Mendelian risk prediction model for pancreatic cancer. Mendelian approaches use prior estimates of the frequency and penetrance of deleterious mutations and apply Bayesian methods to derive carrier probabilities. The PANCPRO model gives carrier probabilities and probability of pancreas cancer developing in a given age interval. Because of the absence of a known gene, we validated our model by comparing predicted new cases to observed cases families from the National Familial Pancreas Tumor Registry. We developed a novel design approach to identify events that would be comparable to the prospective predictions made by the model, and applied an MCMC-based computational approach that accounts for the dependence of predicted cases within families. We found no significant difference in the

observed and predicated number cases, supporting the calibration of this model. We also found good discrimination ability, with an AUC of 0.72 (95% CI: 0.61, 0.82).

85

Strategy for detecting susceptibility genes with weak or no marginal effect

S. Kotti, F. Clerget-Darpoux
Inserm U535, Villejuif, France

Various statistical methods have been developed to test the interaction of two genes but mostly when the main effect of each gene can be detected. Such methods may be inadequate to successfully detect and model the interactive effect in the situations of weak or no marginal effect. We propose here a strategy for detecting two susceptibility genes in such a situation. Our method applies to trio data set genotyped for two loci A and B. The control genotype is formed by the non-transmitted parental alleles to the affected child. Patient and control genotypes are first used to estimate the relative marginal penetrances of the genotype at each locus F_A and F_B and of the joint genotype F_{AB}. In a second step, we test the null hypothesis of no interaction which is equivalent to test the following equality: $F_{AB}=F_A \times F_B{}^T$. For two biallelic loci, the test of no interaction may be achieved by a chi-square test with four degree of freedom. Under the alternative hypothesis of interaction, the statistic follows a non-central chi-square distribution for which we can calculate the non-centrality parameter and the sample size needed to achieve a given power and type I error.
We show that there exist models for which the required sample size is smaller for detecting the joint effect than the marginal effect of two genes. The procedure also allows demonstrating the involvement of two genes with no marginal effect. At a time where genome-wide association studies are fashionable, we think important to consider the alternative strategy of studying good candidate pathways with our approach.

86

Genome-wide association scans for prostate and breast cancer: design and analysis

P Kraft(1), D Hunter(1), S Chanock(2), R Hoover(3), D Gerhard(4), G Thomas(5)
(1) Prog. in Mol. and Genet. Epi., Harvard SPH, (2) Core Genotyping Facility, NCI, (3) Div. of Cancer Epi. and Genet., NCI (4) Office of Cancer Genomics, NCI, (5) INSERM U434, Fondation Jean Dausset-CEPH

We discuss several problems in the design and analysis of genome-wide association scans using the Cancer Genetic Markers of Susceptibility (CGEMS) study as an illustration. CGEMS has adopted a multi-stage approach to scan the genome for prostate and breast cancer susceptibility loci. Initial stages involve nested case-control samples from ongoing cohort studies that have been enriched for advanced disease; over 5,000 cases of each cancer are available. We present a strategy for selecting SNPs to be genotyped at each stage and corresponding multi-stage power calculations, adjusting for the fact that susceptibility loci are unlikely to be directly observed but can be detected due to linkage disequilibrium (LD) with genotyped SNPs. In particular, we examine whether follow-up scans should have greater ''breadth'' or ''depth'': whether many promising SNPs should be tested without adding nearby SNPs to improve LD with unobserved SNPs or whether the few most promising SNPs should be comprehensively tested by genotyping more nearby SNPs. We suggest that a ''broad'' approach is appropriate for early stages where thousands of SNPs will be typed on a limited number of subjects. We also present a simple analytic strategy (based on 3×3 contingency tables) that retains power whether a locus broadly influences cancer risk or only influences risk of advanced disease, highlighting the potential gain from oversampling advanced cases.

87

The ATGL gene is associated with free fatty acids, trigylcerides and type 2 diabetes

F. Kronenberg(1), V. Schoenborn(1), I.M. Heid(2), C. Vollmert(2), A. Lingenhel(1), T.D. Adams(3), P.N. Hopkins(3), T. Illig(2), R. Zimmermann(4), R. Zechner(4), S.C. Hunt(3)
(1) Innsbruck Medical University, Austria, (2) GSF, Neuherberg, Germany, (3) University of Utah, Salt Lake City, UT, (4) Karl-Franzens-University, Graz, Austria

Adipose triglyceride lipase (ATGL) was recently described to predominantly perform the initial step in triglyceride hydrolysis (Science 306:1383, 2004) and therefore seems to play a pivotal role in the lipolytic catabolism of stored fat in adipose tissue. This is the first study which investigates genetic variations within the ATGL gene in humans.
We investigated in 2434 Caucasians from Utah, USA, twelve polymorphisms in the ATGL gene identified via sequencing. These polymorphisms and their statistically reconstructed haplotypes were analysed for association with plasma free fatty acids (FFA), triglycerides, glucose and type 2 diabetes (T2DM).
FFA concentrations were significantly associated with several SNPs and haplotypes of the ATGL gene (decreased FFA levels: SNPs 3,6,8,10,11,12, p-values from 0.015 to 0.00003), consistent with additive inheritance. The pattern was similar when we considered triglyceride concentrations in our secondary analysis. Furthermore, SNP5 showed associations with glucose levels ($p<0.00001$) and risk of T2DM (OR=2.65, p=0.02). Haplotype analysis supported and extended the shown SNP association analyses.
In summary, genetic variation within the ATGL gene shows strong associations with FFA, and less pronounced but still detectable associations with triglycerides, glucose and T2DM.

88

Simple Retrospective Approaches for Detecting Interaction Effects in Case-Control Studies

LC Kwee(1), GA Satten(2), AK Manatunga(1), R Duncan(3), and MP Epstein(3)

(1) Dept of Biostat, Emory Univ, USA, (2) CDC, USA, (3) Dept of Human Genetics, Emory Univ, USA

Case-control association studies of disease often focus inference on interaction effects originating between SNPs (or haplotypes) and environment. For such interaction analyses, one can implement methods based on either a retrospective likelihood (modeling the probability of genotype and environment given disease) or a prospective likelihood (modeling the probability of disease given genotype and environment). Retrospective approaches are generally more powerful than prospective approaches (Satten and Epstein, 2004), but they seemingly require an explicit model of the joint distribution of genetic and environmental factors in the control sample. While the modeling of genetic factors is straightforward under assumptions such as HWE, the modeling of environmental factors (particularly those that are quantitative) is challenging and unappealing. To resolve this problem in interaction analyses, we propose the use of a simple likelihood framework, which is proportional to the full retrospective likelihood. This proposed approach maintains improved power over prospective approaches, but does not require any complicated modeling of the environmental distribution. Our approach is applicable to interaction analyses of both single SNPs and haplotypes. Extensions to the interaction analysis of finely-matched case-control data will also be described. We compare the power of our approach with those from approaches based on both full retrospective and prospective likelihoods. We also apply our approach to case-control data from the Finland-United States Investigation of Non-Insulin Dependent Diabetes Mellitus (FUSION) genetic study.

89

Modeling genetic association with a complex disease: the case of sepsis syndrome and death in severely injured trauma patients

I.R. König(1), T. Menges(2,3), S. Little(2,3), H. Hossain(3,4), H. Hackstein(3,5), I. Franjkovic(3,5), T. Colaris(2), F. Martens(2), K. Weismüller(2), J. Stricker(2), G. Hempelmann(2,3), T. Chakraborty(3,4), A. Ziegler(1), G. Bein(3,5)
(1) Inst. of Medical Biometry & Statistics, University at Lübeck, (2) Dept. of Anesthesiology, Intensive Care Medicine & Pain Therapy, Univ. Hospital, Gießen, (3) National Genome Research Network (NGFN), Gießen Research Center of Infectious Diseases, (4) Dept. of Microbiology, Univ. Hospital, Gießen, (5) Dept. of Clinical Immunology & Transfusion Medicine, Univ. Hospital, Gießen, Germany

Modeling the genetic association with a complex disease is an intricate process where different risk factors, various outcomes in a temporal order, and a number of candidate genes need to be considered. In our prospective cohort study on 159 patients, we investigated the role of candidate genes in severely injured patients. Several outcomes following one another were of interest: plasma level of tumor necrosis factor-alpha, development of sepsis and septic shock, and mortality within the first month. A number of possible risk factors such as age, gender, body weight, and severity of polytrauma were assessed. In addition, occurence of earlier outcomes was modeled as risk factor for later outcomes.
Using generalized linear models and path analyses, we investigated the influence of single loci and haplotypes on different outcomes controlling for the influence of risk factors. The result was a comprehensive prediction model which may have direct clinical implication for therapy.

90

On the haplotype uncertainty from genotyping and reconstruction error and its impact on association analysis

C. Lamina(1), H. Kuechenhoff(2), F. Bongardt(1), B. Paulweber(3), F. Kronenberg(1,4), H.-E. Wichmann(1), I.M. Heid(1)
(1) GSF-Inst. of Epidemiology, Neuherberg, Germany, (2) Inst. of Statistics, Univ. of Munich, Germany, (3) Paracelsus Private Medical Univ. Salzburg, Austria, (4) Innsbruck Medical Univ., Innsbruck, Austria

Haplotypes are of increasing interest in genetic association studies. Due to genotyping error in each SNP and statistical haplotype reconstruction, there is uncertainty in the haplotypes that can cause a bias in association studies or can deflate the power. Different error measures are proposed, e.g. overall error rate, or R^2, explaining the goodness of reconstruction, as well as sensitivity and specificity. To quantify these errors, we conducted simulation studies with different scenarios (e.g. real data haplotype frequencies) comparing the reconstructed with the given haplotypes and their association with quantitative or qualitative phenotypes. Genotyping error of various sizes and different error models was incorporated. We present the impact of the haplotype uncertainty resulting from genotyping error and statistical haplotype reconstruction on association estimates. For example, in a real data scenario based on a gene with high haplotype diversity, a random genotyping error of 1% per SNP reduced the sensitivity of the reconstructed haplotypes substantially by up to 25%. However, since the specificity was not greatly affected, the overall underestimation of the association of the gene with a quantitative endpoint was moderate. A practical method to correct haplotype association estimates for misclassification is presented and exemplified.

91

Classical meta-analysis applied to quantitative trait locus mapping - Genomewide linkage scan for height in the GenomEUtwin project

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

Individual loci influencing a complex quantitative trait are most likely to explain only a small proportion of its total variance. Most linkage studies published to date only consist of a few hundred pedigrees with a limited number of individuals and consequently little power to detect linkage of any but the largest quantitative trait loci (QTL).

In order to enhance power, it is now common practice to retrospectively pool evidence for linkage from several studies. We describe how classical methods for meta-analysis of clinical trials can be adapted to this pooling exercise. Provided individual quantitative trait locus estimates and associated standard errors are available on a common chromosomal grid, estimates can be pooled under the assumption of size homogeneity or heterogeneity of the QTL effects while homogeneity can itself be tested. We show also how a simple two-point mixture distribution can be employed as a novel way to allow for between-study locus heterogeneity. The methods may be applied to studies having different marker maps, family structures or different sampling schemes. We illustrate the methodology using multiple data sets for height originating from the GenomEUtwin project and representing 3212 informative families from Australia, Denmark, Finland, The Netherlands, Sweden and the United Kingdom.

92

Potential Bias in Generalized Estimating Equations Linkage Methods under Incomplete Information

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

Since Liang et al. [1] introduced the use of Generalized Estimating Equations (GEE) with the purpose of estimating the position of a locus linked to a trait, there has been increasing interest in this methodology. The approach has attractive features, in particular, it allows researchers to set a confidence interval around the estimate of the locus position. In the meantime, some refinements and extensions of the approach are being developed (e.g. [2,3]) and it bears potential for a wider use in the future. Strictly speaking, the GEE linkage method is only valid when markers are fully polymorphic, in other words, when identity-by-descent (IBD) status at markers is known with certainty. As far as we are aware, little has been done to assess how robust the method is under more realistic conditions of marker information. Using both simulations and theory, we identify some realistic conditions about marker information under which the validity of the GEE linkage methods may be arguable. Namely, researchers should not trust the GEE parameters' estimates and their associated confidence intervals in areas of the genome where IBD information is sparse or when this information changes abruptly. We show that properly standardized statistics based on IBD sharing provide a valid alternative.

[1] Human Heredity, 51, 2001, pp. 64–78
[2] Genetic Epidemiology, 24, 2003, pp. 107–117
[3] Genetic Epidemiology, 28, 2005, pp. 33–47

93

Efficiency comparisons of estimates from Classical and EM Haseman-Elston regressions when IBD sharing is ambiguous

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

The classical Haseman-Elston regression (HE) detects linkage between quantitative trait loci and a marker by regressing the squared sib-pair phenotypic trait difference on the proportion of alleles shared identical by descent (IBD). A negative regression coefficient is indicative of linkage. When marker information is incomplete, the proportion of alleles IBD is not known with certainty. Haseman and Elston proposed estimating the proportion of alleles IBD with its expected value. An alternative method, which accounts for the missing IBD sharing, is Expectation-Maximization Haseman-Elston regression (EMHE). Under the assumption of normality, EMHE simultaneously estimates the regression parameters and the posterior proportion of alleles IBD given the observed data, using the estimated IBD probabilities as weights. The variance of the EMHE regression coefficient can be estimated using Louis–method. We compared the efficiency of HE and EMHE by simulation under several genetic models and a range of sample sizes. We found no significant difference between the mean regression coefficient estimates of linkage from the two methods. However, the EMHE estimate was slightly more efficient than the HE estimate. In the case of complete marker information (i.e. unambiguous IBD sharing), HE and EMHE lead to identical parameter estimates. The EM algorithm can be applied to other extensions of the HE regression such as the revisited HE.

94

Genetic analysis of memory in relation to Alzheimer disease

J. H. Lee(1), H-S Lee(1), V. Santana(1), J. Williamson(1), R. Lantigua(1), M. Medrano(2), B. Tycko(1), E. Rogaeva(3), Y. Stern(1), P. St. George-Hyslop(3), R. Mayeux(1)
(1) Taub Institute, Columbia Univ, New York, USA, (2) Univ. Tecnologica de Santiago, Dominican Republic, (3) Centre for Research in Neurodegenerative Diseases, Univ of Toronto, Toronto, Canada

Alzheimer disease (AD) is a complex trait arising from multiple genetic and environmental factors. To enhance power to detect AD genes, we conducted linkage analysis using memory traits. Because memory traits comprise several subtypes, we explored each memory trait separately, then as factors derived from principal component analysis.
We studied 1066 affected and unaffected family members from 210 Caribbean Hispanic families with late-onset AD. In addition to a neurological evaluation, we administered neuropsychological tests to measure memory and other cognitive functions. We conducted a variance component linkage analysis as implemented in SOLAR. We adjusted for age, sex, and education, and corrected for ascertainment. We conducted two sets of linkage analysis: each memory trait separately, and three memory factors.
We observed that delayed recognition was a sensitive marker for AD with linkage peaks for the two traits often

overlapping. On the other hand, non-verbal recognition yielded low LODs, and its pattern of LODs differed from others. For long-term recall and long-term storage, the pattern of linkage across the genome was similar. We observe strong signals for several regions such as 18p11 ($2.6 < LOD < 4.5$) for recall, storage, and recognition tests, but not for non-verbal recognition and AD. In contrast to our AD scan, we observed only modest LODs for 12p13 ($LOD \sim 1.66$) and for 10q ($LOD \sim 1$). In general, the LODs for 3 memory factors were lower than the LOD for each test. Here we explore the genetic basis of AD subcomponents by examining memory traits, and show that we may be able to localize genes associated with memory, but not AD.

95

Is there a Survival Benefit of Parental Longevity?

KE Lee, BEK Klein, R Klein
Department of Ophthalmology and Visual Sciences, University of Wisconsin, Madison, WI

Most research on family history of longevity has focused on children of centenarians or on specific diseases. We explore survival using various models of parental age achieved among participants in the Beaver Dam Eye Study population. All persons aged 43–84 years and living in Beaver Dam, WI in 1988 were invited into the study. Participants (N=4926) were asked numerous health and family questions, including parental age at death, or current age if still living (vital status ignored for these analyses). All participants have been contacted annually and dates of death confirmed. Proportional hazards models were performed to predict survival by parental age after adjusting for age, gender, measures of obesity, cardiovascular disease, diabetes, cancer and other lifestyle factors. Having at least one parent surviving to age 90 years improves survival with a hazard ratio (HR) and 95% confidence interval (CI) of 0.83 (0.72, 0.96). Having both parents surviving to age 90 years further reduces the HR, but the CI is wide due to small numbers. Categorizing parental age into decades (<70, 70–79, 80–89, 90–99 and 100+ years), we find the HR and 95% CI for each decade is 0.93 (0.87,0.99). Family history of parental survival has a significant impact on survival with benefit seen even at modest degrees of ''longevity''.

96

Using SLINK to generate haplotype data in linkage disequilibrium with a trait locus and conditional on trait values in arbitrary pedigree structures

M. Lemire
McGill University and Genome Quebec Innovation Centre

The SLINK software for the simulation of genetic data in families has been widely used, with over 300 citations in the litterature to date. The fact that SLINK can simulate linkage disequilibrium (LD) between markers and between markers and the trait locus has not been documented, with the consequence that SLINK has only been used seldomly outside linkage studies over the years. In this work I review the features of SLINK that makes it one of the most flexible and powerful tool for genetic data simulation in pedigrees, as well as its most serious limitation: in practice, only a handful of markers can be simulated with SLINK. I describe how, with a simple two-step algorithm, it can be used to generate arbitrarily dense haplotypes in LD with a trait locus and conditional on trait values, as long as recombination is negligible in the region. This can be used as a simulation-based procedure to compute the power of tests of association in the presence of linkage in existing cohorts, or to help provide a correction of the type I error for multiple association testing in candidate regions. I illustrate the usefulness of SLINK in the context of LD mapping with a sample of families ascertained for asthma.

97

Controlling the Family-Wise Error Rate in Multistage Genome-Wide Association Studies

J.P. Lewinger, D.C. Thomas
Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

Most genome-wide association studies (GWAs) currently underway or in the planning stages will use a multistage design to reduce the overall genotyping cost. In a multistage design a large number of markers are first screened and a set of 'promising' markers is then tested. Several authors have shown that by lessening the impact of multiple testing adjustment, power comparable to that of a single-stage design can be achieved at a reduced genotyping cost. Although analytical results on multiple testing adjustment for multistage GWAs are available for simple settings, it remains unclear how to control the overall type I error in studies with hundreds of thousands of potentially correlated markers. Simple Bonferroni-style adjustments are not immediately available since valid p-values that properly account for the screening stages are required for each tested marker.

We propose two resampling based approaches for controlling the family-wise error rate in multistage GWAs. The first approach computes valid marginal p-values via the nonparametric bootstrap allowing subsequent adjustment using Bonferroni or its step-wise generalizations. The second is an extension of the methods introduced by Pollard, van der Laan, and Dudoit that use information from the joint distribution of the test statistics. The main practical difficulty in both approaches is obtaining bootstrap resamples from the distribution of the test statistics conditional on the outcome of the screening stages. We present importance sampling schemes to efficiently carry out the resampling.

98

Testing for heterogeneity of linkage in meta-analysis studies

C.M. Lewis(1), D.F. Levinson(2)
(1) Guy's, King's and St. Thomas' School of Medicine, KCL, UK, (2) University of Pennsylvania School of Medicine, USA

Meta-analysis of genome-wide linkage studies for complex traits is a valuable tool for identifying linked regions that individual studies may lack power to detect. The most widely used method is the Genome Scan Meta-Analysis (GSMA) method, which used a non-parametric summed rank (SR) statistic. Recently, Zintzaras and Ioannidis developed a method for testing for genetic heterogeneity in the GSMA (Gen Epi, 28:123–37, 2005), with the Q, Ha and B test statistics, used in meta-analysis of epidemiological studies. We show here that these heterogeneity statistics are correlated with the SR used to detect linkage, which must be accounted for to assess significance correctly. For genetic regions with a high SR, the unadjusted Q statistic provides a highly conservative test of genetic heterogeneity. For example, with 10 studies, using a global 5% critical value for Q produces a test of size 1.7% for regions with SR in the top decile.
In simulation studies for 20 scans, the testing procedure adjusted for SR produces a test of the correct size for Q. We used the GSMA and heterogeneity tests to analyse simulations of 20 scans, with a single locus with sibling relative risk λ_S =1.3. The power to detect heterogeneity using Q was low: a power of 13.5% when 75% of studies are linked, 13.0% with 50% of studies linked, and 8% with 25% of studies linked. Test statstics Ha and B had slightly lower power. Testing for heterogeneity is a valuable addition to meta-analysis of linkage studies using the GSMA, but the currently available methods lack power.

