Genetic Epidemiology 25: 236–279 (2003)

ABSTRACTS FROM THE

TWELFTH ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

Redondo Beach, California, USA
November 3–4, 2003

1

Association Between Polymorphisms Upstream of Pituitary Growth Hormone and Term Birth Weight

R. M. Adkins(1), C. Campese(2), R. Vaidya(2), J. Krushkal(3), T. K. Boyd(4)
(1) Dept. of Pediatrics, Univ. Tenn. – Memphis, USA (2) Dept. Biology, Univ. Mass. USA (3) Dept. Preventive Medicine, Univ. Tenn. – Memphis, USA (4) Dept. Pathology, Baystate Med. Center, USA

Pituitary growth hormone is one of 5 genes in the growth hormone (GH) locus, and is unusually polymorphic with 27 SNPs in the span of 1,750 nucleotides. We first provide compelling evidence that the high polymorphism and complex haplotype structure of pituitary GH is the consequence of recurrent gene conversion. Next, we demonstrate the superiority of Bayesian inference (program PHASE) of haplotypes over that of various implementations of the EM algorithm in the presence of gene conversion. The accuracy of haplotype inference is validated using empirically determined haplotypes from a large independent study. Because anencephalic fetuses can achieve nearly normal body length, the accepted wisdom is that pituitary GH does not regulate fetal growth. We demonstrate suggestive association between upstream polymorphisms of GH in key functional sites with variation in term birth weight. These results suggest that pituitary GH may play a role in the regulation of fetal growth late in gestation when GH is expressed by the fetus and GH receptors become widely expressed in many fetal tissues.

2

Strong association of intragenic SNPs of the 6q25 region with leprosy in two independent samples

A Alcais(1), M Mira(2), N Thuc(3), M Moraes(5), C Di Flumeri(1), V Thai(3), A Verner(4), A Montpetit(4), T Hudson(4), E Schurr(2), L Abel(1)
1) INSERM U.550, France, 2) Dept Biochemistry, McGill Univ, Canada; 3) Dermato-Veneorology hospital, Vietnam; 4) Genome Quebec Innovation Centre, McGill Univ, Canada; 5) Dept Tropical Medicine, IOC-Fiocruz, Brazil.

Leprosy is a chronic infectious disease caused by Mycobacterium leprae and is still a major global health problem with 700,000 new cases occurring each year. In a recent genome scan, we mapped a susceptibility gene for leprosy per se (all clinical forms) to chromosome region 6q25 (Mira et al, Nat Genet, 2003). Within the linked 6q25 region, one gene presented appealing characteristics in the context of leprosy control. We constructed a dense map of this gene and identified 74 informative SNPs. These SNPs were used to perform a family-based association study in a sample of 208 Vietnamese trios with 2 parents and one affected offspring. Significant association ($p < 0.005$) was found with several SNPs located in the promoter area of this gene. Fine linkage disequilibrium (LD) map of the promoter identified three main blocks of strong LD. Multivariate analyses showed that only two SNPs were needed to capture almost all association information. Based on these two SNPs, we identified a risk haplotype associated with a 5-fold increase in the risk of leprosy when compared to the protective haplotype ($p < 0.001$). In order to validate this result, a case-control study was carried out in an independent sample from Brazil including 708 cases and 394 controls matched on ethnicity. The very same pattern of results was observed with 10 out of the 15 selected SNPs being strongly associated with leprosy per se in this new sample (p-values ranging from 0.03 to 10-6). Based on the predicted function of this gene, we postulate that a new physiopathological pathway implicated in the onset of leprosy has been identified.

3

Mathematical Models of Microdeletions using SNPs or STRPs

C.I. Amos (1), S. Shete (1), J. Chen (1), R. Yu (1, 2)
(1) Department of Epidemiology, U.T. M.D. Anderson Cancer Center, Houston, TX, USA and (2) Computer Science Department, Rice University, Houston, TX, USA

Microdeletions have been associated with several complex diseases including autism and schizophrenia and also cause numerous simple genetic diseases such as neurofibromatosis. To date, no statistical basis has been available for the identification of microdeletions in family studies. Here, we present an approach to the identification of either stably inherited or de novo microdeletions for parent-affected offspring trios. The method is sufficiently general for application using either single nucleotide polymorphisms or microsatellites. In addition to general algorithmic development, we have also studied the behavior of our methods for SNPs under varying

Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/gepi.10276

conditions including heterogeneity in causation, with only some cases being caused by microdeletions as well as variable error rates in the SNP genotyping. The introduction of genotyping errors or causal heterogeneity was found to have approximately linear effects upon loss of information to detect microdeletions. Because we anticipate that error rates will generally be low (less than a few percent), while heterogeneity could occur at high frequencies we find that heterogeneity has a more severe influence upon the ability to identify de novo microdeletions.