99

Genome-wide linkage scan in keratoconus sib-pair families

X. Li(1), Y.S. Rabinowitz(2), K.D. Taylor(1), Y.G. Tang(1), M. Hu(1), Y. Picornell(1), H. Yang(1)
(1) Medical Genetics Institute, (2) Cornea-Genetic Eye Institute, Division of Ophthalmology, Cedars-Sinai Medical Center, USA

To identify the susceptibility gene loci for keratoconus (KC), we performed a genome-wide linkage analysis using data from 67 KC sib-pair families with 107 affected-sib pairs. A total of 356 subjects were genotyped for 408 microsatellite markers along the genome at ~10 cM density. Multipoint linkage analysis was performed using all pedigrees and Caucasians only (40 families, 217 individuals) by non-parametric methods and maximum likelihood estimates of identity by descent sharing as implemented in GeneHunter. Most linkage peaks were consistent between Caucasian and all pedigrees. The strongest evidence of linkage was observed at the telomere (159 cM) of chromosome 9 (LOD=4.5) for all pedigrees. Other suggestive linkages were identified at 176 cM of chromosome 4 (LOD=2.66), 143 cM of chromosome 5 (LOD=2.0), 7 cM of chromosome 9 (LOD=2.8), 12 cM of chromosome 11 (LOD=2.3), 27 cM of chromosome 12 (LOD=2.3) and 14 cM of chromosome 14 (LOD=2.9). These results indicate that several loci may contribute to KC susceptibility. In addition, it is intriguing that we have observed evidence of linkage at 103 cM on chromosome 5 in a four-generation pedigree (published) as well as a maximum LOD of 2.0 from sib-pair families at 143 cM on chromosome 5. Although these two peaks do not perfectly overlap, their close proximity raises the possibility that these genomic regions might be the same KC susceptibility locus.

100

Selection of SNPs for Evaluating Gene-Disease Associations Using Haplotypes

N. Li and M. Li
Div. of Biostat., School of Public Health, Univ. of Minnesota, Minneapolis, USA

Multiple single nucleotide polymorphisms (SNPs) have seen widespread usage in evaluating potential association between candidate genes or regions with diseases. Because SNPs are diallelic, it is widely thought that haplotypes, i.e., combinations of SNPs on the same chromosome, would provide greater power. However, when the number of SNPs is large, many of them are not informative regarding the gene-disease association. Using the haplotypes from all SNPs result in a large number of parameters and loss of degrees of freedom and power. Much attention has been paid to the problem of selecting tagSNPs in the study design stage. However, it is not yet clear whether the tagSNPs are sufficient to capture disease-gene association information or which selection method is the best. Furthermore, the number of tagSNPs can still be too large. In this paper we propose a strategy for selecting a subset of SNPs to use in a regression-based haplotype analysis after the genotyping has been completed. The optimality is defined based on an information-theoretic criterion and a greedy search is performed. The goodness-of-fit of the final selected model (with a suitably defined loss function) is assessed using a parametric bootstrap method which takes into account the uncertainty associated with the model selection process. The final model can be then used to evaluate the gene-disease association, in particular, to compare the strength of association of the disease with several candidate genes and to evaluate potential gene-gene and gene-environment interactions.

101

Two-Locus Analysis of Gene-Gene Interaction Using Case-Parents Trios

C. Li
Dept of Biostatistics, Center for Human Genetics Research, Vanderbilt Univ, USA

Association analysis using affected individuals and their parents is powerful to detect disease-gene association. For a single marker, the transmission/disequilibrium test has been commonly used to detect such an association. Complex diseases often involve multiple genes and environmental factors. When we suspect two genes interact to modify disease risk, we want to test for their interaction and to estimate disease relative risks associated

with two-locus genotypes. For case-parents trios, we propose a general, likelihood-based method for two biallelic markers that are either unlinked or linked but in linkage equilibrium. The method uses log-normal models and can be used to test for statistical interaction between two loci and to estimate relative risks for two-locus genotypes. Given a locus that is associated with disease, we also may want to test if a second locus has additional effects on disease risk. In this situation, two competing tests exist, one using a fully saturated interaction model as the alternative hypothesis and the other using a model of independent, multiplicative effects between loci. The former is more powerful when gene-gene interaction exists, while the latter is more powerful when the two genes influence disease risk independently. In reality, however, we often do not know which effect is correct; correcting for multiple comparisons may lead to loss of power because the two tests are positively correlated. To address this problem, we develop a single, combined test, which is almost as powerful as the optimal test in either situation. The method has been implemented in an R package that is freely available.

102

Heritability of endothelin secretion, and the influence of polymorphism at the chromogranin A locus, a regulator of catecholamine storage and release

EO Lillie(1), M Mahata(1), J Wessel(1), S Khandrika(1), F Rao(1), G Wen(1), BA Hamilton(1), M Cockburn(2), BK Rana(1), DW Smith(1), SK Mahata(1), NJ Schork(1), DT O–Connor(1)
(1) University of California at San Diego, USA, (2) University of Southern California, USA

Endothelial dysfunction predisposes to vascular injury in association with hypertension. Endothelin (ET-1) is a potent vasoconstrictor that is synthesized and released by the vascular endothelium. Elevated ET-1 is a marker of endothelial dysfunction. Chromogranin A (CHGA) regulates catecholamine storage and release and may also have direct action on the microvasculature. CHGA is a candidate gene for intermediate phenotypes that contribute to hypertension, and shows a pattern of SNP variations that alter the expression and function of this gene. In a study of Southern California twins (N=238 pairs), plasma ET-1 is 63+/-4% heritable ($p<0.0001$). We, therefore, hypothesized that variation in the CHGA gene may influence ET-1 variation. In a sample of Caucasian twins and siblings (N=371), using generalized estimating equations adjusted for age and gender, promoter SNP A-462G (a polymorphism with demonstrated effects on transfected CHGA promoter transcription) shows increasing levels of plasma ET-1 with each promoter A allele (p trend=0.016). Haplotype analysis of 7 promoter SNPs shows a significant association with plasma ET-1 in subjects carrying the common haplotype (freq>1%) that contains the A allele ($p<0.0001$). These results are novel and suggest that common variation in expression of the CHGA gene influences ET-1 secretion.

103

A genome-wide scan for quantitative trait loci of serum gamma glutamyltransferase – the Framingham Offspring Study

J.-P. Lin(1), C.J. O'Donnell(2), L.A. Cupples(3)
(1) Office of Biostatistics/DECA/National Heart, Lung and Blood Institute, USA, (2) Framingham Heart Study/DECA/National Heart, Lung and Blood Institute, USA, (3) Dept. of Biostatistics, Boston University School of Public Health, USA

In addition to its diagnostic uses in liver function tests, variation in serum gamma glutamyltransferase (GGT) in the population is independently associated with risk of death, development of cardiovascular disease, type 2 diabetes, stroke, and hypertension. Twin studies have reported that serum GGT variation is significantly determined by genetic factors with heritability estimates ranging from 35–65%. Four GGT isoforms have been previously mapped to chromosome 22q11 and 22q13. To date, no linkage analysis on serum GGT levels has been reported. We carried out a 10 cM genome-wide scan for quantitative trait loci of GGT in a community-based Caucasian cohort, the Framingham Heart Study. GGT was measured in the second examination of the offspring cohort. Our study population consisted of 330 families with 1096 individuals being both genotyped and phenotyped, including 621 sibling pairs, 414 cousin pairs, and 35 avuncular pairs. Using variance-component linkage methods implemented in SOLAR, the heritability was estimated as 33% after age, sex, serum albumin and total serum protein adjustment. The genome-wide linkage analysis by empirical P value method yielded several chromosomal regions with LOD scores between 1–2: LOD scores of 1.54, 1.73, 1.14, 1.28, 1.29 and 1.20 on chromosomes 3, 8, 9, 10, 13 and 15, respectively. Our study suggests that instead of a gene with a large effect, there may be several genes with small effects in controlling the variation of serum GGT. Those genes seem to reside in chromosomal regions that differ from those containing the genes encoding GGT isoforms.

104

Exploring the Effect of Interactions of GAPD Genes on Late-onset al.zheimer Disease

P.I. Lin(1), E.R. Martin(1), P.G. Bronson(1), C.A. Browning-Large(1), G.W. Small(2), D.E. Schmechel(3), K.A. Welsh-Bohmer(3,4), J.L. Haines(5), J.R. Gilbert(1), M.A. Pericak-Vance(1)
(1) Center for Human Genetics, Department of Medicine, Duke University Medical Center, USA, (2) Department of Psychiatry and Biobehavioral Sciences, USA, (3) Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, USA, (4) Joseph and Kathleen Bryan Alzheimer's Disease Research Center, Duke University Medical Center, USA, (5) Center for Human Genetics Research, Vanderbilt University Medical Center, USA

Previous candidate gene studies on chromosome 12 for late-onset al.zheimer disease (LOAD) have been inconclusive.

Recently, Li et al. (2004) reported the association between the glyceraldehyde-3-phosphate dehydrogenase (GAPD) gene on chromosome 12p and the risk of LOAD. In the current study, we propose to test the effects of GAPD gene and its paralogs in an independent data set. We carried out family-based and case-control association studies on 12 single nucleotide polymorphisms (SNPs) in GAPD gene and members of the GAPD family. One 5′UTR SNP in GAPD gene was associated with the LOAD risk ($p=0.03$) particularly in the subgroup with age at onset >71 years of age, but the association was not significant after we performed the false detection rate procedure. However, a significant interaction effect of GAPD gene and its paralog, PPM1H gene was observed by using Extended Multi-dimensionality Reduction method, in which modified odds ratio was assessed. Our findings suggest that the gene-gene interaction approach may be particularly useful for identifying genes of small effect.

105

Maximum Likelihood Methods for Haplotype Sharing Studies

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

It is highly challenging to study associations between haplotypes and disease phenotypes because of unknown gametic phase in genotype data. The common practice of probabilistically inferring individual haplotypes leads to biased and inefficient analysis of association.
We present a unified likelihood-based approach to this problem for all study designs, including cross-sectional, case-control, cohort, nested case-control, and case-cohort designs. The phenotypes can be disease indicators, quantitative traits, or potentially censored ages at onset of disease. The effects of haplotypes on the phenotype are formulated through flexible regression models, which can accommodate a variety of genetic mechanisms and gene-environment interactions. We construct appropriate likelihood functions under Hardy-Weinberg disequilibrium and show that the maximum likelihood estimators are consistent, asymptotically normal, and statistically efficient. We develop stable and efficient numerical algorithms to implement the corresponding estimation and testing procedures. Both candidate genes studies and genomewide scans are considered. Simulation studies demonstrate that the new methods perform very well in practical settings. Applications to several genetic epidemiological studies reveal important haplotype effects and haplotype-smoking interactions. A general computer program is freely available.

106

Haplotype Analysis in the Presence of Missing Data

N. Liu(1), H. Zhao(1,2)
(1) Department of Epidemiology and Public Health, Yale University School of Medicine, USA, (2) Department of Genetics, Yale University School of Medicine, USA

It is common to have missing genotypes in practical genetic studies, yet the exact underlying missing data mechanism is generally unknown to the investigators. Currently, data are assumed to be missing at random (i.e. different genotypes and different alleles are missing with the same probability) in most statistical approaches, both for haplotype frequency estimation and for haplotype association analysis. However, very few studies have examined the magnitude of potential biases based on these methods when this simplifying assumption is violated. In this study, we show that haplotype frequency estimates can be biased using methods assuming missing at random if genotypes/alleles are not missing at random. Similarly, haplotype association analysis can be biased as well, inducing both false-positive and false-negative evidence of association. We propose a general missing data model to characterize missing data patterns across a set of two or more markers simultaneously. We use simulations of two SNPs (single-nucleotide polymorphisms) to illustrate that our proposed model can reduce the bias caused by incorrectly assuming missing at random, and have reliable estimates. We also prove that haplotype frequencies and missing probabilities are identifiable if and only if there is linkage disequilibrium (LD) between these markers under our general missing data model. Finally, we illustrate the utilities of our method through its application to a real data set.

107

Nonsteroidal Anti-inflammatory Drugs (NSAIDs) and advanced prostate cancer: modification by LTA+80

X. Liu(1), S.J. Plummer(2), N. Nock(2), G. Casey(2), J.S. Witte(1)
(1) Department of Epidemiology & Biostatistics, University of California, San Francisco, USA, (2) Department of Cancer Biology, Lerner Research Institute, Cleveland Clinic Foundation, USA

Nonsteroidal anti-inflammatory drugs (NSAIDs) may protect against prostate cancer partly through inhibition of cyclooxygenase (COX) enzymes and nuclear factor kappa B (NF-kB) as well as other important molecules involved in the inflammatory process. Therefore, any potential protective effects of NSAIDs may be modified by variants in inflammatory genes. A recent study demonstrated that a single-nucleotide polymorphism (SNP) in lymphotoxin-alpha (LTA+80) can serve as a main predictor of LTA protein production through allele-specific recruitment of activated B-cell factor-1 (ABF-1). To determine if LTA+80 modifies the protective effect of NSAIDs on prostate cancer risk, we conducted a case-control study of 510 men diagnosed with advanced prostate cancer and 538 age- and ethnicity-matched controls. We observed a significant protective effect of using aspirin or ibuprofen on the risk of advanced prostate cancer (OR=0.67, 95%CI: 0.52 $\rangle$ 0.86). This association was substantially modified by LTA+80 genotypes. In particular, the association between NSAIDs use and disease was stronger for men carrying the CC genotype (OR=0.42, 95%CI: 0.28 $\rangle$ 0.65), and weaker for those with the AA/AC

genotypes (OR=0.86, 95%CI: 0.62 〉 1.18) ($P_{interaction}$=0.01). Our findings suggest that any potential chemoprevention of advanced prostate cancer by NSAIDs may be most appropriate for men with particular genotypes.

108

The importance of "Uninformative" models in Bayesian Linkage Analysis

MW Logue, Y Li, VJ Vieland
Center for Statistical Genetics Research, University of Iowa

The posterior probability of linkage, or PPL, is a Bayesian statistic that directly measures the probability that a disease gene is linked to a marker (2-point), or genomic location (multipoint). It allows for an unknown genetic model without the inflationary effects of maximization by placing a Bayesian prior over the elements of the genetic model and integrating them out of the likelihood. To this point, we have used essentially uniform priors over the elements of the penetrance vector. However, much of the parameter space corresponds to models which seem, on the surface, unlikely to yield substantial evidence for linkage: for example, models with very high disease allele frequencies and/or very high phenocopy rates. Although the prior has no effect on the asymptotic behavior of the PPL, we investigate the effect of several new priors on the elements of the genetic model in realistic sample sizes. These priors place higher weight over genetic models with higher sibling relative risk, λ. In a series of simulations, under various generating conditions we show that placing more weight over the high λ models does tend to increase the mean PPL under linkage, and to decrease the mean PPL under no-linkage. However, these priors can also cause the PPL to occasionally yield unacceptably high values under no linkage. On balance, and somewhat surprisingly, it appears important to retain prior probability over apparently "uninformative" (low λ) models when integrating over trait parameters.

109

Bivariate linkage analysis of triglycerides and HDL-cholesterol in Hispanic families from the GENNID study

A. Malhotra(1), J. Xu(2), J.K. Wolford(1)
(1) Diabetes & Obesity Unit, TGen, Phoenix, AZ, (2) Center for Human Genomics, Wake Forest U, Winston-Salem, NC

Genome scans for quantitative lipid traits have shown regions of linkage for multiple traits in the same region of a chromosome. This may suggest pleiotropic effects of a gene present in that region. A previous univariate genome scan using the program GENEHUNTER showed evidence of linkage for triglycerides (TG; LOD=2.26) on chromosome 11p15.4-11p11.3 utilizing data from 113 Hispanic families from the Genetics of NIDDM (GENNID) study. Linkage was also observed for HDL-C (LOD=1.15) in this region, suggesting the presence of a common underlying gene affecting both traits in this population. To address this possibility, we performed a bivariate genome scan of HDL-C and TG using the program SOLAR. The trait values from the initial univariate analyses were utilized (i.e., adjustment of covariate and standardization were performed prior to the analysis). The highest bivariate LOD score (1.92) was found on chromosome 11p. A test for pleiotropy, as implemented in SOLAR, showed that the genetic correlation was significantly different from -1, but not significantly different from 0. These results suggest that while linkage for both HDL-C and TG is present on chromosome 11p, they may be independently affected by different genes in this region. We also observed a bivariate LOD score of 1.90 on chromosome 1p. Univariate analyses showed only nominal evidence for linkage in this region (HDL-C: LOD=0.15; TG: LOD=0.24), indicating that bivariate analysis may be useful for detecting linkage when individual genetic effects are weak.

110

A comparison of clinico-pathological factors in sporadic and familial African-American and Caucasian prostate cancer cases in Louisiana

D.M. Mandal, S.L. Halton, J.J. Hunter, O. Sartor, J.E. Bailey-Wilson, W. Rayford

The exact causes of prostate cancer are unknown due to the complex nature of the disease. Important risk factors include family history and race of the patients. African-American men have an incidence rate of prostate cancer almost two times greater than Caucasian American men and in addition, they are two to three times more likely to die from this disease. The aim of the present study is to compare clinical/pathological characteristics in African-American and Caucasian males with and without family history of prostate cancer. Clinical records were reviewed for 250 familial and sporadic prostate cancer cases (approx. 110 African-American and 140 Caucasian) and the clinical characteristics, including age at onset, Prostate Specific Antigen (PSA) values at the time of diagnosis, Gleason scores and clinical stages were compared in African-American and Caucasian males. Preliminary analyses show that there is a significant median difference in the age at onset between familial and sporadic prostate cancer cases in African-American (p=0.02) males and age at onset in familial African-American and Caucasian males (p=0.03). These results provide evidence for the positive association between family history and early age at onset in African-American prostate patients in Louisiana.

111

Effects of single SNPs, haplotypes, and whole genome LD maps on accuracy of association mapping

N. Maniatis, J. Gibson, S. Ennis, A. Collins, N.E. Morton
Human Genetics Division, Southampton General Hospital, UK

We have developed a simple yet powerful approach for disease gene association mapping by linkage disequilibrium (LD). This method is unique since it applies a model with evolutionary theory that incorporates a parameter for the location of the causal polymorphism. The method is based on single marker tests and LD maps, which describe

the pattern of LD by assigning a location in LD units (LDU) for each marker. As a proof of principle, we tested our method using 27 SNPs that cover an 890 kb region flanking the CYP2D6 gene with known location on chromosome 22. Previous LD mapping studies have identified a 390 kb region associated with the poor drug metabolising phenotype. Using a metric LDU map, the commonest functional polymorphism within the gene was estimated to be located only 14.9 kb from its true location, surrounded within a 95% CI of 172 kb. The kb map had a relative efficiency of 33% compared to the map in LDU. Despite the low resolution and the very strong LD in the region, our results provide evidence of the substantial utility of LDU maps for disease gene association mapping. We examined the performance of this mapping approach based on high density LDU maps constructed from the HapMap data. Expressing the locations of the 27 SNPs in LDU from the HapMap LDU map of chromosome 22 yielded a much lower location error (0.5 kb). These tests are robust to large numbers of markers and are applicable to whole-genome association or candidate region studies. We are also comparing the power between single SNPs and haplotypes.

112

Interpretation of a combined two-point and multipoint linkage statistic in the genome screening of a multifactorial disease

P. Margaritte-Jeannin, M.C. Babron, F. Clerget-Darpoux
INSERM U535, Villejuif, France

Systematic linkage test on the whole genome is a widely used approach for localizing susceptibility genes of multifactorial diseases. The linkage statistic may take into account the information either on each marker successively (two-point), or on all the markers simultaneously (multipoint).
It has been argued that two-point statistics may be more reliable because of the uncertainty on the inter-marker distances. However, two-point analyses lead to a loss of power, as they use only part of the information compared to the multipoint ones. Consequently, in a second step, multipoint analyses are often carried out to reinforce the results of two-point analyses. The markers with a high two-point HLOD but with a low multipoint one are likely to be false positives. However, the overall significance for the markers which are replicated by the multipoint analysis is not provided.
This is illustrated by a recently published genome scan of a large sample of Multiple Sclerosis families (Babron et al., Am J Hum Genet, 75:1070–1078, 2004) in which, apart from HLA (a known risk factor for MS), the only marker highlighted by both two-point and multipoint approaches was D1S547 on 1q.
In this study, we assess the probability under H0 (no risk factor) of being over the 5% threshold in the same region for both two-point and multipoint HLOD statistics in the MS family sample by simulations. The 5% genome wide threshold are estimated to 1.90 and 1.98 for the two-point and multipoint HLOD, respectively. The type I error corresponding to these thresholds for the combined two-point and multipoint statistics is equal to 2%.

113

A Quantitative Trait Locus for Bicuspid Aortic Valve and Associated Cardiac Anomalies Localizes to Chromosome 6p

LJ Martin, L Cripe, G Andelfinger, M Tabangin, K Shooner, DW Benson
Cincinnati Children's Hospital Medical Center & the University of Cincinnati School of Medicine

Bicuspid aortic valve (BAV) describes an aortic valve with two rather than three leaflets. BAV is the most common cardiovascular malformation occurring in 1% of the population. We have demonstrated that BAV and associated heart anomalies are highly heritable; however, no genes have been identified. Therefore, our objective was to identify loci linked to BAV and associated cardiac anomalies. We recruited 6 probands with BAV and obtained a 3-generation family history and performed echocardiograms on first-degree relatives. If additional individuals with BAV were identified, their first-degree relatives were also invited to participate. Out of 120 individuals, 13 had BAV and 9 had associated heart anomalies. Individuals were genotyped for 400 markers spread across the genome. Linkage analysis was performed for parametric and non-parametric models using GENEHUNTER and SOLAR. The maximum LOD score detected was 3.5 and 2.9 for the recessive and non-parametric models, at D6S1610 on human chromosome 6p21.2. The one LOD unit support interval is 35.5 Mb and contains 313 genes. Several genes encoded in the interval have a role in cardiogenesis/valvulogenesis *BMP5, VEGFA, NOTCH4, and TNXB* making them interesting biologic candidates. No other chromosomal region showed evidence of linkage. The identification of this novel QTL is the first step in understanding the etiology underlying not only BAV but valve malformations in general.
This work was supported by NIH grants HL74728 and MH59490.

114

Estimating Familial Correlations from Pedigree Data

G. Mathew(1,2), Y. Song(1), R.C. Elston(1)
(1) Department of Epidemiology and Biostatistics, Case Western Reserve University, U. S. A., (2) Department of Mathematics, Missouri State University, U. S. A.

Familial correlations can be estimated by a weighted version of Pearson's product moment correlation. Two weights that can be assigned, which represent extremes, are equal weights $w1$ to each pair of persons in the pedigree and equal weights $w2$ to each pedigree. We consider finding the best weighted average of these two estimates by finding the linear function $aw1+(1-a)\ w2$ of the weights that minimizes the variance of the familial correlation estimate. We do this by fitting a quadratic to the variance of the estimate, as a function of a, and finding the value of a that minimizes the quadratic function.

We also consider a computationally feasible method to estimate the variance of this linear function. We examine the effectiveness of this approach by simulation to estimate parent-offspring, sib-sib, grandparent-grand child, avuncular and cousin-cousin correlations. We provide an analysis of the bias, the variance and the coverage probabilities for confidence intervals of the estimates.

115

Distribution of MOD scores under no linkage

M. Mattheisen(1), K. Strauch(1,2)
(1) Institute for Med. Biometry, Informatics, and Epidemiology, Univ. of Bonn, Germany, (2) New address: Institute for Med. Biometry and Epidemiology, Univ. of Marburg, Germany

The asymptotic distribution of MOD scores under the null hypothesis of no linkage is only known for affected sib pairs (ASP). When imprinting is not taken into account, a P-value of 0.0001 corresponds to a MOD score of 3.29 and, when allowing for imprinting, to a MOD score of 3.70. This follows from the equivalence between a MOD score analysis and Holmans– possible triangle test or the T_{ILR} test, respectively. We performed simulations with families of different size (one million replicates, each with 500 families) to investigate the impact of the pedigree size on the null distribution of MOD scores, in case that larger families are analyzed. We found that adding one unaffected child to the ASP families (resulting in discordant sib triplets) has a much greater impact on the critical MOD score value than adding a third affected child (critical MOD scores of 3.91 vs. 3.41, both corresponding to a P-value of 0.0001). Allowing for imprinting in the analysis of the discordant sib triplet families has a further substantial impact on the asymptotic distribution of the MOD scores under no linkage (critical MOD score of 4.25 vs. 3.91). When analyzing a 3-generation family (three phenotype-unknown founders, four affected non-founders, one unaffected non-founder) and allowing for imprinting, we found a critical MOD score of 4.45 corresponding to a P-value of 0.0001. In conclusion, we could show that the null distribution of MOD scores is strongly influenced by the size and structure of the pedigrees under study.

116

Power and false positive rates for genetic association tests: impact of familial relatedness

P.F. McArdle(1), J.R. O–Connell(1), A.R. Shuldiner(1,2), B.D. Mitchell(1)
(1) University of Maryland School of Medicine, USA, (2) Geriatrics Research and Education Clinical Center, VA Hospital Medical Center, Baltimore MD, USA

Assessments of SNP-trait associations are commonly made within the context of family studies. We evaluated the impact of accounting for relatedness between individuals on detecting SNP associations in family-based case control studies by performing simulations among 28 large families from the Amish Family Diabetes Study. To assess power to detect associations, we simulated traits using a threshold liability model whose variation was attributable to a single SNP and a polygenic background. SNPs were simulated for a variety of effect sizes and parameterized as odds ratios. The power to detect the additive effect of an allele with a frequency of 40% and effect size of odds ratio=1.6 (common variant, small effect) on a trait with heritability of 40% was 0.87 when accounting for family structure and 0.89 when ignoring it (alpha=0.05). False positive rates were assessed by simulating unlinked markers and evaluating their associations with a measured phenotype (in this case presence of diabetes). The heritability of the diabetic phenotype was 60%. Ignoring family structure leads to a doubling of the false positive rate (10.6% from 5.2%). These results indicate that failure to adequately account for family structure has little influence on power to detect SNP associations, but can markedly elevate false positive (type 1 error) rates.

117

SimHap: A comprehensive modeling framework and simulation-based approach to haplotype analysis for population data

P.A. McCaskie(1,2), K.W. Carter(1,2), L.J. Palmer(1,2)
(1) Laboratory for Genetic Epidemiology, Western Australian Institute for Medical Research, Perth, Australia, (2) School of Population Health and Centre for Medical Research, University of Western Australia, Perth, Australia

SimHap is a statistical analysis package that we have developed for genetic association testing. It can perform single SNP and multi-locus (haplotype) association analyses for continuous normal, binary, longitudinal and right-censored outcomes measured in population-based samples. A range of genetic models may be tested for genotypes and diplotypes. SimHap provides a holistic framework for statistical modeling, and also allows analysis of epidemiological models without the inclusion of genetic effects. SimHap uses current estimation-maximisation techniques for inferring haplotypic phase in individuals, and incorporates a novel simulation-based approach to deal with the uncertainty of imputed haplotypes in association testing. The program can accommodate large data sets, enables simple and intuitive subset analyses, and can model both genetic and environmental (covariate) effects, including complex haplotype:environment interactions. SimHap has been written in R and Java to allow the incorporation of complex statistical techniques while still providing cross-platform functionality and a user-friendly graphical user interface, so users need not have a comprehensive knowledge of command line operation to perform complex analyses (www.genepi.com.au/projects/simhap). Extensive simulations have been conducted to investigate the empirical properties of this analytic technique.

118

Genetic Variation Associated with Left Ventricular Traits in Hypertensive African-Americans

KJ Meyers(1), TH Mosley(2), E Boerwinkle(3), IJ Kullo(4), S Turner(4), 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, (4) Cardiovascular Diseases, Mayo Clinic and Foundation

As part of the Genetic Epidemiology Network of Arteriopathy(GENOA) study, hypertensive African-American sibships were screened using two-dimensional echocardiography and 416 SNPs in candidate genes were genotyped. Three echocardiographic traits that are associated with cardiovascular mortality and morbidity were studied: left ventricular mass index(LVMI), relative wall thickness(RWT), and aortic root diameter(ARD). To identify SNP associations with replicate effects across samples, we took two sibs from each sibship and divided them into two samples, each with 448 unrelated individuals. We identified 94 SNPs(36 with LVMI, 18 with RWT, and 40 with ARD) in the first sample and 108 SNPs(35 with LVMI, 32 with RWT, and 41 with ARD) in the second sample that were associated with the three traits after adjusting for appropriate covariates. Three SNPs - one each in the APOE, SCN7A, and SLC20A1 genes - were significantly associated in both samples with LVMI and had replicate genotype-phenotype relationships. One SNP in the ADRB1 gene was significantly associated with RWT with replicate effects. Finally, 4 SNPs in the KCNJ13 gene and 1 SNP in the SLC12A3 gene were significantly associated with ARD with replicate effects. This study capitalized on the affected sib design for genetic association studies and identifies genetic variation that influences LV traits with replicable effects in an African-American population.

119

GenetSim: Software for Simulation of Familial Data in Genetics and Epidemiology

M.B. Miller(1), N. Li(2)
(1) Div. of Epidemiology, (2) Div. of Biostatistics, University of Minnesota

GenetSim provides flexible simulations of family data within an easy-to-use, high-level programming language. GenetSim was developed first within the MATLAB-like environment of the free software package Octave (Eaton, 1997), but newer versions are designed to work with the R statistical language. GenetSim has no limit on pedigree sizes or structures (these can be imported from LINKAGE-format files), or number of families, no limit on number of marker or trait loci, no limit on number of chromosomes (nonhuman diploid species can easily be modeled). Genetic transmission is modeled by first generating the locations of recombination events (according to nearly any multilocus feasible model desired - Haldane, Sturt, etc., or a user-specified model), and then performing gene dropping according to the given recombination pattern. Any pattern of missing data can be specified and genotyping errors (or other kinds of errors) can be simulated. GenetSim can simulate multiple QTLs with pleiotropic effects, multivariate polygenic background and any number of environmental factors, age effects, epistasis and variable expression. Traits could be quantitative (continuous) or one could use penetrance functions and/or liability threshold models for affection-status (binary) traits. Users can also select families based on ascertainment schemes by repeating simulations (e.g., retain only families where at least two members have trait values exceeding 140). GenetSim is freely available under the GNU General Public License.

120

Why do results from individual-level (IA) and family-based (FA) association analyses of candidate genes and time-to-event phenotypes differ?

L. Mirea(1,4), S.B. Bull(1,4), A.D. Paterson(2,4), A.P. Boright(3), M. Liu(2), B. Zinman(1,5), The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) Study Research Group
(1) Samuel Lunenfeld Research Inst, Mt Sinai Hospital, Toronto, Canada, (2) Program in Genetics & Genomic Biology, Hospital for Sick Children, Toronto, Canada, (3) Dept of Medicine, University Health Network, Toronto, Canada, (4) Dept of Public Health Sciences, Univ of Toronto, Canada, (5) Leadership Sinai Centre for Diabetes, Mt Sinai Hospital, Toronto, Canada

We performed a comparative case study for renal complication-free survival and SNP rs1800764 in the ACE gene in a cohort of 1365 ''White'' probands with type 1 diabetes recruited in the DCCT/EDIC study. We found no evidence for population structure using the program STRUCTURE with 25 unlinked markers. Cox model analyses of the 1365 probands with adjustment for important covariates revealed lower risk for the TT vs CT genotype (HR=0.61, pvalue=0.003). Of the 1365 probands, 837 have DNA available from family members. In the subset of 448 probands with relatives informative for FA at this SNP, the IA association was similar but less significant (HR=0.50, pvalue=0.03). FA analyses, treating residuals from the Cox model as a quantitative trait, showed weak evidence of excess C allele transmissions (pvalue=0.06) under a dominant model. In this case, the observed discrepancies can be substantially explained by reduced efficiency in the subset of probands with available relatives that are informative for FA at this SNP.

121

ALBERT: A Likelihood-Based Estimation of Risk in Trios

Adele A. Mitchell(1), Eileen S. Emison(2), Aravinda Chakravarti(2), Elizabeth A. Thompson(1)
(1) Departments of Statistics and Genome Sciences, University of Washington, Seattle, WA, (2) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

The original transmission disequilibrium test (TDT) developed by Spielman et al., is a popular test for linkage in the presence of association in case-parent trios. Two major problems with the TDT are that its false positive rate is inflated in the presence of undetected genotyping error

and it suffers from low power, particularly when the penetrance of the disease of interest is not multiplicative with the number of copies of the risk allele.
Here, we present ALBERT, A Likelihood-Based Estimation of Risk in Trios. ALBERT tests for association between a diallelic marker and a dichotomous trait in case-parent trios. The method, which provides estimates of genotype relative risks, genotyping error rate and risk allele frequency, is more powerful than the TDT for most sample sizes, genotype relative risks, allele frequencies and genotyping error rates. It has a low false positive rate and is robust to the presence genotyping errors. We have implemented the method in a software package, ALBERT, which is freely available through our website.
We applied our method to 29 diallelic markers spanning the RET region on chromosome 10q11.12 in 130 trios ascertained through a child affected with Hirschprung disease. We found a stretch of nine consecutive markers showing significant association ($p < 10^{-4}$) with Hirschprung disease; the middle four of these markers were significant at $p < 10^{-9}$. Thus, ALBERT has shown itself to be an effective tool in identifying alleles associated with increased risk of disease.

122

A Flexible Data Mining Framework for Detecting and Interpreting Gene-Gene Interactions

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

Detecting and interpreting gene-gene interactions in studies of human disease susceptibility is a computational and a statistical challenge. To address this problem, we have previously developed a multifactor dimensionality reduction (MDR) method for combining multiple SNPs into a single predictor (i.e. constructive induction). Here, we describe a flexible framework for detecting and interpreting interactions that utilizes 1) advances in information theory for selecting interesting SNPs, 2) MDR for constructive induction, 3) machine learning methods for classification, and 4) graphical models for interpretation. We illustrate the usefulness of this strategy using artificial datasets simulated from several different two-locus and three-locus interaction models. We show that the accuracy, sensitivity, specificity, and precision of a naïve Bayes classifier are significantly improved ($P < 0.05$) when SNPs are selected based on their information gain (i.e. class entropy removed) and reduced to a single predictor using MDR. We then apply this strategy to detecting, characterizing, and interpreting epistatic models in a genetic study (n=500) of atrial fibrillation and show that both classification and model interpretation are significantly improved ($P < 0.05$). A strength of this approach is the ability to plug-and-play different statistical and computational methods at each of the four steps. Open-source software is available from www.epistasis.org/mdr.html.

123

Allowing for haplotype-environment interaction effects in population-based association studies

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

Population-based association studies have been widely recognised as having the potential to efficiently map genetic polymorphisms contributing susceptibility to complex human diseases. However, in reality, success has been limited to a handful of major gene effects, with otherwise little evidence of replication of ''positive'' association signals. One possible explanation for this observation is gene-environment interaction (GxE), where the genetic contribution to disease risk is modified by exposure to different environmental conditions, which may vary from one study to another.
To overcome this problem, we extend the method of Morris (2005) to allow for interaction between environmental covariates and SNP haplotypes in a candidate gene. The log-odds of disease of each individual is modelled in a logistic regression framework, parameterised in terms of haplotype main effects and haplotype interaction effects with covariates. Unphased SNP genotype data can be utilised by considering all possible haplotype configurations, weighted in the logistic regression model by the corresponding estimated phase assignment probabilities. To reduce the number of model parameters, haplotypes are clustered according to their similarity in terms of allelic make-up so that each haplotype in the same cluster is assigned the same main effect and interaction effects with covariates. A Markov chain-Monte Carlo algorithm is developed to sample from the space of haplotype clusters and model parameters, and hence to estimate the posterior probability of haplotype association with disease, allowing for GxE. The results of a detailed simulation study illustrate the increased detection of true positive associations by allowing for haplotype-environment interactions by this approach over a range of GxE models.
Morris AP (2005). Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. Gen Epidemiol (in press).

124

Association Analysis for Sets of Correlated Markers

Nitai D Mukhopadhyay and Nicholas J Lewin-Koh
Eli Lilly and Co.

High density genome scans often show random patterns of association due to sampling variability and small samples. The large amount of noise and the nonstationary patterns of linkage observed across the genome, make traditional methods of picking the peaks, such as the pvalue below a certain threshold, are rather prone to false discovery. The dependence structure of the markers across the genome should be exploited to gather strength from individual markers into an association analysis of an

extended genomic region. Haplotype analysis provides one option for this kind of analysis, but is restricted to small haplotype blocks with low recombination. A more generalized multiple gene association analysis will reduce the high dimensionality of the association analysis, as well as leading to inferences from the association analysis, which are more meaningful to the medical community. We propose an analysis of a set of multiple markers to get a combined analysis of one or multiple genes that are correlated through location or pathways. Performance of the methods will be presented with a simulation study.

125

A novel methodology for screening and replication in the same dataset for family-based designs using affection status

A.J. Murphy(1) and C. Lange(1,2)
(1) Dept. of Biostatistics, Harvard School of Public Health, USA, (2) Channing Laboratory, Harvard Medical School, USA

Recently, Van Steen et al. (2005) introduced a two stage-testing strategy for family-based designs that substantially outperforms standard methodology, e.g. false-discovery rate (FDR, Benjamini and Hochberg, 1995), making genome-wide association testing feasible, even for relatively moderate sample sizes. Their approach is built upon the conditional mean model (Lange et al., 2003), and therefore requires variation in the phenotype or trait to be tested for association. However, in most family-based studies, affection status is the primary phenotype of interest. Typically, only affected probands are recruited, making an application of the testing strategy by Van Steen et al. (2005) impossible. Here, we propose a completely novel approach that can be applied in such situations to circumvent the multiple testing problem. Our method has the same advantages as the approach by Van Steen et al. (2005), but does not require any variation in the phenotype or specification of a phenotypic mean model. We assess power of our screening method by analytical power calculations and simulation studies. Its practical importance is illustrated by an application to a 100K-scan of the Framingham Heart Study.

References:
Benjamini Y and Hochberg, Y (1995). J. R. Statist. Soc. B 57:289–300.
Lange C, et al. (2003). Am J Hum Genet 73:801–11.
Van Steen K, et al. (2005). Nature Genetics, 37(7):683–691.