4

Polygenic inheritance of breast cancer: implications for the design and analysis of association studies.

AC Antoniou, DF Easton
CRUK, Genetic Epidemiology Unit, Cambridge Univ, UK

Susceptibility to breast cancer is likely to be the result of susceptibility alleles in many different genes. In particular, one segregation analysis has suggested that disease susceptibility in non-carriers of BRCA1/2 mutations may be explicable in terms of a polygenic model with large numbers of susceptibility polymorphisms acting multiplicatively on risk. We considered the implications for such a model on the design of association studies to detect the susceptibility polymorphisms, in particular the efficacy of utilizing cases with a family history of the disease, together with unrelated controls. Relative to a standard case-control association study with cases unselected for family history, the sample size required to detect a common disease susceptibility allele was typically reduced by more than twofold if cases with an affected first degree relative are selected, and by more than fourfold if cases with two affected first degree relatives are utilized. The relative efficiency obtained by using familial cases was greater for rarer alleles. Analysis of extended families indicated that the power is most dependent on the immediate (first degree) family history. Bilateral cases may offer a similar gain in power to cases with two affected first degree relatives. In contrast to the strong effect of family history varying the ages at diagnosis of the cases across the range 35-65 did not strongly affect the power to detect association. We also investigate the implications of the polygenic model for the parameter estimation in association studies based on familial cases.

5

Resampling distribution of the location of the maximum lodscore

L.D. Atwood, N.L. Heard-Costa
Boston University Medical School, Boston, MA, USA

Once a significant linkage is found and confirmed interest turns to the estimated location of the linked locus. Fast computers now make it possible to compute the resampling distribution of the estimated location. To explore that distribution we used GASP to simulate a quantitative trait, affected by a QTL that accounted for 40% of the total variation, in 200 nuclear families of size 5, on a 160cM chromosomal segment, with markers every 10cM. Variance components linkage analysis (Genehunter) was performed on each replication until 200 were obtained with maximum lodscore greater than 3.0. For each replication we re-sampled 800 times. Each resample comprised 200 families sampled from the original 200 with replacement. Linkage analysis was repeated on each resample. Since maximum lodscores tend to occur at marker locations (marker bias), we examined two cases; one in which the QTL was between two markers (5cM from either flanking marker) and one in which the QTL was at a marker (0cM from the marker).The resulting 200 distributions showed a wide variety of shapes, skewed and symmetric, one peak and multiple peaks. When the QTL was between two markers, 73.5% of the replications had their mode at one of the closest flanking markers. In those replicates where the mode accounted for at least half of the distribution (i.e. 400 of the 800 re-samples), 90% had their mode at one of the flanking markers. When the QTL was at a marker, 66.5% of the replicates had their mode at that marker (i.e. the closest marker). In those replicates where the mode accounted for at least half the distribution, 84.8% had their mode at that closest marker. We conclude that the mode of the resampling distribution is a good indicator of the reliability of the location estimate.

6

A Franco-American genome scan for Multiple Sclerosis

M-C. Babron (1)J.L. Haines (2), the French and American Multiple Sclerosis Genetics groups, J.R. Oksenberg (3), M.A. Pericak-Vance (3), B. Fontaine (4), F. Clerget-Darpoux (1)
(1) INSERM U535, Villejuif, France, (2) Vanderbilt University Medical Center, USA, (3) Duke University Medical Center, USA, (4) Fédération de Neurologie, Paris, France

Multiple sclerosis (MS) is a chronic inflammatory disorder of the central nervous system. MS is a multi-factorial disease with a strong genetic component. The only genetic risk factor identified to date resides in the HLA region. To evidence other risk factors, a genome scan was performed on the sample of 243 multiplex families issued from a French and US collaboration. The families were recruited in France (94 families) and the US (149 families), using the same clinical criteria. A genome scan with 356 microsatellite markers was carried out to search for linkage using the Z statistics of Kong and Cox (1997). The strongest signal is observed on chromosome 1q in the total sample (Z=3.38). Among the 5 other regions of interest with $Z > 2$, are the HLA region (Z=2.26), and the 5q31-33 region (Z=2.17). The presence of a risk factor in region 5q31-33 is particularly interesting since it has already been shown in several auto-immune diseases. Furthermore, this region is homologous to the rat chromosome 10 region where a susceptibility locus for Experimental Autoimmune Encephalomyelitis has been mapped.