126

Assessing sensitivity and percent agreement of posterior subpopulation assignment across clustering methods using both a multiethnic and Caucasian-based sample

K.K. Nicodemus(1), M.F. Egan(1), J. Meyers(1), D. Rujescu(2), D.R. Weinberger(1)
(1) Genes, Cognition & Psychosis Program, CBDB, NIMH, NIH, Bethesda, MD, (2) Department of Psychiatry, Ludwig Maximilians University, Munich, Germany

Case-control studies of genotypes and their association with disease status are a popular and powerful method to detect genetic association. When population substructure exists within a study sample, results can be spurious. Several new methods have been developed to control for population substructure in the design and analysis phases of case-control genetic association studies. However, in order to control for population substructure, sampled individuals must be correctly assigned to subpopulations. Three clustering methods (STRUCTURE, L-POP, GENELAND) have been proposed to assign individuals to subpopulations based on genotype data; one method uses spatial sampling information to help cluster individuals into homogeneous subpopulations. We sought to test the sensitivity of these methods and quantify percent agreement of posterior subpopulation assignment using a multiethnic sample (Caucasians, African-Americans, Eastern Asians) and also within a Caucasian-only sample of individuals from the eastern coast of the United States and Germany. Preliminary results showed disagreement across methods in estimating the number of subpopulations in the multiethnic sample; in addition, posterior subpopulation assignment varied for certain individuals. Results from several models within each method will be presented, along with results varying the number of assumed subpopulations.

127

Polymorphisms in PAH metabolizing and conjugating genes, interactions with smoking and prostate cancer risk

N.L. Nock(1), B.A. Rybicki(2), X. Liu(3), G. Casey(4), J.S. Witte(3)
(1) Dept. of Epi & Biostat, Case Western Reserve Univ., USA, (2) Dept. of Epi, Henry Ford Health System, USA, (3) Dept. of Epi & Biostat, UC San Francisco, USA, (4) Dept. Cancer Biology, Cleveland Clinic Foundation, USA

Polycyclic aromatic hydrocarbons (PAHs) are found in cigarette smoke, which has been implicated equivocally in prostate cancer. PAHs, however, require metabolic activation and subsequent binding to DNA to exert their carcinogenic action; therefore, the ambivalence may be explained, in part, by inter-individual variation in PAH metabolism and conjugation. We investigated polymorphisms in genes that metabolize (CYP1A1 Ile462Val, CYP1B1 Ile432Val and mEPHx His139Arg) and detoxify (GSTM1 +/−, GSTT1 +/−, GSTP1 Ile105Val and Ala114Val) PAHs in a family-based case-control study with 439 prostate cancer cases and 479 brother controls. Among men with less aggressive disease (Gleason score (GS) <7 and clinical tumor stage (CTS) $<$T2c), the GSTM1 null genotype compared to the functional genotype was associated with an increased risk of prostate cancer (OR=2.01; 95% CI:1.06-3.81). However, in men with more aggressive disease (GS ? 7 or CTS ? T2c), the GSTM1 null genotype decreased prostate cancer risk (OR=0.60; 95% CI:0.36-0.99). We also observed a statistically significant multiplicative interaction between the GSTM1 polymorphism and smoking (p=0.029), which was enhanced when restricting the analysis to Caucasian men with more aggressive disease

(p=0.005). Interestingly, GSTM1 is the most effective GST in inhibiting DNA adduct formation by the ''ultimate'' PAH carcinogen, (+) -*anti*-BPDE.

128

Genetic architecture of adiposity: Evidence for epistatic interactions on chromosomes 7 and 13 in the NHLBI Family Heart Study (FHS)

KE North(1), RH Myers(2), M Feitosa(3), W Tang(4), C Lewis(5), P Hopkins(6), J Hixson(7), L Wagenknecht(8), A Kraja(3), IB Borecki(3)
(1) Epi, UNC, (2) Neurogen, BU, (3) Bios Div, Wash U Med, (4) Epi Div, U Minn, (5) Prev Med, UAB, (6) CVD Genetics, U Utah, (7) SPH, UT (8) PHS, Wake Forest U Med School

There is growing evidence that complex interactions among genetic factors influence the pathogenesis of obesity. In previous work in the Caucasian sample of the FHS, two QTLs for body mass index (BMI) were localized, on chromosome 7q32 (LOD=4.7) and on chromosome 13q14 (LOD=3.2). To refine the localization of these putative QTLs and further describe the genetic architecture of BMI, additive interaction effects between chromosome 7q and 13q variants on BMI were assessed.
FHS families were recruited from five US study centers. The Mammalian Genotyping Service typed a total of 404 microsatellite markers. Among 3,104 Caucasians comprising 8,301 relative pairs, variance component linkage analysis extended to include additive interaction effects both in the full sample and stratified by primary evidence for linkage was performed using SOLAR. Using marker allele frequencies derived from pedigree founders, multipoint IBD sharing was estimated using GeneHunter. BMI was adjusted for study center, age, age2, and age3, within sex. Significant evidence for an additive QTL interaction at 7q32 and 13q14 ($sigma_E{}^2=0.30\pm0.08$, P=0.00001) was detected. When stratifying by primary evidence for linkage, evidence for alternate nearby peaks on chromosomes 7 and 13 were localized, a finding consistent with multiple genes at each locus. Although these analyses did not narrow the focal regions of chromosomes 7 and 13 as hoped, these findings demonstrate strong main and epistatic effects on BMI. These findings will be further tested in the African American FHS sample, once genotyping has been completed, and may be important for future interpretation of the genetic architecture, once likely genes are identified.

129

The power of the molecular haplotype for the study of complex diseases in genetic isolates

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

Genetic isolates such as the Old Order Amish offer many advantages for studying complex diseases, including large sibship sizes, extensive and accurate genealogical records, homogeneous lifestyle and a relatively small number of founder haplotypes. Exploiting the rich recombination history of large 10–14 generation pedigrees for linkage and linkage disequilibrium has been computationally intractable given the number of individuals with missing data and the number of loops. The multipoint likelihood problem is intractable due to two sources of combinatorial complexity. First, as the number of heterozygous genotypes increases the number of possible phases increases by a factor of 2. Second, the multiple generations of missing data in the top portion of the pedigree that connects the oldest individuals with genotype data. This portion is used to compute the probability that alleles in the oldest individuals are shared identical by descent from a common ancestor. The molecular haplotype, however, has the power to completely eliminate both sources combinatorial complexity. First, molecular haplotypes trivially eliminate phase ambiguity in individuals with genotype data. Second, dense SNP genotyping on a haploid background allows regions of identity by descent to be inferred directly by modeling the probability of the length of consecutive of allele matches on each haploid. Thus we eliminate the top portion in the probability calculation. We present our statistical models, new likelihood algorithms and current progress on medium-throughput methods to separate chromosomes for molecular haplotyping.

130

Linkage disequilibrium analysis of the recombination hotspot located upstream region of the beta-globin gene in Japanese population

J. Ohashi, I. Naka, K. Tokunaga
Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Recombination hotspot located upstream of the beta-globin (HBB) gene has been intensively studied by pedigree analysis, sperm typing approach, and linkage disequilibrium (LD) analysis so far. Our previous study based on patterns of linkage disequilibrium (LD) among 44 biallelic markers in Thai suggested a putative recombination hotspot spanning 2.0 kb in length at the upstream of the HBB gene. In this study, to localize the recombination hotspot more detail, we analyzed several polymorphic sites in the putative recombination hotspot for 31 Japanese. The following linkage disequilibrium analysis suggested that this hotspot consisted of approximately 700 bp. Although confirmatory studies need to be conducted, our results shed light on the mechanism of recombination hotspot in the human genome.

131

Heritability of cognitive traits in a young genetically isolated Dutch population

L Pardo(1), P Sanchez(1), YS Aulchenko(1), I de Koning(2), K Sleegers(1), JC van Swieten(2), BA Oostra(1), CM van Duijn(1)
(1) Genetic Epidemiology Unit, Departments of Epidemiology & Biostatistics and Clinical Genetics, Erasmus

Medical Center Rotterdam, The Netherlands, (2) Department of Neurology, Erasmus Medical Center Rotterdam, The Netherlands

We studied the heritability of cognitive function in 2900 participants (ages from 18 to 89 years) of the Erasmus Rucphen family (ERF) study. This family-based study was conducted in a young genetically isolated Dutch population. We evaluated memory, learning, visuospatial and executive functions using 5 neuropsychological tests scores (Auditory verbal learning〉AVLT, trail making test-TMT A and B, Stroop color, WAIS -III block design and verbal fluency tests). Additive heritability was estimated for each test score adjusted for individual characteristics including level of inbreeding. To analyze the effect of the APOE gene, we also estimated heritability of the tests, adjusting the models for APOE*4 allele in a sample of 1000 subjects. We found significant additive heritability in all tests with estimates that ranged from 12–30% (AVLT-derived scores) to 40% (WAIS 〉III block design, verbal fluency tests). The effect of APOE was significant in TMT-B explaining about 2% of the total variance and 11% of genetic variance. Inbreeding effect was significant in TMT, semantic fluency and Stroop color tests scores. Our findings suggest that genes explain a substantial part of cognitive functioning in which APOE gene may have a contribution. The significant inbreeding effect indicates that some cognitive traits may be mediated by recessive mutations.

132

Evidence for Association of Chromosomal Regions 2q34-37, 3p25-26, 5q31, and 5q35-qter with Nonsyndromic Oral Clefts

J.W. Park(1,2), R. Ingersoll(1,2), J.B. Hetmanski(1), I. McIntosh(2), M.D. Fallin(1), A.F. Scott(2), T.H. Beaty(1)
(1) Dept. Epi and (2) Inst. Genet Med, Johns Hopkins Univ., USA

Multiple levels of etiologic heterogeneity may control the risk to nonsyndromic oral clefts, a common but heterogeneous group of birth defects. Different genes (locus heterogeneity) and different mutations at one gene (allelic heterogeneity) may affect the risk to oral clefts, and these genes may interact with one another and/or with environmental risk factors. To identify genes or chromosomal regions involved in nonsyndromic oral clefts, we performed transmission disequilibrium tests (TDT) with fine mapping panels of 490, 229, 157, and 121 single nucleotide polymorphism (SNP) markers in chromosomal regions 2q34-37, 3p25-26, 5q31, and 5q35-qter, respectively, in 58 case-parent trios from Maryland. A number of suggestive regions in each chromosomal region (26, 9, 5, and 6 regions, respectively) were identified from either individual markers or sliding windows of haplotypes consisting of 2 to 5 SNPs. The most statistically significant evidence for linkage in the presence of disequilibrium occurred with SNP markers rs1991161 (P=0.0031), rs10980 (P=0.001), rs968057 (P=0.0004), rs329304 (P=0.0071), and rs1643811 (P=0.0031). These SNPs are located in the TNS1 (2q35-36), MRPL44 (2q36.1), CNTN4 (3p25-26), PHF15 (5q31.1), and ADAMTS2 (5qter) genes, respectively. While these results are consistent with the complex etiologic heterogeneity of nonsyndromic oral clefts, a number of potential new candidate genes identified through this analysis warrant future study.

133

High Throughput SNP and Expression Analyses of Candidate Genes for Nonsyndromic Oral Clefts

JW Park(1,2), J Cai(2), I McIntosh(2), JB Hetmanski(1), M Vekemans(3), M Lovett(4), AF Scott(2), EW Jabs(2), TH Beaty(1)
(1) Dept Epi & (2) Inst Genet Med, Johns Hopkins Univ, USA, (3) Dept Genet, Hôpital Necker Enfants Malades, France, (4) Dept Genet, Washington Univ, USA

Current research suggests that multiple genes and environmental factors influence risk to nonsyndromic oral clefts. With recent advances in high-throughput genotyping technology, it is possible to test many candidate genes simultaneously. We present findings from analysis of single nucleotide polymorphism (SNP) markers in 64 candidate genes genotyped using the BeadArray in 58 case-parent trios from Maryland to illustrate how multiple markers in multiple genes can be analyzed using family based association tests. To assess whether these genes were expressed in human craniofacial structures relevant to normal palate and lip development, we analyzed data from the Craniofacial & Oral Gene Expression Network consortium and searched other public databases. Thirteen candidate genes showed significant evidence of linkage in the presence of disequilibrium, and ten of these were expressed in relevant embryonic tissues: SP100, MLPH, HDAC4, LEF1, C6orf105, CD44, ALX4, ZNF202, CRHR1, and MAPT. Three genes showing statistically significant evidence of association, but no evidence of expression included: ADH1C, SCN3B, and IMP5. Many of the candidate genes reported here have not been previously studied for oral clefts, and this approach demonstrates how statistical evidence on large numbers of SNP markers can be combined with gene expression data in a new strategy for identifying candidate genes for complex disorders.

134

Mitochondrial (mtDNA) Haplogroups and Complications of Type 1 Diabetes (T1D)

A.D. Paterson(2,4), B. Bharaj(2), S. Khalid(2), L. Mirea(1,4), S.B. Bull(1,4), A.P. Boright(3), The Diabetes Control & Complications Trial/Epidemiology of Diabetes Interventions & Complications (DCCT/EDIC) Study Research Group
(1) Samuel Lunenfeld Research Inst, Mt Sinai Hospital, Toronto, (2) Genetics & Genomic Biology, Hospital for Sick Children, Toronto, (3) Dept Medicine, University Health Network, Toronto, (4) Dept Public Health Sciences, Univ Toronto, Canada

Variations in the genes involved in oxidative phosphorylation may confer risk for the development of diabetic complications. We hypothesized that common mtDNA haplogroups are associated with diabetes complications, specifically, nephropathy and retinopathy. The subjects were 1,368 white probands with T1D from the DCCT/EDIC, a long-term prospective clinic trial and follow-up of intensive diabetes treatment. We genotyped 6 mtDNA SNPs which distinguish haplogroups >5% in Caucasians (H/H2/H3/H4, H1, J1/J3, U, K, T, and others). Association of single SNPs and haplogroups with diabetic complications was tested using multivariate Cox Proportional Hazards models, including appropriate covariates. There were no significant associations with nephropathy outcomes. For the development of macular edema we observed higher hazards for J1/J3, H1, and K haplogroups (HR (95%CI) 1.92 (0.98-3.75), 1.68 (0.99-2.85), and 1.87 (1.02-3.43) respectively), compared to the most common haplogroup, but these borderline associations (p=0.05) were non-significant after a Bonferroni adjustment for multiple comparisons (n=6). No significant evidence for association between mtDNA haplogroups and type 1 diabetic nephropathy or retinopathy were observed in this study.

135

Segregation analysis of Bone Mineral Density in European pedigrees selected through male osteoporotic probands

C Pelat(1), I Van Pottelbergh(2), M Cohen-Solal(3), A Ostertag(3), JM Kaufman(2), M Martinez(1,4), MC de Vernejoul(3)
(1) INSERM EMI0006, France, (2) End. Dept. Gent Univ., Belgium, (3) INSERM U606, France, (4) Northwestern Univ. Evanston, USA

Osteoporosis is mainly characterized by low Bone Mineral Density (BMD). BMD is a highly heritable trait but its underlying genetic mechanisms remain unclear. To this aim, we conducted segregation analyses under the regressive models using Finesse (O–Connell et al., 1998 Genet Epidemiol 15:521) in 100 pedigrees (424 measured subjects) from France and Belgium, selected through a male osteoporotic subject. BMD values at Lumbar Spine (LS) and Femoral Neck (FN) were adjusted on age, sex and Body Mass Index (BMI) and segregation analysis was conducted on the residuals, involving or not Gene-Covariate (GxC) interactions. Considering models without GxC interaction, no significant effect of a major factor on BMD-LS was detected. GxAge interaction effects were found significant (p=0.02). Models including this term provided evidence for a major factor (p=0.0021) but the Mendelian transmission model was rejected. For BMD-FN, models without interaction term showed significant effects of a major factor (p=1.7E-7) but rejected the involvement of a single Mendelian gene. Significant GxBMI interaction effects were identified (p=0.045). Including this term, the best segregation model involved the effects of both residual correlations (rpo=rss=0.34) and a Mendelian major gene accounting for 41% of BMD-FN variance. These results highlight the site-specific etiology of BMD and the impact of considering GxC interaction on genetic analysis.

136

Heritability estimates for phenotypes related to obesity, cardiovascular disease and type 2 diabetes mellitus – The CANHR Study

R Plaetke(1), A Goropashnaya(1), G Antunez de Mayolo(1), S Hutchinson(1), Y Wang(1), JR Herron(1), DB Allison(2), HK Tiwari(2), GV Mohatt(1), AG Comuzzie(3), BB Boyer(1)
(1) CANHR, University of Alaska, USA, (2) University of Alabama, USA, (3) Southwest Foundation for Biomedical Research, USA

We have ascertained 37 pedigrees including 648 study participants (Yup–ik Eskimos from Southwest Alaska; 55% women; age:14–94 years). One pedigree consists of 507 (78%) participants; the remaining pedigrees have on average 3.9 individuals (range:3–22). Narrow-sense heritabilities were estimated based on a variance component approach (SOLAR). Results for h^2/~variance due to covariates~ ($\pm SE(h^2)$; P-value; covariates included in the analysis):(1) **Obesity related phenotypes:** leptin: 0.37/0.62 ($\pm$0.1; 1.2*10-5; sex, age, bmi, insulin), adiponectin: 0.66/0.18 ($\pm$0.09; 3.8*10-14; sex, age,%bodyfat). (2) **CV related risk factors**: ldl: 0.37/0.18 ($\pm$0.09, 3.7*10-6; sex, age, bmi), hdl: 0.39/0.36 ($\pm$0.09,2*10-7; sex, age, fatmass, insulin, adiponectin), cholesterol: 0.44/0.26 ($\pm$0.1, 2.1*10-6; sex, age, bmi), triglycerides: 0.33/0.17 ($\pm$0.09, 7.5*10-6; sex, glucose, insulin, adiponectin). (3) **Diabetes related traits**: fasting glucose: 0.21/0.15 ($\pm$0.09, 3.9*10-3; sex, age, triglycerides), plasma insulin: 0.03/0.36 ($\pm$0.08, 0.3; age, bmi, hdl, triglycerides, leptin), HbA1c: 0.24/0.46 ($\pm$0.09; 6.1*10-4; age, glucose, insulin). Heritability estimates are mainly intermediate indicating that genes may have a substantial additive effect on these traits and may be identified in a genome wide linkage analysis.

137

Meta-Analyses of Correlated Genomic Scans

M.A. Province
Washington Univ Med School, St. Louis, MO, USA

Meta-analysis is an attractive way to post-hoc combine evidence across genomic scans. As denser genomic scans are generated, requiring more severe corrections for multiple comparisons, harnessing the combined power of meta-analysis can be a cost effective way to overcome the inevitable individual study power losses. Several linkage consortia have already been formed (e.g. the NHLBI GENELINK, the Breast Cancer Linkage Consortium, etc.). But there can be hidden statistical dependencies between component studies which challenge the meta-analysis iid assumptions (e.g. the Framingham Heart Study was one recruitment pool for both the NHLBI Family Heart Study and the HyperGEN study, so one cannot simply meta-analyze data from all 3 studies as

though they were independent). Ignoring such correlations can artificially inflate meta-analysis evidence, creating false-positive signals. We present a simple method to meta-analyze potentially correlated genomic scans, which can be used with either linkage or association data. The idea is to use the fact that the vast majority of the loci are under the null hypothesis for any given phenotype. Thus, we can estimate and use the empirical correlation matrix between scans as weights in a meta-analysis statistic to combine evidence properly. When scans turn out to be independent after all, the method reduces to the traditional sum of independent Z-scores statistic, so it can be safely used in all cases, whether correlated or not. We show simulation results demonstrating the method–s utility and statistical power and type I error, as well as results combining several real genome scans.

138

Genetic epidemiology of the Mitsuda reaction in leprosy sibships

B Ranque(1), A Alcaïs(1), NV Thuc(2), VH Thai(2) E Schurr(3) and L Abel(1)
(1) INSERM U550, Paris, France, (2) Dermatology Hospital, Ho Chi Minh City, Vietnam, (3) McGill University, Montréal, Canada

Leprosy is a chronic infectious disease caused by Mycobacterium leprae still affecting 600,000 new persons each year. The Mitsuda reaction is a delayed skin granulomatous reaction elicited by intradermal injection of lepromin that reflects cellular immune response to M. leprae.
We conducted a segregation analysis of quantitative Mitsuda reaction in 168 nuclear Vietnamese families ascertained through leprosy patients, using regressive models. Different analyses provided strong evidence ($p < 10\text{-}9$) for a recessive major gene controlling the Mitsuda reaction independently of the leprosy status, with the recessive allele ($q=0.35$) predisposing to high Mitsuda values.
We then performed a genome-wide scan of the quantitative Mitsuda reaction in a subsample of 19 families (110 children), using both model-based (the model being estimated by the segregation analysis) and model-free (quantitative Maximum Likelihood Binomial) analyses. We found some evidence for linkage (lod score > 1.1) to 5 chromosomal regions, which were saturated by 30 markers. The model-free fine mapping found suggestive linkage to the NRAMP1 gene (lod score=2.1). In addition, both model-based and model-free analyses identified a novel chromosomal region. Interestingly, a negative correlation between the family-specific lod scores at the two regions suggested an interaction between them.
This genetic dissection of the Mitsuda reaction is expected to cast novel lights into the mechanisms of granuloma formation in leprosy.