7

Polymorphisms of Phase I and II Detoxification Enzymes and Lung Cancer Risk

A.B. Baffoe-Bonnie, C. Spittle, R. Michielli, H. Wang, S. Brusstar, A. Balshem, M. Unger, W.T. London, J. Testa, M. Clapper
Division of Population Science, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, PA 19111

Individual susceptibility to lung cancer from exposure to tobacco smoke may be influenced by polymorphic variation in Phase 1 and II detoxification enzymes. We examined the relationship between susceptibility to lung cancer and polymorphisms in three genes whose enzyme products are involved in the metabolism of procarcinogens found in tobacco smoke. We evaluated the effect of the MPO (Myeloperoxidase) -463 (G → A) polymorphism, CYP1A1 and GSTM1. We tested 96 incident lung cancer cases (mean age 61.4 ± 9.7 years) and 96 controls (mean age 61.06 ± 9.9 years) among Caucasians. Controls were healthy volunteers who participated in the Bio-sample Repository at Fox Chase Cancer Center. Cases and controls were comparable in terms of demographic factors. There were more cases who were current smokers (13.4%) compared to 8.4% of controls. Cases were more likely to be long-term heavy smokers compared to controls (85.9% vs. 57.8%). Genotyping was performed by pyrosequencing. The statistical significance of the differences in the frequencies of each gene between groups was calculated by Fisher's exact test (1 tailed). Binary logistic regression was used to assess the odds ratio (*OR*) and confidence intervals between genotypes. Investigation of the relationship between cases and controls comparing GSTM1 wild type to null was not significant. Comparing cases and controls, CYP1A1 homozygous A/A was found to be significantly protective in lung cancer (*OR*, 0.31, 95% CI, 0.11–0.91, p=0.018). Similarly, a single copy of the MPO (A/A and A/G) was inversely associated with susceptibility to lung cancer (*OR*, 0.58, 95% CI, 0.32–1.02, p=0.04). Our preliminary results confirm previous reports showing the protective effect of both the MPO (A/+) and CYP1A1 (AA).

8

The relative efficiency of haplotype and genotype data for fine-scale LD mapping

D.J. Balding (1), A.P. Morris (2), J.C. Whittaker (1)
(1) Department of Epidemiology and Public Health, Imperial College London, UK (2) Wellcome Trust Centre for Human Genetics, Oxford, UK

A common approach to the analysis of multi-locus genotype data for linkage disequilibrium (LD) fine mapping is a two-stage strategy of first inferring the haplotype pairs underlying the genotypes, and then implementing a haplotype-based analysis. This is unsatisfactory for a number of reasons. In particular, the uncertainty arising in the haplotype inferral stage is usually ignored in the second stage. Because true haplotypes are more informative for fine mapping than the corresponding genotypes, users may erroneously believe that inferred haplotype pairs are also more informative than genotypes. We present the results of a simulation study which indicate that genotype data, appropriately analyzed, often provide almost as much information for LD fine mapping as do true haplotype pairs, and generally provide more information than do inferred haplotypes. To perform the study, we have extended our existing coalescent-based Markov chain Monte Carlo algorithm for LD fine mapping, which previously accepted only haplotype data (Morris et al., 2002). The resulting COLDMAP program treats the unknown phases as latent variables and approximates an integration over their possible assignments.

9

Efficiency of the Cladistic Association Analysis

C. Bardel(1), V. Danjean(2), P. Darlu(1), E.Génin(1)
(1) U535, INSERM, France, (2) LaBRI, UMR 5800, Bordeaux, France

With the development of molecular techniques to identify genetic polymorphisms, numerous markers are now available within and between candidate genes. Different haplotypic methods have been proposed to use the joint information provided by different markers to detect the role of a given genomic region in disease susceptibility. One such method has been proposed by Templeton and uses a cladistic approach. The method consists in building a phylogeny of the haplotypes and comparing the proportion of haplotypes carried by affected and unaffected individuals within each clade. If a clade contains an excess of affected individuals, one can conclude that mutations defining this clade may be involved in the disease. This strategy has been applied on some real data by different authors, but its statistical properties have never been assessed. In this study, we have performed extensive simulations to determine the power of the method to detect an association with a genomic region involved in disease susceptibility and once the association is detected, the efficiency of the method to identify the functional variant among the different polymorphisms tested. We also carried out a comparison with two other methods: one that compares the whole distribution of haplotypes in cases and controls and the other that consists in the separate analysis of the different sites. We have also applied our method to a real data set concerning the CARD15 gene and Crohn disease. We were able to successfully identify the three variant sites that are involved in the disease susceptibility.

10

Accuracy of statistical tests for haplotype analysis using Mantel statistics

L. Beckmann(1), C. Fischer(2), D. C. Thomas(3), J. Chang-Claude(1)
(1)German Cancer Research Center DKFZ, Heidelberg, Germany (2)Dept. of Human Genetics, Univ. of Heidelberg, Germany (3)Dept of Preventive Medicine, Univ. of Southern California, Los Angeles, CA, USA

Complex diseases are attributable to the joint effect of genetic and non-genetic risk factors. We used Mantel statistics to correlate genetic with phenotypic similarity across pairs of haplotypes while accounting for covariates. Mantel statistics have the form $\Sigma_{i,j}X_{ij}Y_{ij}$. In our context X_{ij} is a measure of genetic similarity defined as the shared length between haplotypes i and j,Y_{ij} a measure of phenotypic similarity, defined by affection status or a covariate and the summation is over all pairs of haplotypes. The basic idea is that individuals who have similar phenotypes also have similar haplotypes around a disease-causing variant. Several different variants of Mantel statistics were considered in case-only and case-control studies. Simulated data showed that incorporating covariates improved the power of haplotype-sharing analysis to localize disease genes. We analyzed three tests for statistical significance with respect to type I error and power: (1) based on the assumption of asymptotic normality; (2) by fitting Pearson curves based on the exact first four moments; and (3) a Monte Carlo permutation procedure. Simulation studies indicate that the permutation procedure is the most suitable test. To further improve the power we are also investigating different definitions of genetic similarity under varying conditions such as marker spacing and variations in linkage disequilibrium.

11

Gene-environment modeling in incomplete trios

J. Beyene, C.M.T. Greenwood, J. Lee, A.D. Paterson
Hospital for Sick Children and University of Toronto, Canada

Comparison of rates of transmission of alleles from parents to affected offspring is often carried out using family-based association tests such as the Transmission Disequilibrium Test (TDT). For complex traits adjustment for covariates may enable detection of gene-environment interactions and can improve the power of these tests. However, missing parental marker genotypes pose problems and usually lead to exclusion of families that could potentially result in a significant loss of power. In this study we apply the log-linear modeling framework of Weinberg (1999, AJHG 65:229–235) with one binary covariate. We have extended the covariate model to allow for trios missing one parent by using an extension of the Expectation-Maximization (EM) algorithm also described by Weinberg. This approach allows us to investigate different modes of inheritance, to estimate the relative risk at the marker, and also provides a framework for multilevel modeling. The multilevel modeling approach can also be useful in elucidating the effect of measurement error on estimation and testing of model parameters. We will demonstrate the potential of the approach with real and simulated data sets.

12

One gene or two? Methods for estimation and testing for two linked disease genes

J.M. Biernacka (1, 2), S.B. Bull (1, 2)
(1) Dept. of Public Health Sciences, Univ of Toronto, Canada, (2) Samuel Lunenfeld Research Institute, Toronto, Canada

Consideration of multiple susceptibility genes may increase the power to detect genes in complex disorders. We developed a model for simultaneous localization of two linked trait genes. Expected allele sharing in affected sib pairs (ASPs) in a region containing two susceptibility genes is a function of four parameters: the locations and expected IBD sharing in ASPs at the two genes. Generalized estimating equations (GEE) can be used to estimate these parameters as was proposed by Liang et al (Hum Hered 51, 2001, 64–78) for the case of one disease gene in a region. We developed an algorithm that uses marker IBD sharing in ASPs to estimate the locations and expected IBD sharing parameters. We studied properties of estimates obtained by this method using simulation. Here we propose several procedures to evaluate the evidence for two versus one-disease loci in a region based on wald, quasi-likelihood, and quasi-score statistics. Simulation results show that variances of parameter estimates based on the GEE are biased downward, which has implications for test statistics that rely on these estimates. We applied the proposed methods to data from a diabetes genome-scan (Nature Genet 19: 297–300, 1998) and found suggestive evidence for two linked disease genes on chromosome 16. Our results indicate that the power to identify two linked disease genes depends on IBD sharing at the two loci, the distance between them and sample size. The proposed method can improve disease gene localization when two disease genes are present in a region.