139

A linkage analysis test statistic which models 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, USA, 2) Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA

Misspecified relationships can lead to reduced power for linkage analysis. Possible solutions are to discard the erroneous data or to use the most likely alternative pedigree structures or to statistically model the uncertainty by weighting over the possibilities. Consider the situation where we have collected affected relative pair (ARP) data and carried out a genome-wide scan for linkage. Our method is an extension of the affected relative pair LOD score analysis approach of Cordell et al. (2000). In our extension, we statistically model any relationship uncertainty via weights, the weights being the conditional probability of a true relationship type given the apparent one and the genome-wide marker data. The estimates of the weights can be used to fit a likelihood model to the IBD sharing among ARPs. To assess our method and to compare it to the maximum likelihood statistic of Cordell et al. and a nonparametric linkage statistic from MERLIN, we are performing a simulation study. Each simulated dataset consists of 300 pedigrees with 660 affected apparently full sib pairs having several underlying true relationship types with marker data having 367 markers and a disease locus on chromosome 10 with a dominant reduced penetrance mode of inheritance. For each simulation, we also construct both the true relationship structure and the pedigree after discarding the individuals with erroneous relationships. Preliminary results indicate our method perfoms well: Based on 960 replicates using genome-wide empirical 0.01 significance thresholds, power at the disease locus is: 60% (our method); 77% (MERLIN, true structure) and 53% (MERLIN, discarded structure).

140

Association Tests based on Haplotype Similarity in Case-Parent Trio Studies

G.A. Satten(1), A.S. Allen(2)
(1) CDC, Atlanta, GA, (2) Dept. of Bioinformatics and Biostatistics, Duke University, Durham NC

Association tests based on the identity length contrast for case-parent trios have been proposed that detect excess sharing among all transmitted vs. all untransmitted haplotypes (Bourgain et al. 2000 Ann Hum Genet) or equivalently that use a TDT that ''scores'' each haplotype by the amount of sharing between it and all parental haplotypes (Zhang et al. AJHG 2003). Here we take a different approach by modeling an increase in disease risk when transmitted haplotypes are similar to one or more fixed haplotypes we call archetypes. Testing is carried out using the haplotype regression approach of Allen, Satten and Tsiatis (Biometrika, 2005). Similarity can be measured by identity length contrast or by other measures such as parsimony. We choose the archetype haplotype(s) by examining patterns of haplotype similarity among parental haplotypes, or by using multidimensional scaling.

A simple choice is to average the similarity between each fixed haplotype and all parental haplotypes, and then select those haplotypes having the greatest similarity on average to the parental haplotypes as the archetypes. Model selection can be used to infer the best number of archetypes to use. Preliminary simulation results of populations with a recently-introduced disease mutation indicate that this approach may have more power than existing methods, and more power than regression models based on the risk of specific haplotypes. Our approach can be used for both quantitative and qualitative traits, accounts for phase ambiguity, and allows for missing genotype information.

141

Statistical analysis of gene-environmental interactions on an additive scale in case-control studies

H. Schäfer, H.-H. Müller, T.T. Nguyen
Institute of Medical Biometry and Epidemiology, University of Marburg, Germany

We consider a genetic epidemiological case-control study for testing the association between a disease D and a genetic variant G (0/1), in which additionally an environmental exposition factor E (0/1) is studied. By usual logistic regression, G*E interactions can be studied on a multiplicative scale for the odds. For medical decision making purposes, it may be more relevant to determine G*E interactions on an additive scale as given by the term d=P(D | G1,E1) - P(D | G1,E0) - P(D | G0,E1) + P(D | G0,E0). The interaction term d can be understood as a utility index measuring the utility of genotype-dependent reduction of environmental risk exposition as compared to random reduction of environmental risk exposition.
With additional information on the individual pre-test risk for the disease, or, on the population level, on the prevalence of the disease, estimates for d can be obtained from case control data in an obvious way. We derive formulae for the variance of this estimate as well as confidence intervals for d and a statistical test for d=0. We also discuss a tentative ''rare disease'' solution, i.e., P(D) - >0. We apply the method to published case-control data on factor V Leiden mutation and oral contraceptives as risk factors for deep venous thrombosis. There was a debate on the relevance of factor V Leiden mutation for medical decision making in the literature (Vandenbroucke et al., Lancet 1994, 344:1453–1457; Westhoff, Lancet 1996, 347:396; Bloemenkamp et al., Lancet 1996, 347:396–397).

142

U-statistics for Testing the Association of Genotype Similarity with Trait Similarity: Methods for Quantitative and Censored Traits

D.J. Schaid, J.P. Sinnwell, S.N. Thibodeau
Mayo Clinic College of Medicine, Rochester, MN

Nonparametric U-statistics have been used to test the association of multiple genes with case/control status, as well as to evaluate the association of the similarity of haplotypes with the similarity of a trait (including quantitative traits); recently, Beckmann et al. showed that Mantel statistics for space-time clustering, a type of U-statistic, can be used for quantitative traits. Although this eases computation of a permutational variance of the test statistic, simulations show that the test statistic fails to achieve the asymptotic standard normal distribution for reasonable sample sizes. Without knowing the null distribution of the test statistic, it is difficult to evaluate the power of competing kernels. Further analysis illustrates that the cause of failure can be due to the type of kernel used to measure either genetic similarity or phenotypic similarity. In fact, the standard use of trait covariance as the phenotype kernel is a major cause of failure. However, use of alternative asymptotic methods allows derivation of a much improved approximation of the distribution of the test statistic, hence avoiding the need to perform permutations to determine statistical significance. More importantly, the theory of U-statistics provides guidance on the choice of kernels, as well as their covariances when combining U-statistics across multiple genetic markers to devise a global test statistic. We also illustrate extensions to evaluate the association of genetic similarity with survival time similarity, allowing for censored times.

143

Mexican-American Genetic Markers in Alzheimer Disease

R.S. Schiffer, X. Yin, E. Drigalenko
Dept. of Neuropsychiatry, Texas Tech University Health Sciences Center, Lubbock, TX

The burdens of the diseases of late life in the Twenty First Century, according to the World Health Organization, are likely to be overwhelmingly neuropsychiatric in nature. Foremost among these diseases will be Alzheimer disease. The effects of ethnic group membership upon risk factors, disease expression, diagnosis, and treatment is of increasing interest for such diseases as develop slowly across the lifespan. In the United States, the populations deriving in part from the eight Southwestern Indian tribes constitute the ''Mexican-Americans'' ethnic group. Very little is presently known about Alzheimer disease in this population. We are performing an initial genetic epidemiologic comparison between Mexican-American Alzheimer disease subjects and matched Anglo Alzheimer subjects. We are evaluating the joint distribution of allele frequencies for candidate genes for Alzheimer disease. As candidate genes, we took genes that have been found associated with the disease (APP, PSEN1, PSEN2, APOE, TFCP2, PLAU, UBQLN1, A2M, CST3, TF, HFE, ACE) and added genes from the amyloid metabolic pathway. Alleles frequencies of the candidate genes are calculating for Anglos and Mexican-Americans, and between-group comparisons is made. We also study first-degree relatives of our subjects, to gain some indication of linkage for these genetic markers and Alzheimer disease that might exist in our populations. From these initial data, we will hope to develop leads for

future studies concerning genetic and other risk factors for Alzheimer disease within the Southwestern Mexican-American population of the United States.

144

Variance components models for ordinal family data: Baldness and the androgen receptor gene

K.J. Scurrah(1,2), J.A. Ellis(1), J.E. Cobb(1), J.L. Hopper(2), S.B. Harrap(1)
(1) Department of Physiology and (2) Centre for Genetic Epidemiology, University of Melbourne, Melbourne, VIC, Australia

Variance components models are frequently used in genetic epidemiology to assess evidence for genetic effects on a trait, and they may also be used for association studies of family data. However, these models become very difficult to fit when phenotypes are not normally distributed. We have previously described methods for fitting these models to binary and censored survival time data obtained from nuclear and extended families (1), using Markov chain Monte Carlo techniques as implemented in the WinBUGS software package. We have now extended these models to allow analysis of ordinal categorical data obtained from families. The models are used to investigate associations between male pattern baldness (which may be categorised as none, frontal or vertex only, and frontal and vertex) and a polymorphism in the androgen receptor gene, in 1200 males from the Victorian Family Heart Study. Inclusion of data from the large variety of relative pair types in the study (brother, father-son, uncle-nephew, MZ and DZ twins, and sons of MZ and DZ twins) increased power and provided substantially more information than the initial case-control study of this polymorphism, which used unrelated individuals from the same study.

(1) Genetic Epidemiology 2000;19:127–141

145

New Association Tests based on Haplotype Similarity

Q. Sha, H.-S. Chen, S. Zhang
Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931

The difference of haplotype distributions in affected and unaffected individuals is sound evidence for a disease-liability mutation in the region. Tests for differentiation of haplotype distributions often take the form of either Pearson's chi-squared statistic or tests based on the similarity among haplotypes in the different populations. Under a case-control design, haplotype-similarity based methods essentially compare the average similarity among cases with the average similarity among controls. However, with a complex trait, it seems plausible that similarity within affected individuals and similarity within normal individuals could be similar, but with different haplotype distribution in each, e.g. controls could be more likely to share protective haplotypes. In this article, we propose new association tests based on haplotype similarity. These new tests compare the average similarity within cases and within controls with the average similarity between cases and controls. These methods can be applied to either phase-known or phase-unknown data. We compare the power of the proposed methods with Pearson's chi-squared test and the existing similarity-based tests by simulation studies under a variety of scenarios. The simulation results show that, in most cases, the new proposed methods are more powerful than both Pearson's chi-squared test and existing similarity-based tests. Only in some extreme cases Pearson's chi-squared test and existing similarity-based tests may be slightly more powerful than the new proposed methods.

146

Mendelian Randomisation and Instrumental Variables for Causal Inference

NA Sheehan(1) and V Didelez(2)
(1) Dept. Health Sciences and Dept. Genetics, University of Leicester, UK, (2) Dept. Statistical Science, University College London, UK

In epidemiological research, the effect of a potentially modifiable phenotype or ''exposure'' on a ''disease'' is often of public health interest. Inferences on this effect can be distorted in the presence of confounders affecting both phenotype and disease. Issues of confounding require causal rather than associational arguments. Mendelian randomisation (see Davey Smith & Ebrahim, 2003, for example) is a method for deriving unconfounded estimates of such causal relationships and basically exploits the fact that a gene known to affect the phenotype can often be reasonably assumed not to be itself associated with any confounding factors and thus has an indirect effect on the disease. It is well known in the economics and causal literature (e.g. Pearl, 2000) that these properties define an instrumental variable but they are minimal in the sense that they only permit unique identification of the causal effect of the phenotype on the disease status in the presence of additional and fairly strong assumptions. These assumptions relate to the distributions of the variables e.g. multivariate normality, and the nature of the dependencies between them, e.g. linear. We explore these assumptions in the context of standard epidemiological applications and make some suggestions as to when these may, or may not, be relaxed. The ideas are illustrated using directed acyclic graphs with interventions.

Davey Smith, G and S Ebrahim. (2003). International Journal of Epidemiology 32:1–22.
Pearl, J. (2000). Causality. Cambridge University Press.

147

Estimating Identity by Descent Using Monte-Carlo Markov Chain Methods

Sanjay Shete, E. Warwick Daw, Jianzhong Ma, Yue Lu, Christopher I. Amos
Department of Epidemiology, U.T. M.D. Anderson Cancer Center

For large or complex pedigrees in which multiple genetic markers have been genotyped, exact computation of identity by descent (ibd) conditional on marker data is not computationally tractable. Therefore, Monte-Carlo Markov Chain (MCMC) procedures have been developed that can provide estimates of ibd sharing, conditional on the marker data and have been implemented in the programs LOKI and SIMWALK 2. Both of these programs also calculate identity by descent sharing according to parental origins of alleles, but using LOKI for this purpose required modification to the output files. We compared the results from applying these estimation procedures to the exact method provided by the GENIBD subroutine of SAGE. Results of applying LOKI and SIMWALK 2 to five families that have been studied for genetic linkage analysis of Wilms Tumor showed excellent correlation (rho=0.990) between the MCMC programs. However, we have noted that there are some pairs for which the ibd sharing for either program is uninformative (i.e. it equals the sharing given by the relationship of the pair) for one program but not for the other. Comparison in small families between results using the MCMC programs versus GENIBD indicated a slightly higher correlation between results using LOKI (rho=0.99996) compared with SIMWALK 2 (rho=0.99688). LOKI appeared to provide more reliable results when analyzing very tightly linked loci (recombination fraction < 0.01), while results of the two approaches were virtually identical for more sparsely located markers (recombination fraction > 0.02).

148

Semiparametric Maximum Likelihood Inference of Disease Associations With a Genetic Factor and Independent Continuous Attribute in a Case-Control Study

J.-H. Shin(1), L. Bekris(2,3), A. Lernmark(2), B. McNeney(1), J. Graham(1)
(1) Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada, (2) R.H. Williams Laboratory, University of Washington, (3) Environmental Health, University of Washington, Seattle, USA

In case-control studies, covariate information is often collected on a genetic factor and a nongenetic attribute which can be assumed to occur independently of the genetic factor in the underlying population. Under this independence assumption, we develop a maximum likelihood method which can more efficiently estimate multiplicative interaction between a genetic factor and nongenetic attribute for a rare disease. We apply our method to a case-control study of age-specific associations between type 1 diabetes and a variant of the glutamate-cysteine ligase catalytic subunit. The results are compared to those from a logistic regression analysis, which does not assume independence.

149

Correction for Asymptotic P-Values in Model-Free Linkage Analysis

Moumita Sinha, Yeunjoo Song, Robert C. Elston, Jane M. Olson, Katrina A.B. Goddard
Dept of Epidemiol & Biostat, Case Western Reserve University, Cleveland, OH

Model-free linkage analysis using the conditional logistic model provides unreliable asymptotic p-values under the assumption of a simple mixture of chi-square distributions of the test statistic. Permutation tests are often performed to obtain empirical p-values, but this is not possible in this instance because all individuals in the analysis are affected. Alternatively, we can develop regression models to predict more accurate p-values, an approach we consider here. Let P_τ be the empirical p-value, which is the proportion of statistical tests whose lod score under H_0 exceeds a threshold determined by τ, the asymptotic p-value. We obtained the P_τ for simulated data, and compared them with those calculated under the asymptotic distribution. We developed a regression model, based on sample size, the asymptotic p-value, and marker density to derive predicted p-values for single point and multipoint analyses for each of baseline and covariate (1-4) models under 10 different models. To evaluate our predictions, we used another set of simulated data to compare the empirical p-values for the data with those obtained by using the prediction model, which can be referred to as predicted p-values. Under all circumstances the difference between the predicted and the empirical p-values is less than that between the asymptotic and the empirical p-values. Thus the regression models suggested by our analysis provide a more accurate alternative to using asymptotic p-values for model-free linkage analysis using the conditional logistic model.

150

Genetic Map Estimation: A Unified Approach

W.C.L. Stewart, E.A. Thompson
Department of Statistics, University of Washington, Seattle, USA

A unified method for genetic map estimation based on the marker data of individuals of extended families is presented. The method uses Markov chain Monte Carlo Techniques to find the maximum likelihood estimate of the genetic map. It represents a significant advancement in the area of genetic map estimation in that the method can analyze marker data observed at multiple polymorphic linked loci in some individuals of large extended families, with or without the assumption of typing error. In addition, the precision and accuracy of existing genetic map estimates are guaranteed to improve when map estimates are combined over data sets. This should reduce marker model misspecification and improve the inference of any multipoint analysis based on pedigree data.

151

Identification of Disease-Associated SNP Clusters Using a Scan Statistic

Y.V. Sun(1), K.J. Meyers(1), T.H. Mosley(2), E. Boerwinkle(3), I.J. Kullo(4), S.T. Turner(4), S.L.R. Kardia(1)
(1) Dept. of Epidemiology, Univ. of Michigan, (2) Dept. of Medicine, Univ. of Mississippi Medical Center, (3) Dept.

of Human Genetics, Univ. of Texas Health Sciences, (4) Cardiovascular Diseases, Mayo Clinic and Foundation

Scan statistics have been receiving attention as a method for genomic analysis. However, most methods are limited by the assumption of a uniform density between genetic elements. We developed a scan statistic using a compound Poisson process that consider the complex landscape of human genome variations in the identification of gene regions associated with disease. Using hypertensive African-American sibships from the GENOA study, we divided two sibs from each sibship into two samples of 448 unrelated individuals and analyzed patterns of SNP associations with echocardiographic phenotypes to identify SNP regions with replicable effects. We tested 84 SNPs in 6 genes to identify associations with left ventricular relative wall thickness (RWT) and aortic root diameter (ARD). With a p-value cutoff of 0.05, we detected one region in SLC4A5 significantly associated with RWT in both datasets. Two regions, one each in ADD2 and VCAM1, were significantly associated with ARD in both datasets after adjusting for traditional risk factors. We found 7 out of 30 single-SNP tests were significant in the three regions. However, none of the 15 pairs of single SNP p-values were consistently significant. This result suggests that SNP regions detected by the scan statistic may be more robust than single SNPs in disease association studies. This scan statistic can be used to make more consistent region-based inferences about association.

152

Power Calculations for a Genetic-Model Free Method for Linkage Analysis of a Disease Related Trait

H.J. Sung(1), S.J. Finch(1), K.Q. Ye(2) and N.R. Mendell(1)
(1) Dept. of Applied Math and Stat, SUNY at Stony Brook, USA, (2) Albert Einstein College of Medicine, 10461, USA

In this paper we investigate the power of detecting linkage to a disease locus through analysis of traits related to the disease. We propose a family-based gene-model-free linkage statistic. This statistic involves considering the distribution of the number of alleles identical-by-descent with the proband through comparison of disease related trait positive siblings to disease related trait negative siblings. Upon assuming an allele pleiotropic for both disease and disease related trait, the identical-by-descent distribution in these two groups can be obtained numerically as a function of (1) the frequency of the disease/disease related trait allele and (2) the penetrance of each of the 4 disease/disease related trait phenotypes for each genotype and (3) the recombination fraction. We report findings for a range of realistic pleiotropic alleles and sample sizes.

153

Gene Mapping Through Hierarchical Bayesian Models

M.D. Swartz(1,4), M. Kimmel(2), P. Mueller(3), C.I. Amos(4)
(1) Dept. of Statistics, Texas A&M University USA, (2) Dept. of Statistics, Rice University USA, (3) Dept. of Biostat. and Applied Math, U. T. M.D. Anderson Cancer Center, USA, (4) Dept. of Epidemiology, U. T. M. D. Anderson Cancer Center, USA

Hierarchical Bayesian models have the flexibility to model the complexities present in mapping the genes for a complex disease, such as Rheumatoid Arthritis (RA). We present a hierarchical Bayesian model for case-parent triad data that uses a conditional logistic regression likelihood to model the probability of transmission to a diseased child. We define hierarchical distributions on the allele main effects to model the genetic dependencies present in the HLA region of Chromosome 6. We first add a hierarchical level for model selection that accounts for both locus and allele selection. This allows us to cast the problem of identifying genetic loci associated with disease into a problem of Bayesian variable selection. To further improve the model, we include linkage disequilibrium as a covariance structure in the prior for model coefficients. Using simulated data, we found that the hierarchical priors for locus and allele selection and including LD as the covariance structure outperformed using maximum likelihood estimates, and put more posterior probability on the true genes than when excluding the LD from the model for our simulated disease. We also saw promising performance when applying the method to map genetic markers in the HLA region to RA, using 65 triads.