13

On testing causality of genotype data for binary traits in incomplete nuclear families

S. Boehringer, A. Steland
Institut für Humangenetik, Essen, Germany

In genetic data sets observations of alleles can be highly correlated on account of high linkage disequilibrium (LD) and therefore hamper the effort to pinpoint causal loci influencing some phenotypic trait. We exploit the interesting fact that conditioning on genotypes the phenotypes of parent-offspring pairs are uncorrelated whenever a causal locus is observed. The dependence structure induced by the genetic mechanism is analyzed in detail. The corresponding likelihood can be generalized to more extensive pedigrees. A likelihood-ratio test is presented to test hypotheses about the conditional correlation as quantified by the LD parameter δ of an allele at an observed candidate locus and an allele at a linked unobserved causative locus. The statistical properties of the test are studied by simulations. Power turns out to be excellent for many common situations (e.g. allele frequencies 0.3 at both loci, LD: δ'=0.4, penetrance: f=0.8 result in power of 0.97 with just 80 pairs). Covariates can be included into the analysis by specifying an appropriate penetrance function. Furthermore the residual covariance of phenotypes caused by additional,

unlinked loci influencing the phenotype can be decomposed from the influence of the loci under scrutiny. Applications of this method include fine mapping efforts and the assessment of contributions of individual polymorphisms to traits. Additionally the method can be used to reconstruct haplotypes in incomplete or ambiguous nuclear families.

14

Haplotype associations on 7q with BMI in the Family Heart Study (FHS)

I.B. Borecki(1), Y. Jiang(2), J.B. Wilk(2), A.L. DeStefano(2), R.H. Myers(2), M.A. Province(1)
(1) Div Biostatistics, Washington Univ Schl Med; (2) Dept Neurology, Boston Univ Schl Med

Strong evidence of linkage with BMI was reported on 7q31-34 (lod=4.9) in the NHLBI FHS. The maximum lod score was within 3 cM of the leptin gene (LEP), an obvious candidate. To identify associations, 99 SNPs were typed within LEP and in the~10 Mb region around the peak lod score. Focusing on the 80 families with the strongest evidence of linkage (accounting for~80% of the maximum LOD score) 640 individuals were typed. Association was assessed by a TDT as implemented FBAT and TRANSMIT. Three regions defined by several SNPs in each were significantly associated with BMI ($P < 0.01$). Only one of these regions is located in LEP, suggesting the possibility of other relevant genes in the region. These initial findings were pursed using several statistical approaches. We tested all possible haplotypes created from the SNPs within LEP showing marginally significant association, using a bootstrap approach to account for the familial dependency. Haplotype effects were also investigated using the inferred block structure and the 2-stage hierarchical method proposed by Conti and Witte. We will compare these results with a variance components analog in which both LD and familial dependencies are simultaneously estimated in a 1-stage procedure. We address the multiple comparisons issue by contrasting several approaches including a newer Benjamini-Hochberg correction and empirical estimates of p-values.

15

Genome scan for asthma and atopy related phenotypes in the French EGEA study

E Bouzigon, MH Dizier, D Torchard, I Annesi-Maesano, C Betard, J Bousquet, F Gormand, M Guilloud-Bataille, J Just, J Maccario, R Matran, N Le Moual, F Neukirch, MP Oryszczyn, E Paty, I Pin, D Vervloet, J Feingold, F Kauffmann, M Lathrop, F Demenais INSERM EMI0006, Evry, France

A genome scan for asthma and atopy-associated phenotypes was conducted in the whole sample of EGEA families selected through at least one asthmatic subject (295 families) and in a subset with at least two asthmatic sibs (110 families), using a panel of 396 markers. The following phenotypes were considered: asthma, positive skin prick test response (SPT) to 11 allergens, Multi-RAST Phadiatop, PD20 methacholine, total IgE levels, eosinophil counts (EOS) and percent predicted FEV1 (%FEV1). Linkage was investigated by model-free methods: Maximum Likelihood Binomial approach for binary traits and both Haseman-Elston and variance components methods for quantitative measures, using GENEHUNTER and MERLIN. Four potential regions of linkage ($p < 0.001$) were detected in the whole sample for two atopy-related phenotypes (IgE, SPT) and a measure of lung function (%FEV1): 6q14 (%FEV1, LOD=2.94), 12p13 (IgE, LOD=2.07), 17q22 (SPT, LOD=1.96), 21q21 (%FEV1, LOD=2.81). Our most significant result was found in the family subset with a higher proportion of asthmatics: 13q21-31 linked to EOS (LOD=3.38, p=0.00004). Evidence for linkage to other phenotypes, including asthma, was weaker ($p < 0.005$). Our results show that the mode of ascertainment of family samples has a major impact on linkage detection by selecting different underlying physiological pathways.

16

Bayesian Investigation of Stoppage Effects in Segregation Analysis

M. Brimacombe
Department of Preventive Medicine & Community Health, New Jersey Medical School - UMDNJ, Newark, NJ

In the case of some complex, relatively rare diseases, for example severe autism, segregation analysis can be biased by the presence of stoppage effects. These refer to settings where the onset of the disease itself affects the desired family size and related family size distribution. The likelihood function for these settings was developed in Slager et al. (2001). Here we extend the analysis to settings where a Bayesian approach is helpful in assessing stoppage related bias effects across various assumed family size distributions. Both standard Bayes and empirical Bayes approaches are discussed. Autism related family size data for a cohort of 100 patients is the basis of discussion.

17

Other cancer risks among BRCA1/2 mutation carriers in Dutch families

R.M. Brohet (1), H. Meijers-Heijboer (2), C.J. van Asperen (3), L.J. van 't Veer (1), M.A. Rookus (1), F.E. van Leeuwen (1)
(1) Dept. of Epidemiology & Pathology, The Netherlands Cancer Institute, Amsterdam, (2) Dept. of Clinical Genetics, Erasmus MC, Rotterdam, (3) Dept. of Clinical Genetics, Leiden University Medical Center

To evaluate the risks of other cancers in BRCA1/2 mutation carriers, we conducted a retrospective cohort study of 5116 and 1260 individuals from 410 BRCA1 and 103 BRCA2 families. Studies based on typed carriers could produce an overestimation of cancer risks if affected family members are more likely to go for testing. To overcome this possible testing bias we selected family members with a 50% probability of being a carrier. The observed cancer incidence was compared with the expected site-, sex-, and period-specific

cancer incidence in the Dutch population. Relative Risks (RR) among carriers of each cancer type were then derived from the RR found in the selected cohort of 50% presumed carriers. There was some evidence of an elevated risk in BRCA1 mutation carriers of cervical (RR=3.0; 95% Confidence interval (CI): 2.3–5.6), endometrial (RR=2.0; 95% CI:1.3–3.0) and bone cancer (RR=7.0; 95% CI:1.9–17.0). In contrast, a decreased risk was found for prostate and lung cancer (RR=0.1; 95% CI:0-0.6, RR=0.2; 95% CI:0–0.5, respectively). BRCA2 was associated with an increased risk of pancreatic (RR=3.7; 95% CI:1.0–8.6) and larynx cancer (RR=6.0; 95% CI: 1.8–13.4). In conclusion, there may be an increased risk for abdominal cancers in BRCA1 mutation carriers and pancreatic and larynx cancer in BRCA2 mutation carriers. No increased risk of prostate cancer was found among BRCA2 mutation carriers.

18

Phenotypic Item Clustering to Find Simpler Inheritance Patterns

S. G. Buyske
Statistics Dept, Rutgers University

There is evidence that at least some behavioral disorders may actually be the conjunction of multiple phenotypes. If we knew how to untangle the overall phenotype, the individual components of the phenotype should be genetically simpler and so easier to map than the original complex trait. We show how the genetic segregation patterns, by family, of phenotype item information can be used to cluster items into genetically simpler phenotypes. These new synthetic phenotypes could lead to greater power in genetic linkage studies. Results are shown for simulated data.

19

Genetic Correlations Between Adiposity and Liver Function in Oman Population

G. Cai(1), R. Bayoumi(2), A.G. Comuzzie(1), F. Al-Ubaidi(2), S. Al-Yahyaee(2), S. Albarwani(2), J. Al-Lawati(3), A. Jaffer(3), M. Al-Kindi(2), H. Al-Barwany(2), and M. Hassan(2)
(1) Dept. of Genetics, Southwest Foundation for Biomedical Research, USA; (2) College of Medicine, Sultan Qaboos University, Muscat, Oman; (3) Ministry of Health, Muscat, Oman