154

Bias in the gene-age interactions in the case-control association studies resulting from unmeasured confounders and competing risks

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

Case-control association studies are often conducted in candidate genes or regions of multifactorial late-onset disease. It is frequently reported that the genetic effects on late-onset diseases are modified by age, however, not being considered competing risks of death or withdrawal and the unmeasured confounders. The author extends previous work concerning the bias associated with competing risks for nested case-control studies, in which cases the relative hazard of the disease for the polymorphism increase or decrease with age and prognostic factors of competing disease are confounders of the relationship between the disease and the polymorphisms. Specifically, the distorting effect of competing risks is illustrated for three methods of control selection; Cumulative sampling, density sampling and case base sampling. The formulas are derived for the bias of the odds ratio and numerical examples show the magnitude of bias when competing risks and confounders cannot be ignored.

155

Multipoint Linkage Analysis Using Combined Panel of SNPs and Microsatellites

B.O. Tayo(1), Y. Liang(2), M. Trevisan(1)
(1) Dept. of Social and Preventive Medicine, SUNY at Buffalo, USA, (2) Dept. of Biostatistics, SUNY at Buffalo, USA

BACKGROUND: With the rapid increase in the number of available SNPs in the dbSNP database, plus the plummeting costs of SNP genotyping through the use of high-throughput technologies, the future and utility of the less abundant microsatellites have began to be questioned. METHODS: As a follow-up to earlier studies comparing these two sets of genetic markers side by side for linkage analysis, we carried out multipoint non-parametric linkage analysis of onset-age of alcoholism in 143 families from the Collaborative Study on the Genetics of Alcoholism using combined panel of SNPs and microsatellites. RESULTS: Our results showed that except for chromosome 1, the locations and linkage signals from the combined panel tended to be closer to those obtained separately from the SNP panel on every chromosome. CONCLUSION: We conclude that the results of this study do not provide evidence of added advantage for using combined panel of SNP and microsatellite in linkage analysis.

156

Evaluation and comparison of gene clustering methods in microarray analysis

A. Thalamuthu(1), I. Mukhopadhyay(1), X. Zheng(1), G.C. Tseng(2)
(1) Department of Human Genetics, University of Pittsburgh, USA (2) Department of Biostatistics, University of Pittsburgh, USA

Microarray technology has been widely applied in biological and clinical studies for simultaneous monitoring of gene expression in thousands of genes. Gene clustering analysis is found useful for discovering groups of correlated genes potentially co-regulated or associated to the disease under investigation. Traditional methods including hierarchical clustering, K-means, PAM, SOM and mixture model-based clustering have been widely used in the literature. To tackle with the nature of high-dimensional, inter-correlated and complex gene network structure in microarray data, resampling based clustering methods are recently proposed. In this paper two recently proposed gene clustering methods, Tight clustering and Penalized K-means, are compared with some of the traditional methods such as hierarchical, K-means, PAM, SOM, and model based clustering for microarray experiments. We propose a modified Rand index that measures similarity of two clustering results with possible scattered genes (i.e. a set noise genes not being clustered). Accuracy and sensitivity analysis of the clustering methods are assessed by simulated data sets and published data sets. Our results show that Tight clustering and Penalized K-means perform better than the traditional clustering methods both for real and simulated data sets. Our analysis provides deeper insight to the complicated gene clustering problem in microarray analysis and can be used as guideline for biologists in daily practice.

157

Recent Developments in Genome-Wide Association Scans

D.C. Thomas, J.D. Buckley, D.V. Conti, W.J. Gauderman, J.P. Lewinger, D.O. Stram
Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

With the advent of affordable high-density SNP genotyping technologies, it has now become feasible to conduct genome-wide association studies using $\sim$500,000 markers. Recent methodologic developments have included the use of multistage sampling designs, haplotype associations with tag-SNPs, and various approaches to the multiple comparisons problem, including testing all possible haplotype windows and all pairwise SNP x SNP interactions. We will briefly review these advances and focus on our discussion on a few of these topics. First, we describe a novel Bayesian approach to prioritizing SNP associations detected in the first stage of a scan of all available markers for testing a subset of them in a new sample of cases and controls, incorporating prior genomic information in a hierarchical model. For gene-gene interactions, we compare exhaustive testing of all possible pairs of SNPs with two-stage approaches based either on testing first for marginal associations or on testing gene-gene association in the combined case-control sample. We describe a unified framework for testing haplotype association and haplotype sharing, which obviates the need for testing all possible haplotype windows. Finally, we describe approaches to population stratification in a genome-wide significance testing framework for multi-stage case-control sampling schemes, discussed in Lewinger et al. (IGES abstract). In general, it appears that the optimal two-stage design typically entails allocating about half the available sample size to stage I with a first-stage $\alpha \sim 0.001$ to yield a genomewide significance.

158

Linkage analysis of sarcoidosis stratified by genetic subpopulations

CL Thompson(1), BA Rybicki(2), MC Iannuzzi(3), RC Elston(1), SK Iyengar(1), C Gray-McGuire(1)
(1) Dept of Epi & Biostats, Case West Res Univ, USA, (2) Dept of Biostats & Research Epi, Henry Ford Health System, USA, (3) Div of Pulmonary, Critical Care & Sleep Med, Mt Sinai School of Med, USA

Sarcoidosis is an autoimmune disorder of unknown etiology. In the US, African-Americans are more commonly and more severely affected than Caucasians. The first genome scan conducted in 229 African-American families indicated linkage to chromosome 5 and showed marginal evidence of linkage to 6 other regions (Ianuzzi et al., Genes Immun 2005). Because of known admixture in African-Americans, we reevaluated these linkages in light of the admixture observed. The population structure of our sample was inferred via the program STRUCTURE, using the 380 microsatellite markers from our genome scan. Evidence of two subpopulations was found, with 7 families estimated to be from subpopulation I, 218 estimated to be from II, and 4 of the families representing a mixture of the two. The linkage analysis was repeated on the families in these two subpopulations. Stratified analysis suggests that several of the regions identified in

the original scan are due to subpopulation I (1p22, 11p15 and 17q21) or II (5p15-13 and 20q13), while others remain significant in both (2p25, 5q11, 5q35 and 9q34). Evidence for a new locus only in subpopulation II (2q37) was also found. This analysis demonstrates the utility of admixture information to identify more homogeneous subsets of families in studies of complex disease. It also provides evidence for multiples genes influencing sarcoidosis and that these genes may be population specific.

159

Multilocus Lod Scores in Large Pedigrees: a New Approach to Combine Exact and Approximate Calculations

L. Tong, E. Thompson
Dept. of Statistics, the Univ. of Washington, USA

To detect the positions of disease loci, LOD scores need to be calculated within a (several) pedigree(s) for a given set of markers at multiple chromosomal positions. Exact LOD score calculations are often impossible when the size of the pedigree and the number of markers are both large. In this case, a Markov Chain Monte Carlo (MCMC) approach is able to provide an approximation. However, the mixing performance, to provide accurate results, within a reasonable amount of time, is always a key issue in these MCMC methods. In this paper, we propose a new approach, which divides a large pedigree into several parts by conditioning on parental haplotypes. We perform exact calculation for the offspring parts where more data are often available, and combine this information to sample the hidden variables for the parental parts. We also improve the parental sampling part using a mixture of several conditional Hidden Markov Chains across loci or meiosis. Our approach is not only more efficient for large pedigree(s) with large number of markers, but also very useful for a looped pedigree, in which case most current methods cannot give satisfactory results.

160

Comparison of Methods for Estimating Admixture Background and Control for Population Stratification Confounding in Admixed Populations

H.-J. Tsai(1), S. Choudhry(1), M. Naqvi(1), W. Rodriguez-Cintron(2), E.G. Burchard(1), E. Ziv(1)
(1) University of California, San Francisco, SF, CA, USA, (2) University of Puerto Rico School of Medicine, San Juan, PR

Population stratification may confound the results of genetic association studies among unrelated individuals from admixed populations. Several methods have been proposed to estimate ancestral information in admixed populations and to adjust for population stratification in genetic studies. We evaluate the performance of three different methods: maximum likelihood estimation, ADMIXMAP and Structure through comprehensive simulation scenarios and real data from Latino Americans participating in a genetic study of asthma. All methods perform similarly on estimating accurate admixture background and controlling for the inflation of type I error rate. The most important factor in determining accuracy of the ancestry estimates and in minimizing type I error rate is the number of markers. We demonstrate that approximately 100 ancestry informative markers (AIMs) are required for the methods to obtain ancestry estimates that have high correlation ($r > 0.9$) with the true ancestry estimates. After accounting for ancestry information in association tests, the excess of type I error rate is controlled at the 5% level when 100 markers are used to estimate ancestry. Using data from Latino subjects, we apply these methods to test the association between body mass index and 44 AIMs. Our works provide practical guidelines for investigators conducting genetic association studies in admixed populations.

161

Use of Covariates in Linkage Analysis: Comparison of Model-Based and Model-Free Methods

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

Covariate inclusion in linkage analysis of complex traits may increase the power to detect linkage. To identify situations with large power gain for a qualitative trait, the effect of covariate inclusion on power and type I error was evaluated. A trait with penetrance determined by a dominant biallelic locus and modified by a quantitative covariate was simulated with G.A.S.P., using various degrees of penetrance, disease allele frequency and covariate effect. For each model, segregation analysis was performed (REGDHUNT) on a large singly ascertained sample to provide a trait model for use in the model-based analysis. The trait and linked (θ=0.01, 0.05) and unlinked markers were generated for 10,000 samples of 300 nuclear families with 4 offspring. Both model-based (LODLINK with/without covariates) and model-free (SIBPAL with/without covariates, MERLIN, ALLEGRO) methods of linkage analysis were used. The results suggest that, when a good estimate of the heredity model is available, lod-score methods are the wisest choice when analyzing a qualitative trait where disease etiology involves covariates. 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. When covariates are involved, use of NPL scores and Kong and Cox LOD analyses with MERLIN and ALLEGRO should be considered last since they do not allow for covariate effects and lose power in the presence of strong environmental effects.

162

A power comparison of multilocus mapping approaches in the presence of epistasis and heterogeneity

J. Tzenova, M. Farrall
Dept of Cardiovascular Medicine, Wellcome Trust Centre for Human Genetics, Univ of Oxford, UK

We investigate the performance of five linkage methods, previously proposed to assess the evidence that multiple regions contribute to a trait, under two-locus genetic models of epistasis and heterogeneity in samples of affected-sib-pairs (ASP).
Two-locus models were specified in terms of the two-locus penetrances, resulting in 9 models of epistasis, 2 of heterogeneity, and 2 additive models. Null models of neither gene contributing to the trait, and single-gene recessive, dominant, or interference models were also generated. Marker genotypes were simulated in 100 ASP assuming a population prevalence of 10%, sib recurrence risk ratio of 2, and equal disease allele frequencies at both disease loci. We analyzed each replicate with 1/ genehunter-two-locus using the two-locus parametric lod (Strauch et al., Am J Hum Genet, 2000, 66:1945–1957), 2/ genehunter-plus using epistatic and heterogeneity discrete weighting schemes (Cox et al., Nat Genet, 1999 21:213–215), 3/ twoloc 〉 a simultaneous two-locus mapping approach (Farrall, Genet Epi, 1997, 14:103–117), 4/ genefinder - a generalized estimating equation method (Liang et al., Genet Epi, 2001, 21:105–122), and 5/ logistic regression (Holmans, Hum Hered, 2002, 53:92–102). We evaluate power for each method and examine the effect of missing data. Overall, models of strong epistasis fare well regardless of the mapping approach, while additive models have less power. We also investigate parameter estimates from three methods in an attempt to obtain a range of values specific to the genetic model simulated.

163

Bivariate linkage analysis of asthma-related phenotypes in 295 French EGEA families indicates a pleiotropic quantitative trait locus (QTL) in the 21q22 region

A. Ulgen(1), E. Bouzigon(1), M.H. Dizier(2), J. Maccario(3), C. Krähenbühl(1), A. Lemainque(4), M.P. Oryszczyn(3), F. Kauffmann(3), M. Lathrop(4), F. Demenais(1)
(1) INSERM EMI0006, Evry, (2) INSERM U535, Villejuif, (3) INSERM U472, Villejuif, (4) Centre National de Génotypage, Evry, France

A genome screen for asthma and seven asthma-related phenotypes was recently conducted in 295 EGEA families ascertained through asthmatic probands. This scan showed potential evidence for linkage of 21q22 region to two asthma-related phenotypes: %predicted FEV1 (forced expiratory volume in 1 second) and SPTQ (number of positive skin test responses to 11 allergens). To investigate whether the 21q22 region may contain a QTL with a pleiotropic effect on %FEV1 and SPTQ, we conducted a bivariate linkage analysis based on Haseman-Elston regression method, using a fine map of 13 microsatellites (average spacing of 3 cM). The multivariate statistic was calculated as the sum of the squared univariate t-test statistics obtained from linkage analysis of the individual principal components of the observed phenotypes. This bivariate analysis indicated linkage in the same region as the univariate analyses of the phenotypes with a p-value of 0.0001 (at 35 cM from pter), this p-value being lower than the p-values previously obtained when analyzing %FEV1 and SPTQ separately (p=0.002 for%FEV1 and p=0.004 for SPTQ). This result suggests that the 21q22 region may contain a QTL underlying these two phenotypes. Further analysis based on bivariate variance component method will be performed to further investigate this finding.

164

Genomic Screening in Family Based Association Testing and the Multiple Testing Problem

K. Van Steen(1), M.B. McQueen(1), A. Herbert(2), C. Rosenow(3), E.K. Silverman(4), N.M. Laird(1), S.T. Weiss(4) C. Lange(1,4)
(1) Harvard School of Public Heatlh, USA, (2) Boston University, USA, (3) Genomics Collaboration Genotyping, Affymetrix, Inc., USA, (4) Harvard Medical School, USA

In this contribution we address the multiple testing problem in the context of genome-wide family-based association studies, using ideas of Lange et al. (2003). Our methods use the entire sample and do not require separate screening and validation samples to establish genome-wide significance, as for population-based designs. The methodology is implemented in the PBAT software (Van Steen and Lange, 2005), which can be freely downloaded from http://www.biostat.harvard.edu/~fbat/default.html.
Via simulations we show that the proposed 2-step screening technique maintains its protective character for extended data sets with a few hundred thousand SNPs. Compared to FDR-methods, our proposed screening-tools are particularly attractive to detect >1 trait-influencing loci in data sets with thousands of SNPs, even for low heritabilities. The Framingham Heart Study Data using a 100 K Affymetrix SNP mapping array is used to illustrate our strategies on real data.
The paper describing the results of this research has been accepted for publication by Nature Genetics.
Van Steen, K. and Lange, C. PBAT: a comprehensive software package for genome-wide association analysis of complex family−based studies. Hum Genomics 2(1):67–69 (2005).
Lange C et al. Using the noninformative families in family-based association tests: a powerful new testing strategy. Am J Hum Genet 73:801–11 (2003).

165

Consent and Population-based Family Cancer Genetic Research: Ethical and Methodological Issues

C.M. van Vliet, M.A. Jenkins, D.M. Gertig, J.L. Hopper
Centre for Genetic Epidemiology, University of Melbourne, Australia

By studying family history and collecting DNA samples and lifestyle information, population-based family cancer genetic research identifies and estimates the frequency of gene variants associated with increased cancer risk; estimates cancer risk for a given family history; and identifies environmental modifiers of risk. This information is important for genetic counselling and genetic

testing and contributes to improved diagnostic and preventive strategies.
Despite the potential for great public good, family studies present some unique ethical and methodological challenges regarding consent. As the unit of research is the family and not the individual, the information collected is jointly owned and relevant to each family member's participation. However since the focus of legislation and ethical guidelines is on individual consent, it is not clear to what extent family history collected on members who decline to participate or from whom consent is not or cannot be sought may be used. The area is contentious as the information collected is sensitive, there is public concern over potential harm, the omission of family data may jeopardize the study's validity and the ability to ''de-identify'' information is limited as individuals remain linked to other identifiable family members. In addition, cancer research requires identifiable information for verification of cancer reports.
The implications of these issues will be discussed in relation to consenting and non-consenting family members, researchers and the broader community.

166
Genetic and Environmental Variation of Prenatal Growth and Birth Weight of Twins
R. Vlietinck(1,2), C. Derom(1), M. Gielen(2), M.C. Neal(3), H. Maes(3), P. Lindsey(2)
(1) Human Genetics, University of Leuven, Belgium, (2) Population Genetics, University of Maastricht, Netherlands, (3) Human Genetics, Medical College, Richmond, Virginia, USA

Background: Birth weight is the major determinant of neonatal mortality. Low birth weight increases the risk of the metabolic syndrome.
Aim: Determine the genetic and environmental variation of the weight at birth and their influence on the intrauterine growth curves.
Subjects: All live born twins born between July 18, 1964 and December 31, 1997 in East Flanders (Belgium) were prospectively ascertained. Birth weight, gestational age, parity, congenital malformations had to be known. Fetal membranes, umbilical cord were examined within 24 hours and the zygosity was determined by blood groups and DNA fingerprints. Of these 10.080 children 9.058 (89.9%) remained after the following exclusion criteria: stillborn, unrealistic birth weight, major congenital malformations and zygosity probability of <0.95.
Methods: The birth weight was controlled for gestational age, maternal age and parity, number of placentas, umbilical cord implantation, vascular anastomoses by Mixed GLM models. The influence of genetic and environmental variation was assessed by path analysis comparing the birth weight of 7groups: dizygotic dichorionic male-male, female, and male pairs with the monozygotic dichorionic male-male and female and the monozygotic monochorionic pairs. The covariables were the sex, zygosity and the chorionicity.
Results: Genes determined 20.8% of the variation, while 64.8% was determined by environmental factors common to both twins: 34.1% by their gestational age, 15.5% by their chorion type and 5.2% by the maternal age, which leaved 14.4% by other common environmental causes.While males weighted 90 gr more than females, DZ twins weighted 80 gr more than MZ and those with a central cord insertion were 160 gr heavier than with a peripheral one.

167
A Spatial Clustering Approach to Fine-Mapping of Disease Genes
E.R.B. Waldron(1) J.C. Whittaker(2) D.J. Balding(1)
(1) Dept. of Epi & Public Health, Imperial College London, (2) Dept. of Epi & Pop Health, London School of Hygiene & Tropical Medicine

Attempts to improve the power of genetic association studies beyond what can be achieved from analysing markers one by one are often based on using a 'haplotype' where a group of markers are treated as the unit of analysis. However, these approaches also suffer limitations in that the similarities between haplotypes may not be accounted for in an effective way that models their recent shared ancestry and it can also be difficult to allow for rare haplotypes. I will describe an approach based on defining and searching through a haplotype space for a case-rich cluster of similar haplotypes. A simulation study has shown this approach to be more effective than alternative approaches and also allows us to identify situations in which multi-point approaches can substantially improve on a single-point approach.

168
A multiallelic test for marker-trait association studies
K. Wang

A constrained likelihood ratio test of association between a trait and a marker that has multiple alleles is introduced. The constraint requires that the genotypic effect of any heterogeneous genotype is neither higher than the larger nor lower than the smaller of those of the two corresponding homogeneous genotypes. This test is proposed in the framework of generalized linear models and is applicable to various types of traits, including quantitative traits and dichotomous traits. This method tests for the association between the trait and the genotypes and it does not rely on the validity of the Hardy-Weinberg equilibrium. Simulation studies are conducted to assess its performance in comparison with some other competing methods.