The genetic correlation between physiological measures associated with nonalcoholic steatohepatitis (NASH) and measures of adiposity were investigated in 630 adult individuals of Omani origin distributed across 6 consanguineous pedigrees. Quantitative genetic analyses were conducted using SOLAR. A substantial proportion of phenotypic variation in adiposity measurements was due to a genetic component, with heritabilities (h2) ranging from 0.31 to 0.68. Parameters of liver function [albumin, total bilirubin, alanine aminotransferase (ALT)] were also heritable (h2 from 0.24 to 0.53). Bivariate genetic analyses showed that albumin was genetically correlated with percent body fat ($rg=-0.48\pm0.16$, p=0.004), marginally with waist circumference ($rg=-0.38\pm0.19$, p=0.06), and with BMI ($rg=-0.47\pm0.14$, p=0.003). Similar genetic correlations were observed between total bilirubin and percent body fat ($rg=-0.33\pm0.13$, p=0.01), waist circumference ($rg=-0.47\pm0.16$, p=0.005), and BMI ($rg=-0.39\pm0.13$, p=0.003). ALT was genetically correlated with BMI ($rg=0.31\pm0.14$, p=0.045), but not with percent body fat or with waist circumference. Therefore, decreased levels of albumin and total bilirubin and elevated ALT were genetically correlated with increased adiposity levels. These results suggest the existence of genes with pleiotropic influence on liver metabolism and adiposity, and offer a potential connection between fatty liver disease and obesity.

20

A new perspective on the prospects of genealogical/ medical data resources - the Utah example

LA Cannon-Albright, JM Farnham, A Thomas, NJ Camp
University of Utah Medical Informatics; Intermountain Health Care, USA.

A Utah population database (UPDB) was created in the 1970's, consisting of genealogical records for descendants of the Utah pioneers. Statewide cancer records were linked, allowing estimation of the familial and genetic nature of cancer. Studies of specific UPDB pedigrees led to localization of multiple cancer predisposition genes. It has been argued that genetic isolates might represent a more valuable resource in the identification of genes for complex phenotypes. Although Utah does not represent an isolate population; pseudo-isolate subpopulations can be "created" by pruning the UPDB. A set of founders of any size can be selected and only matings of founders and/or their resulting descendants considered; matings with ineligible founders occurring in the population are censored. Bottlenecks and immigration can be controlled. The current UPDB has 2 million individuals who are part of 3+ generation pedigrees. We identified the 27,850 Utah Mormon pioneers and have created a Mormon pseudo-isolate of 299,148 individuals. We have examined this and other pseudo-isolates to identify excesses of specific cancers and isolate-specific cancer pedigrees. Analysis of these pedigrees may allow the localization of rare cancer predisposition genes whose effect may be masked in the more heterogeneous full Utah pedigree set. Extensions to other subpopulations (e.g. inbred groups) have also been considered. We discuss extending the utility of the Utah resource by creating relevant subpopulations for genetic study of specific disease targets.

21

Evidence for Modifying Effects of MC1R Gene, Sunburns and Atypical Nevi on CDKN2A Penetrance in 20 French Melanoma-Prone Pedigrees

V. Chaudru(1), A. Chompret(2), A. Minière(2), K. Laud(2), M.F. Avril(2), B. Bressac-de Paillerets(2), F. Demenais(1)
(1)INSERM, EMI0006, Evry, France, (2)Institut Gustave Roussy, Villejuif, France

Germline mutations in CDKN2A gene predispose to melanoma with high but incomplete penetrance, suggesting effects of other genetic and/or environmental factors. Case-control studies have shown that nevus phenotypes, pigmentary traits, skin reactions to sunlight, and sun exposure influence melanoma risk. Variants of MC1R (melanocortin-1 receptor) gene were found to be associated with red hair and fair skin as well as melanoma. Our goal was to examine joint effects of MC1R variants (R151C, R160W, D294H), nevus phenotypes, pigmentary traits and sun-related covariates on CDKN2A penetrance. Clinical, genetic, and covariate data were recorded in 20 French melanoma-prone families with co-segregating CDKN2A mutations. Analysis of the co-transmission of melanoma and CDKN2A alleles was conducted using regressive logistic models, which allow taking into account a variation of disease risk with age and the aforementioned risk factors. Tests for significant risk factors were conducted using a stepwise procedure. Factors which modify significantly CDKN2A penetrance include sunburns (OR=3.97), atypical nevi (OR=3.49), and MC1R variants (OR=2.64). In CDKN2A mutation carriers, the cumulative risk of melanoma is 0.58 by age 80 years and reaches 0.97 by adding presence of MC1R variants and 1.0 with atypical nevi or history of sunburns. Our study shows that several pathways are likely to be involved in the development of melanoma.