169
A quantitative linkage score (QLS) for improving an association study following a linkage analysis
T. Wang and R.C. Elston
Department of Epidemiology and Biostatistics, Case Western Reserve University

Currently, a commonly used strategy for mapping complex traits is to use a linkage analysis to narrow suspected genes, followed by an association analysis to fine map the genetic variation in regions showing linkage. Two questions arise in the design and the resulting inference at the association stage of this sequential procedure: (1) how should we design an efficient association study given the information provided by the previous linkage study? and (2) can an association in a linkage region explain, in part, the detected linkage signal? We derive a quantitative linkage score (QLS) based on Haseman-Elston regression and make use of this score to address both questions. In designing an association study, the selection of a subsample from the linkage study sample can be guided by the linkage information summarized in the QLS. When heterogeneity exists, we show that that selection based on the QLS can increase the proportion of sample individuals from the subpopulation affected by a disease allele and therefore improves power of the association study. For the resulting inference, we frame as a hypothesis test the question of whether a linkage signal can be partly explained by a marker allele. A one sided paired t-statistic is defined by comparing the two sets of QLSs obtained with/without modeling a marker association: a significant difference indicates that the marker can at least partly account for the detected linkage. All results suggest that a careful examination of QLSs should be helpful for understanding the results of both association and linkage studies.

170

Association of CTLA4 gene polymorphisms with age at diagnosis of type 1 diabetes in Newfoundland: a family-based study

K.S. Wang(1), M. Liu(1), B. Bharaj(1), M. Lu(1), H.T. Chen(1), J.A. Curtis(2), L.A. Newhook(2), A.D. Paterson(1,3)
(1) Program in Genetics & Genomic Biol, Hospital for Sick Children, Toronto, Canada, (2) Dept. of Pediatrics, Memorial Univ. of Newfoundland, (3) Dept. of Public Health Sciences, Univ. of Toronto

The CTLA4 gene has been confirmed to be associated with type 1 diabetes (T1D) in several studies and may also be associated with age at diagnosis (AAD). A major challenge to mapping AAD as a quantitative trait is its special distribution. We investigated the association of eight SNPs and (AT) n around the CTLA4 gene with AAD using 472 trios from a relatively isolated population of Newfoundland (NF) which has a high incidence of T1D. Firstly, the observed distribution of AAD is a mixture of two normal distributions with mean AAD of 7.97 (range 0-16.1) and 24.65 (range 17-40) years. According to the two distributions of AAD, the 472 trios were divided into early and late subgroups of T1D (394 and 78 trios, respectively). Furthermore, PBAT demonstrated five SNPs (CTAF343, rs1863800, MH30, JO31 and JO27) are significantly associated with AAD in the early subgroup ($P < 0.0056$) but not in the late subgroup. Moreover, haplotype-specific PBAT indicated haplotype CTAF343-JO27 is most significantly associated with AAD in the early subgroup ($P=0.00024$). In conclusion, the CTLA4 gene has an important effect on the AAD of T1D in the early subgroup and thus may be related to the pathogenesis of T1D in young patients in NF. However, the association analysis in the late subgroup may have limited power because of the small sample size (78 trios).

171

PEDMerge – A program for merging extended pedigrees: The CANHR Study

Y. Wang, B.B. Boyer, A. Goropashnaya, G.V. Mohatt, R. Plaetke
Center for Alaska Native Health Research, University of Alaska Fairbanks, Fairbanks, AK 99775

We ascertain extended pedigrees for a genetic study of obesity and type 2 diabetes mellitus. In our study, pedigrees often become large. Therefore we collect the information in ''sub-pedigrees'' using a pedigree drawing program in the field. Finally these sub-pedigrees need to be merged for data analyses. Merging pedigrees can be tedious and error prone even when using pedigree drawing programs. For example, in our study we have been repeatedly merging 26 ''sub-pedigrees'' to obtain one pedigree consisting of 986 individuals (recruited: 507) for genetic statistical analyses. In order to solve this problem, a Perl program PEDMerge 〉 running across all platforms - was developed that merges multiple sub-pedigrees to one pedigree based on linking individuals that are occurring in (multiple) sub-pedigrees. The size and number of sub-pedigrees to be merged are unlimited. Thus, each sub-pedigree may have multiple key individuals.
PEDMerge takes two input ASCII files: (1) pedin.txt, a standard pedigree file consisting of pedigree ID, ego ID, father ID, mother ID, sex and (2) keys.txt, a file that provides information about the key individuals– specific IDs in each sub-pedigree. PEDMerge generates two output files: (1) pedout.txt, a file containing all merged pedigrees and unmerged pedigrees and (2) master.txt, a candidate Master file for the database PEDSYS developed by the Southwest Foundation for Biomedical Research (San Antonio, TX).
We will show (1) our approach to collect pedigree data in the field and (2) the application of PEDMerge.

172

Problems in attempts to account for a linkage signal using candidate polymorphisms

D.M. Warren, H.H.H. Göring, J. Blangero, L. Almasy
Dept. of Genetics, Southwest Foundation, San Antonio, TX, USA

After identification of a locus by linkage analysis, linkage analyses can be performed conditional on a measured genotype to examine whether a variant is the only functional polymorphism in the candidate region. If it is, incorporating the measured genotype into the mean effects model should absorb most or all of the QTL variance in the region and thus greatly reduce or eliminate the LOD score. It is often assumed that a conditional LOD

>0 indicates additional functional polymorphisms in the region or measured genotype nonfunctionality. To test this, we simulated in SOLAR a diallelic functional site and a fully linked marker. Frequency of one functional site allele was set at 0.10 or 0.01. Locus-specific heritability (h2q) was varied from 0.05 to 0.40 in 0.05 unit increments. We simulated 5000 replicates for each setting. For h2q generating values <0.10, only 3% of replicates had a LOD >3. In those cases, h2q was always overestimated (as previously described by Göring et al., AJHG 2001;69:1357–69) and 45% of conditional LODs were >0.50 (a conservative estimate of nonsignificance). The tendency to overestimate h2q and to fail to absorb the QTL variance in conditional linkage analyses decreased for increasing h2q generating values. Our results suggest that a conditional LOD >0 may not indicate additional functional polymorphisms in a region. This is of particular concern when power to detect linkage is low (e.g. when h2q is low), and linkage is likely to be identified through a ''lucky bounce'' that is unlikely to be eliminated when analyses are conditioned on measured genotype.

173

Application of Self-Organizing Maps to detect Population Structure

N. Wawro(1), I. Pigeot(1,2)
(1) Dept. of Mathematics & Computer Sciences, Univ. of Bremen, Germany, (2) Bremen Institute for Prevention Research & Social Medicine, Bremen, Germany

In the context of association studies undetected population structure may lead to false positive results. Well known methods to cope with this problem are Genomic Controls and the Structured Association approach. We propose the use of exploratory Self-Organizing Maps (SOMs) to detect population structure and to derive estimates of the individuals' genetic backgrounds. These estimates may be used for a stratified test for association.
We present the results of a simulation study where a discrete and an admixed structure have been investigated. The focus was the correct identification of the number of the involved subpopulations and the evaluation of the background estimates. A varying number of highly informative loci was used as well as different map sizes and sample sizes. We investigate metric repeat scores. For this purpose, the SOM algorithm is extended appropriately to handle such paired genotypic data. In the discrete population model SOMs worked nearly perfect in most of the settings, whereas in the admixed settings the correct identification of the number of subpopulations seemed far more challenging. Here the estimates of the genetic backgrounds were strongly biased.

174

Human Genetic Association Analysis Based on the Allelic Composition of Multiple Loci

J. Wessel(1,2,3) and N.J. Schork (1,2)
(1) Polymorphism Research Laboratory, Department of Psychiatry, (2) Department of Family and Preventive Medicine, University of California, San Diego, and (3) School of Public Health, San Diego State University,

The completion of the Human Genome Project and soon the HapMap, have facilitated the production of large amounts of genotype, haplotype and tagging SNP data. The success of association studies to identify polymorphisms that have an affect on complex traits has been limited. A number of design and analytical methods have been proposed to overcome some of the limitations of association studies, namely issues concerning linkage disequilibrium, haplotype analyses and study designs. Likewise using phylogenetic methods or ancestry information has been suggested as a method for grouping individuals, only it can ignore recombination events that may make individuals appear to be ancestrally related but not genetically or phenotypically related. We have developed methods of analysis for examining the multi-allelic composition of genotype and haplotype information. Furthermore we use functional and sequence conservation information to place individuals in groups whose phenotypes can be contrasted. We illustrate our methods of calculating genetic similarity with 16 common SNPs in the CHGA gene and testing for maximal grouping with Chromogranin A (CgA) levels in 370 unrelated, white, non-Hispanics from Southern California. Ultimately, pairs of individuals can have similar values of genetic similarity, but differ in their actual alleles. The addition of functional information served to distinguish individuals with similar phenotypic effects. Cluster analysis grouped individuals into more manageable groups to identify the optimal grouping for association testing with CgA levels. We found our methods of analysis to be better at capturing genetic variation, and as such can improve the ability to identify genetic associations.

175

Estimation and testing of genotype and haplotype effects in family-based analyses of quantitative traits: comparison of prospective and retrospective approaches

E. Wheeler, H.J. Cordell
Department of Medical Genetics, University of Cambridge, UK

The case/pseudocontrol approach is a convenient framework for family-based association analysis of case-parent trios, incorporating several previously-proposed methods such as the transmission/disequilibrium test and log-linear modelling of parent-of-origin effects. The method allows genotype and haplotype analysis at an arbitrary number of linked and unlinked multiallelic loci, as well as modelling of more complex effects such as epistasis, parent-of-origin effects, maternal genotype and mother-child interaction effects, and gene-environment interactions. Here we extend the method to perform analysis of quantitative as opposed to dichotomous (disease) traits. The resulting method can be thought of as a retrospective approach, modelling genotype given trait value. Application of this method to quantitative traits involves several complications not encountered when analysing disease

traits, such as issues related to selected sampling and unmodelled population stratification. Through simulations and analytical derivations, we examine the power and properties of our proposed approach, and compare it to several related methods for single-locus quantitative trait association analysis. All methods are found to give correct type 1 error rates and unbiased parameter estimates when applied to randomly ascertained trios from a single homogeneous population. With randomly ascertained families, in the presence of population stratification, a prospective approach (modelling trait value given genotype) is generally more efficient and interpretable than our retrospective approach. However, our method is found to have some advantages with regard to estimation and interpretability of parameter estimates when applied to selected samples.

176

Association of polymorphisms in SOD2 with occurrence of Alzheimer's Disease

H.W. Wiener, R.T. Perry, R.C.P. Go
Univ. of Alabama at Birmingham, Birmingham, AL

Oxidative damage due to free radicals and reactive oxygen species (ROS) appear to contribute to the pathogenesis of Alzheimer–s Disease (AD). Evidence indicates oxidative damage may be increased in the AD brain. SOD2, an antioxidant enzyme in the mitochondria, scavenges free radicals. Both elevated and decreased SOD2 activities have been reported in the AD brain. We previously genotyped four SNP loci located in the 5– UTR (rs2758346), exon 2 (rs1799725), intron 3 (rs2855116), and 3– UTR (rs5746136) of the SOD2 gene. These were genotyped in three study groups: a subset of the families ascertained by the NIMH AD Genetics Initiative; a group of African American (AA) AD patients and age matched controls; and a collection of Caucasian controls ascertained from a study of age related macular degeneration (AMD). We found significant evidence for association between several of these polymorphisms and occurrence of AD (NIMH group association p-values=.03 for rs2758346, .03 for rs1799725, .03 for rs2855116, .03 for rs5746136; AA group p-values=.02 for rs2758346, $<.01$ for rs2855116), with the intronic polymorphism demonstrating the most significant and consistent association. Motivated by these results, we are pursuing an affective polymorphism in the coding and other significant sequences surrounding rs2855116, and will present these results at the meeting.

177

No evidence of major population substructure in the Framingham Heart Study

JB Wilk, AK Manning, J Dupuis, LA Cupples, MG Larson, C Newton-Cheh, S Demissie, AL DeStefano, SJ Hwang, C Liu, Q Yang, KL Lunetta
Boston University Schools of Medicine and Public Health, The National Heart, Lung and Blood Institute's Framingham Heart Study, and Massachusetts General Hospital

Population stratification continues to be a concern for genetic association studies, but often limited information is known about the structure of the sample. The Framingham Heart Study original and offspring cohorts consist of white Americans of European descent. The most frequent self-reported ancestry of participants was western European, which included individuals of mixed ancestry, and ancestry specifically from Great Britain, Italy, and Ireland was also commonly reported. We sought to use genome-wide microsatellite markers to detect population substructure, hypothesizing that we might detect evidence of subpopulations reflecting ancestral European geography. We studied 346 randomly selected unrelated participants using the Structure v.2.1 software. An admixture model, allowing individuals to have ancestry from multiple subpopulations, was implemented with 74 unlinked genome-wide markers under the assumption of correlated allele frequencies in subpopulations. In a second analysis, we used a set of 601 markers on all autosomes and an admixture model that allowed linked markers. The MCMC algorithm was run to independently test the fit of the assumption of (K) between 1 and 5 subpopulations, with 50,000 burn-in and 100,000 replicates. Both the unlinked and linked admixture models provided the same result: the data from these Framingham participants are most consistent with a single population. The estimate of F_{st}, the average genetic drift of the population away from the ancestral mean, was 0.02 for both models. These results suggest that the geographically and racially homogeneous Framingham sample does not have detectable genetic substructure using these genetic markers.

178

Substantial Advantage for Longevity in Siblings of Okinawan Centenarians

B.J. Willcox(1,5), W.-C. Hsueh(2), Q. He(1), D.C. Willcox(3), J.D. Curb(1,5), M. Suzuki(4)
(1) Pacific Health Research Inst., USA, (2) UC San Francisco, USA, (3) Okinawa Prefecture Univ., Japan, (4) Okinawa Gerontology Research Center, Japan, (5) Univ. of Hawaii, USA

Okinawa, an isolated island prefecture (state) of Japan, has among the world's longest life expectancy and what may be the world's highest prevalence of centenarians ($\sim$5 times the U.S. prevalence rate). Whether such a phenomenon is partially due to genetic factors is unclear. We hypothesized that exceptional longevity in Okinawa is, in part, familial. To obtain a quantitative estimate of familiality, we analyzed a population-based sample including 969 siblings (507 females and 462 males) of 348 centenarians born between 1874 and 1902. The median sibship size was 3, ranging from 1–10. The age of the proband was verified by a public family registry and that of the siblings was reported by the proband or family members. The prevalence of centenarians in their contemporary cohort was obtained from a public registry. The sibling relative risk estimates (λ_S) of centenarianism for Okinawan centenarians were 6.5 (95% CI: 3.9-10.7) for females and 5.1 (95% CI: 1.8-14.2) for males. The weighted

sex combined λ_S was 6.3. These estimates in Okinawans appear to be higher than that obtained in an earlier study based on U.S. whites and suggest a substantial familial component to Okinawan longevity. Further research is needed to determine the basis for the potential genetic contributions to longevity in Okinawa.

179
Seeing the Forest for the Trees in Genome-wide Association Studies
J. Witte, E. Jorgenson

Greater knowledge about human genetic variation and advances in genotyping technology have led to an escalating interest in genome-wide association studies. Undertaking these studies requires large SNP sets with comprehensive coverage of either the entire genome or at least all known genes (i.e., positional cloning or candidate gene based sets). We first consider the rationale for developing such SNP sets, the structure of linkage disequilibrium in gene-rich and non-gene-rich regions of the genome, and the resulting implications for genome-wide association studies. We then compare the sample sizes needed for—and relative cost of—positional cloning versus candidate gene studies. The positional cloning approach requires substantially more SNPs, slightly larger sample sizes, and thus is considerably more expensive to perform than the candidate gene approach. However, the candidate gene approach may miss a limited number of causal variants outside of genes, specifically variants in cis-regulatory regions which may be located far away from coding regions. The expected ratio of causal variants in genes versus outside of genes helps distinguish which approach is most appropriate.

180
Do disorders of speech and language share common genetic components?
J.K. Wittke-Thompson(1) and N.J. Cox(1,2)
(1) Dept. of Human Genetics, The University of Chicago, USA, (2) Dept. of Medicine, The University of Chicago, USA

Speech and language disfluency is a phenotype that is commonly found in numerous complex disorders that include, but are not limited to, stuttering, Specific Language Impairment (SLI), autism, dyslexia, and Tourette–s syndrome (TS). Numerous genome-wide linkage analyses have been conducted, but as is common with many complex disorders, very few regions have been replicated in independent samples. We have conducted a meta-analysis of numerous studies in which some component of speech and language disfluency was assessed for linkage to identify a shared component of speech and language with disorders that appear not to be genetically connected. One region identified is on 13q21, in which there is evidence of linkage in several studies of SLI, TS, autism, and stuttering. Stuttering and the delay of phrase-based speech in individuals with autism share evidence for linkage on chromosome 2 (180-210 cM). There are also several regions with evidence exclusive to stuttering on chromosomes 3, 5, 7, 9, 13, and 15. By identifying the common regions in the genome that contain genetic variation that increases susceptibility to speech and language disfluency, it may be easier to detect the unique genetic, and environmental, components that are exclusive to a single speech and/or language disfluency disorder.

181
Computational models for the emergence of genomic instability: implications for aging and the incidence of cancer
D. Wodarz and N. Komarova
Department of Ecology & Evolution, University of California Irvine, USA

Genomic instability is defined as an elevated rate with which cells acquire genetic changes. Genetically unstable cells are found in a variety of conditions, including Werner's syndrome, ulcerative colitis, and cancer. Genetic instability occurs when checkpoint mechanisms, which are responsible for DNA repair and genome maintenance, become corrupted. We use computational models to investigate how the level of DNA damage can determine whether tissue cells remain stable, or whether unstable cells emerge. Cells can receive DNA damage through exposure to environmental carcinogens, or through oxidative radicals which accumulate with age. If DNA damage is received, stable cells repair the damage, while unstable cells do not. Both strategies are associated with advantages and disadvantages for the cells. If cells repair the damage, deleterious mutations are rare. On the other hand, the cells enter cell cycle arrest upon repair, and this can slow down the rate of cell division. Unstable cells, on the other hand, avoid cell cycle arrest and a slower rate of cell division upon damage because they do not repair. However, unstable cells can readily accumulate deleterious mutations. In the context of this tradeoff, we determine the effect of DNA damage on the growth kinetics and relative fitness of stable and unstable cells. This has implications for understanding how aging and the presence of environmental carcinogens correlate with the incidence of cancer in the human population, and for designing prevention strategies.

182
Effect of altering population parameters on linkage statistics in extended pedigrees using Merlin-Regress
R Wojciechowski(1), D Stambolian(2), JE Bailey-Wilson(1)
(1) IDRB/NHGRI, USA, (2) Ophthalmology, Univ of Pennsylvania, USA

Background: Regression-based linkage methods can offer good power to detect quantitative trait loci and are robust to trait distribution assumptions. However, the method implemented in Merlin-Regress requires correct specification of population parameters of the trait distribution. We estimated the effect of alternate specifications of the

underlying population mean (μ), variance (V) and heritability (h2) of ocular refraction on linkage statistics in Merlin-Regress.
Methods: The sample consisted of multigenerational Ashkenazi Jewish families, selected to contain multiple myopic individuals. Refractions were log-transformed to achieve zero skewness prior to analysis. Initial population parameters were chosen based on population-based studies of refraction and our sample trait distribution. The values of μ, V and h2 were initially set to 3.12,.02 and.6, respectively. μ, V, and h2 were then individually varied to: μ=3, 3.06, 3.19, 3.29; V=.0067,.01,.04,.06; and h2=.1,.3,.45,.75,.9 and 1.
Results: Our best-guess model yielded a maximum multipoint lod score of 8.7. This corresponds to a simulation-based genomewide significance of $p < 0.01$. Varying h2 from.3 to.75, μ from 3 to 3.19 and V from.0067 to.04 had negligible effects on the maximum lod score.
Summary: The Merlin-Regress QTL linkage method appears to be robust to misspecification of population parameters within a reasonable range of values. We are conducting simulations to study the effect of alternate parameter specifications on empirical p-values.

183
Extension of Haseman-Elston Regresion Model to Longitudinal Data
Sungho Won(1), Taesung Park(2)
(1) Dept. of Epi & Biostat, Case Western Reserve Univ., USA, (2) Dep. of Statistics, Seoul National University, Korea

Although Haseman and Elston (HE) method is one of the most commonly used model-free linkage methods, not many extensions to longitudinal data have been developed. We propose an extension of HE method for linkage analysis of the longitudinal data. The proposed model is a mixed model having several random effects. As a response variable, we use the squared difference, the mean corrected cross-product, and the weighted sum of the squared difference and the squared mean-corrected sum. The proposed model allows us to investigate the time effect by treating it as a pair-specific covariate. Thus, we can test whether or not the genetic variation changes by age. Also, the proposed model can test for the gene time interaction.
The proposed model was applied to the analysis of GAW13 simulated dataset for a quantitative trait of the systolic blood pressure. A simple independence model and four mixed models were compared. For each model, three basic models were defined: model without time effect, model with time effect, and model with the interaction between gene and time (gene time). Independence model did not preserve the sizes, while four mixed models preserved sizes. In general, the models with the gene time interaction tended to preserve sizes well. Mixed models with sibpair and common sib random effect tend to have higher powers. The proposed models seem not only quite useful in detecting linkage with the longitudinal data for the trait but also quite flexible to use. They could handle a wide class of correlation structures. Models with more general class of covariance structure are desirable.

184
Characteristics and Familial Patterns of Sleep in the Old Order Amish
S.-H. Wu(1), B.D. Mitchell(2), A. Mody(3), K. Hairston(2), P. Sack(2), A.R. Shuldiner(2), S. Snitker(2), W.-C. Hsueh(1)
(1) UC San Francisco, CA, (2) Univ. of Maryland, Baltimore, MD, (3) UC Berkeley, CA

Although under influence of some external factors (daylight, temperature, etc.), sleep/wake patterns are tightly regulated. Humans are unique in that they will curtail their sleep to accommodate their lifestyle, which may reduce the genetic influence on sleep regulation. The goal of this study was to investigate sleep patterns in the Old Order Amish, whose agrarian lifestyle is more homogeneous than modern societies.
The sample included 258 subjects (49% male) from 18 families, with a mean age of 45 ± 15 yrs (range: 20-80 yrs). The most common occupation for women was homemaking (82%) and for men farming (39%). We determined sleep and wake-up times using 7-day accelerometer records, and computed sleep duration as the difference between the sleep and wake-up times. Heritability was estimated by variance-component method using the SOLAR program.
The mean sleep time, wake-up time, and sleep duration were 9:50 PM $\pm$ 44 min, 5:12 AM $\pm$ 46 min, and 7.4 ± 0.9 hrs, respectively. The seasonal effect on the 3 sleep traits was significant, accounting for 12–15% of the variance. Sex and age were not significantly associated with these traits. After adjusting for the aforementioned factors, there was no significant heritability in the sleep time or the sleep duration, while the heritability for the wake-up time was 0.37 ± 0.22 ($p=0.04$).
In conclusion, we detected familial aggregation in wake-up times, but it remains to be determined whether this is due to shared environment or genes.

185
Joint Effects of Germline p53 Mutations and Sex on Cancer Risk in Li-Fraumeni Syndrome
C.C. Wu(1), S. Shete(1), C.I. Amos(1), L.C. Strong(2)
(1) Dept of Epidemiology, (2) Section of Clinical Cancer Genetics, Dept of Molecular Genetics, UT M. D. Anderson Cancer Center, Houston, Texas

To characterize the cancer incidence related to germline p53 mutations, we have analyzed longitudinal data from a study of familial cancer in 159 kindreds systematically - ascertained through childhood soft tissue sarcoma patients treated at M. D. Anderson Cancer Center between 1944 and 1975. We have followed these kindreds for >20 years and identified 7 families with germline p53 mutations. To evaluate quantitatively the effects of p53 mutations on cancer risk, and to identify any risk modifiers, we utilized segregation models accommodating p53 mutation and

other factors as covariates to analyze the data. p53 mutations had been identified a significant effect on cancer risk: carriers had 110-fold greater odds of developing cancer than noncarriers in males and 478-fold higher odds in females. Sex had a significant effect on cancer risk in people with p53 mutations: female carriers had 4.70-fold higher odds than male carriers on cancer risk. To determine the most likely mode of disease inheritance in the absence of germline p53 mutations, we also performed a segregation analysis of 100 kindreds from the series with no p53 mutations. Our results provide important insight into the genetic etiology of the presence or absence of germline p53 mutations and valuable reference for localizing or identifying non-p53 disease susceptibility loci or risk modifying loci and clinical counseling.

186

Distribution and Magnitude of Type I Error of the Model-based Multipoint Lod Score over Genetic Models

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

The magnitude of type I error and methods for correcting multiple testing when maximizing the single point lod score over both the recombination fraction and genetic model parameters have been widely investigated. The mod score method is recommended not only for detecting linkage, but for referring the mode of inheritance of the disease. However, in the situation of multipoint lod score, where there is no need of maximizing likelihood over the recombination fraction, the asymptotic theory of distributions and so the former analytical work fail. In this study we simulated a map of two markers 10 cM apart and one unlinked disease locus under both dominant and recessive model with reduced penetrance of 0.8, 0.5, and 0.2, respectively. We analyzed the data under both dominant and recessive models with penetrances of 0.1 $\sim$0.9 by the multipoint linkage analysis. The results show that the distribution and magnitude of type I error are correlated with the genetic models specified. Low penetrance models tend to generate more false positive results. Therefore, when employing the multipoint lod (mod) score approach, caution should be taken to specify (refer) the genetic models.

187

Segregation analysis reveals a Mendelian co-dominant model for age-of-onset of lung cancer in families with young probands

H. Xu, M.R. Spitz, S. Shete
Department of Epidemiology, UT MD Anderson Cancer Center

Lung cancer risk is largely attributed to environmental exposures, but genetic predisposition also plays an etiologic role. Several studies have investigated the importance of genetic predisposition in lung cancer aggregation. Our previous study modeling the age-of-onset distribution of lung cancer using a large data set did not provide evidence of Mendelian inheritance. In this study, we performed further segregation analysis of lung cancer age-of-onset distribution using families of young (<65) lung cancer probands to reduce the heterogeneity. The analysis was performed on 8,050 individuals from 904 lung cancer families, allowing for the effects of smoking and sex. Genotypes were assumed to affect the age-of-onset but not susceptibility. Results indicated that the best-fitting model was a Mendelian co-dominant model of a major autosome gene that produces earlier age-of-onset for carriers. The results were largely consistent with the findings of Sellers et al. (1990), with the exception that the allele frequency was found to be higher in our model. The results may act as a guide for further studies to localize genes underlying the age-of-onset of lung cancer.

188

Model Selection in Genetic Association Studies

Y.-C. Yen(1), P. Kraft(1,2)
(1) Dept. of Epi & (2) Biostat, Harvard School of Public Health, Boston, MA, USA

Genetic association studies have become popular tools to uncover the genetic etiology of complex human diseases. To test for association between a diallelic variant and dichotomous disease, three 1-degree-of-freedom (d.f.) tests are possible, corresponding to dominant, recessive, and additive genetic models. When the underlying inheritance mode is unknown, often all three models are tested, with ''significance'' declared if one test is significant at the nominal level, and parameter estimates from the ''most significant'' model reported. This procedure is known to have higher than nominal Type I error rates; the nominal confidence intervals (CIs) for parameter estimated will also be too narrow. An alternative procedure is fitting a single genotypic-risk model which estimates genotype-specific relative risks without specifying a genetic model; for a diallelic marker, the genotypic-risk analysis yields a 2-d.f. test. Using simulation studies, we assess the Type I error rates, power, and CIs for these procedures. Simulations confirm that ignoring multiple testing inflates Type I error rates (ranged from 6% to 15% at 5% nominal significance level), and ignoring model selection leads to underestimated CIs (64.5% coverage for 95% CI, while 95.5% coverage in genotypic-risk model, under our most extreme scenario). The genotypic-risk model has correct Type I error rates while losing very little power relative to the true genetic model. Thus, we recommend that the genotypic-risk analysis should be used when the underlying genetic model is unknown, as long as the number of subjects with the rarest genotype is not too small.

189

Null Distribution of Maximized LOD Score over Genetic Models

Y.J. Yoo(1), N.R. Mendell(2), B. Nemesure(1), S.J. Finch(2), Q.K. Ye(3)
(1) Dept. of Preventive Medicine, Stony Brook University, USA, (2) Dept. of Applied Math and Stat, Stony Brook

University, USA, (3) Department of Epidemiology & Population Health, Albert Einstein College of Medicine of Yeshiva University

The LOD score is a likelihood ratio statistic that is used to detect linkage of a disease gene to a marker in genetic linkage analysis. In the calculation of LOD score, a genetic model should be assumed even when that is unknown. To avoid biased result, maximized LOD scores over several parameters or entire parameter space have been suggested. In this work, the asymptotic null distributions of LOD scores maximized over 2 or 20 parameter values (LOD2, LOD20) and over the entire parameter space (MLOD) are investigated. The theoretical asymptotic null distributions are obtained using the asymptotic theory of likelihood ratio statistics with boundary restriction. Also, empirical null distributions are obtained using simulated data and compared favorably with the theoretical distribution. The critical values corresponding to Type I error of 0.0001 are suggested from the obtained null distributions. Those values are 3.3, 3.6 and 4.9 for LOD2, LOD20 and MLOD respectively.

190

The Effect of Marker Density on Haplotype Block Structure

Z. Yu, R. Guerra, H.H. Hobbs, J.C. Cohen

The haplotype block structure has been studied by many groups because of its potential in helping to understand both how diseases are associated with mutations and the human evolutionary history. Different marker selection strategies have been suggested to reduce the genotyping expenses. However, there has been no systematic study about the impact of different maker selection on the haplotype block structure. The SNP data from Dallas Heart Study (DHS) contains gene loci where the entire DNA sequence has been obtained, and all SNPs identified, which makes study of marker density feasible. In this article, we will show the effect of the marker density and the effect of different tagging strategies on the similarity between haplotype blocks estimated from selected SNPs and the haplotype blocks estimated using all the informative SNPs. Haplotype block structure will be lost or misrepresented even using tagging SNPs that can keep a large portion of 'information', which has been defined by previous study.

191

The Effect of Sample Size on Tagging SNP Performance and Consequences for Complex Disease Gene Mapping

E. Zeggini, W. Rayner, L. Cardon, M.I. McCarthy, on behalf of the International Type 2 Diabetes 1q Consortium
WTCHG, University of Oxford, UK

As part of the International T2D 1q Consortium, we are following a linkage disequilibrium (LD) mapping strategy to identify type 2 diabetes susceptibility variant(s) on chromosome 1q. We have genotyped 3000 SNPs in 13.5 Mb of 1q in over 4000 individuals across 7 populations. The majority of markers have been selected from the HapMap to be common variants. A smaller proportion of SNPs have frequencies below 5%. This dataset was used to address the as yet unanswered question of the effect of sample size on tagging SNP (tSNP) performance. We have focused on 3 subregions containing common and rare SNPs, in 411 T2D cases and 408 matched controls. To emulate tagging strategies employing the HapMap, we restricted tSNP selection among common variants, using samples of different sizes (n1=45, n2=60, n3=100, n4=200 individuals) drawn randomly from the control group. Tagging SNP selection was carried out using the Carlson et al. (Am J Hum Genet, 74:106, 2004) method and the aggressive method implemented in Tagger (de Bakker, BROAD Institute). The proportion of all variation captured by different sets of tSNPs was evaluated. The tSNP sets captured similarly high proportions (>85%) of common variation, but only up to 20% of rare variation. Tagging SNP performance improved when increasing the selection sample size from 45 to 60 individuals, but did not change substantially with larger sample sizes. We are currently extending the analyses to simulated data, including a higher proportion of rare variants in the tSNP selection pool.

192

Age Dependent QTL Analysis Using Gibbs Sampling for Random Effects Models

F. Zhang(1,2), Q. Yang(1), L.A. Cupples(1)
(1) Dept. of Biostat, Boston Univ., USA, (2) Dept. of Ambulatory Care & Prevention, Harvard Univ. and Harvard Pilgrim Health Care, USA

Repeated measurements of traits contain more information than a single cross-sectional measurement for inference on age-dependent genetic effects. However, most existing linkage analyses either use derived summary measures or are only applicable to very limited patterns of age-dependent genetic effects. We provide a general extended variance components approach using Markov Chain Monte Carlo (MCMC) methods to fully utilize data collected longitudinally in extended human families to calculate age-dependent heritabilities for quantitative trait loci (QTL) as well as other parameters of interest. The model assumes an ignorable missing mechanism and allows for a polynomial representation of the QTL genetic effect, but limited the residual genetic effect to a scalar. We used the Deviance Information Criterion for model selection.
We found that non-informative priors with large sample sizes yielded the best resulting distributions of the parameters. We applied our method to high-density lipoprotein (HDL) data collected in the Framingham Heart Study. A quadratic age dependent model provided the best fit among the models we considered. In these data there is evidence that the quantitative trait locus effect on chromosome 6 is increasing with age among people aged 40 to 70.

193

Analytical Correction for Multiple Testing in Admixture Mapping, Including genome-scan

S. Zhang, Q. Sha, X. Zhu

Admixture mapping is an efficient approach to localizing disease-causing variants that differ in frequency between two historically separated populations. It may be more powerful than linkage studies and, for a genome search, it typically require only $\sim 1\%$ genotyping effort as many as required by a genome-wide association study. Recently, several methods have been proposed to test linkage between a susceptibility gene and disease locus by using admixture-generated linkage disequilibrium (LD) for each of the typed markers. In this report, we propose an analytical approach for correction of the multiple testing in admixture mapping and for calculating overall p-value for a genome scan. The proposed method is based on the Markov property of the marker-specific ancestry. We use simulation studies to evaluate the performance of the method. In simulation studies, we generate admixture population under different population models by using a dense ancestry-informative marker penal for African-American. The simulation results show that the proposed method gives correct overall type I error rate for genome scan under all the cases. The proposed method provides a useful alternative to more computationally intensive simulation-based tests and conservative Bonferroni correction.

194

Multipoint Linkage Disequilibrium Mapping with Haplotype Block Structure

M. Zheng(1) and M.S. McPeek(1,2)
Departments of Statistics (1) and Human Genetics (2), University of Chicago, Chicago, IL

The HapMap Project is providing a great deal of new information on high-resolution haplotype structure in various human populations. Statistical methods for linkage disequilibrium (LD) mapping are needed that make use of this information. We consider LD mapping in case-parent trios under the assumption that haplotype blocks and their common haplotypes have been identified. We propose methods for fine-mapping a disease-associated variant based on multipoint haplotype or genotype data, in which haplotype block structure is explicitly incorporated into the model. The first association method we propose, the multipoint likelihood method for observed variants (MOV), considers only typed markers. Recognizing that a limited number of ''tag'' SNPs may be typed in each block, we propose a second association method, the multipoint likelihood method for virtual variants (MVV), that considers not only typed markers, but also explicitly tries to detect whether a haplotype block is likely to have untyped variants that are strongly associated with the trait. Both the MOV and MVV methods use a multilocus likelihood for the entire set of data, giving two main advantages. First, hypothesis testing results at different points are directly comparable. As a result, the methods can be much more effective for localization than methods that do not use a multilocus likelihood for the entire set of data. Second, in the likelihood framework, efficient use is made of the data when some genotypes are missing. A goodness-of-fit test is applied to validate the haplotype block structure model used for inference. We also propose single-block methods (single-block observed variants method (SBOV) and single-block virtual variants method (SBVV)) which use only the multipoint information in each haplotype block and detect whether there are typed or untyped variants within this block that are strongly associated with the trait. These single-block methods are computationally much simpler and use some of the multipoint information, but test results at different blocks are not formally comparable when the amount of missing information varies across blocks.We compare, by simulation, the power and accuracy of localization of single-point, SBOV, SBVV, MOV, and MVV association analysis based on trio genotype data, taking into account haplotype uncertainty. These comparisons give some general insight into the value of using different amounts of multipoint information for association mapping, depending on the informativeness of the data and depending on whether the primary goal is detection or localization. The methods are applied to the data set of Daly et al. (2001).

195

A Classical Likelihood Based Approach for Admixture Mapping using EM algorithm

X. Zhu(1), S. Zhang(2,3), H. Tang(4), R.S. Cooper(1)
(1) Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, Maywood, IL60153, (2) Department of Mathematical Science, Michigan Technological University, Houghton, MI 49931, (3) Department of Mathematics, Heilongjiang University, Harbin, China, (4) Divison of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Several disease-mapping methods have been proposed recently, which use the information generated by recent admixture of populations from historically distinct geographic origins. These methods include both classic likelihood and Bayesian approaches. In this study we directly maximize the likelihood function from the hidden Markov Model for admixture mapping using the EM algorithm, allowing for uncertainty in model parameters, such as the allele frequencies in the parental populations. We determined the robustness of the proposed method by examining the ancestral allele frequency estimate and individual marker-location specific ancestry when the data were generated by different population admixture models and no learning sample was used. The proposed method outperforms a widely used Bayesian MCMC strategy for data generated from a various population admixture models. The multipoint information content for ancestry was derived based on the map provided by Smith et al. (2004) and the associated statistical power was calculated. The effects of misspecification of ancestral allele frequencies and linkage disequilibrium between adjacent markers were also considered. Finally, we examined the

distribution of admixture LD across the genome for both real and simulated data and established a threshold for genome wide significance applicable to admixture mapping studies. The software ADMIXPROGRAM for performing admixture mapping is available from authors.

196

First evidence that one or more rare genetic polymorphisms with high penetrance may be involved in the aetiology of endometriosis

K.T. Zondervan(1), J. Lin(2), G. Dawson(3), D. Zabaneh(3), V. Smith(3), S. Bennett(3), A. Lambert(2), A. Carey(3), D.E. Weeks(4), S.A. Treloar(5), G.W. Montgomery(5), D.R. Nyholt(5), N.G. Martin(5), L.R. Cardon(1), I. MacKay(3), J. Mangion(3), S.H. Kennedy(2)
(1) Wellcome Trust Cntr for Hum. Genet & (2) Nuffield Dept. of Obstet. and Gynaecol., Univ. of Oxford, UK, (3) Oxagen Ltd, UK, (4) Div. of Stat. Genet., Univ. of Pittsburgh, USA, (5) Queensland Inst. of Med. Research, Brisbane, Australia

Endometriosis (endometrial-like deposits outside the uterus) is a common complex disease associated with pelvic pain and subfertility in women. Its etiology is unknown, but familial aggregation has been demonstrated in humans as well as non-human primates. For a genome-wide linkage study in Oxford (UK), 256 Caucasian extended families with 2+ sister-pairs with endometriosis were recruited including 52 families containing 3 or more affecteds. Non-parametric genome-wide linkage analyses in these families, using sex-specific information from the Rutgers linkage-physical map, resulted in six LOD score peaks with K&C LODs >1. One of the peaks, on chromosome 7, reached a LOD of 3.49 (genome-wide p-value: 0.007).

The dataset was subsequently increased through collaboration with an Australian genome-wide linkage study consisting of 931 extended Caucasian families with 2+ affected sib-pairs, providing in unique combined dataset of 1187 families (the largest sample size available for studying genetic linkage of endometriosis in the world). The Australian dataset included 196 families with 3+ affecteds.

Using Merlin, parametric linkage models were run using the combined dataset of 248 families with 3+ affecteds. Maximum LOD scores (MOD scores) were found by iterating across different inheritance models, with a set phenocopy rate of 0.04. A MOD score of 4.10 was found at 59.3 cM for a recessive model, with allele frequency=0.01, heterozygous penetrance 0.1 and homozygote penetrance 1.0. The UK and Australian samples both supported the peak with MODs >2.0 at 59.3 cM, with all UK families and 87% of Australian families contributing to the linkage.

The results strongly suggests, for the first time, that one or more rare genetic polymorphisms on chromosome 7 with high penetrance are involved in the aetiology in a subgroup of women with endometriosis.