22

Power of Two Alternative Formulations of the Regressive Models to Detect Gene-Environment Interactions in Complex Diseases

V. Chaudru, M. Rosenberg-Bourgin, F. Demenais
INSERM, EMI0006, Evry, France

We have previously shown that model-based methods are of interest to identify causal genetic variants and to detect gene × environment (G × E) interactions in complex diseases. Our present goal was to compare the performances of two formulations of the regressive models to detect G × E when G is in linkage disequilibrium (LD) with a marker (SNP). Affection status, environmental factor and marker data were simulated in 165 families of varying sibship size. The liability to disease was generated under a model including a common gene (G) with a small effect (VG=5%), a polygenic component (VP=40%) and a binary environmental factor interacting with G (VG × E=15%). We generated a tightly linked SNP with varying degree of LD with G: no LD (D′=0), CLD (D′=1), ILD (D′=0.5). One hundred replicates of the simulated data were analyzed using either the regressive threshold model (RTM) or the regressive logistic model (RLM). RTM is more general than RLM since it allows to adjust the phenotypes of each individual's antecedents for their own genotypes and covariate effects. Power of RTM to detect G × E was between 65–82% when D′=1, 25–28% when D′=0.5, 7–13% when D′=0, depending on the characteristics of G. Power was 0.3 to 0.9-fold decreased when using RLM. Alternatively, evidence for the causal genetic variant (fit of CLD) was reduced by 55–83% when ignoring G × E as compared with taking it into account, this impact being slightly greater for RTM than for RLM. Thus RTM has more power to detect G × E than RLM and taking G × E into account appears of major importance to identify the causal variant.

23

A Comparison of Combined Linkage and Association Methods for Mapping Quantitative Trait Loci

I. Chazaro(1, 2), L. Atwood(1), J. Dupuis(1), R. D'Agostino(1, 2), L.A Cupples(1)
1) Department of Biostatistics, Boston University, USA, 2) Department of Mathematics and Statistics, Boston University, USA

There is growing pessimism over the ability of linkage or association studies used independently to identify quantitative trait loci (QTLs). Linkage studies require large sample sizes to reach moderate power for QTLs with small effects on quantitative traits. In addition, they are often criticized for having large location errors, biased effect estimates and inflated type-I error when the phenotypes are not normally distributed using a variance component approach. Conversely, association studies using unrelated subjects are prone to false positive results in the presence of population admixture. To address some of these concerns, several investigators have proposed combined linkage and association methods for mapping QTLs using nuclear families or extended pedigrees. We compared four of these methods in nuclear families: the generalized linear model approach of Abecasis (*2000* Am. J. Hum. Genet, 71:1330–1341), the variance component extension proposed by Almasy (*1999* Genet. Epi. 17:S31–S36), the family based association test of Rabinowitz (*2000* Hum. Her. 50:211–223) and the measured genotype approach of Boerwinkle (*1986* Ann Hum Genet.;50:181–194). We used simulations to compare these methods in terms of power and bias assuming normally distributed phenotypes and diallelic QTLs. We also considered QTLs with more than two alleles and/or several polymorphic sites. Finally, we examined the sensitivity to the normality assumptions of these four methods.

24

Meta-analysis of four Coronary Heart Disease genome-wide linkage studies confirms a susceptibility locus on chromosome 3q

B.D. Chiodini, C.M. Lewis
Div. of Genetics and Development Guy's, King's and St. Thomas' School of Medicine, London, UK

Objective: In coronary heart disease (CHD), four independent genome-wide screens have now been published, using Finnish, Mauritian, European and Australian families. Results from these studies are

inconclusive. We performed a meta-analysis to identify genetic regions that show evidence for susceptibility genes across studies. Methods and Results: The rank-based Genome Scan Meta-Analysis (GSMA) method was applied to the four CHD genome-wide linkage studies. The strongest evidence for linkage was found on chromosome 3q26–27 (p=0.0001) and 2q34–37 (p=0.009). An analysis weighted by study size confirmed linkage in these regions (3q26–27: p=0.0002, 2q34–37: p=0.014). Conclusions: The genetic regions 3q26–27 and 2q34–37 may contain susceptibility genes for CHD. Linkage to 3q26-qter region has previously been shown in type 2 diabetes mellitus, metabolic syndrome, cholesterol concentration in LDL size fractions and renal function in hypertensive subjects. The 2q34–37 region lies close to the type-2 diabetes NIDDM1 locus. Both of these regions harbor several candidate genes involved in the homeostasis of glucose and lipid metabolism. These results are particularly intriguing given the growing evidence of an association between CHD risk and metabolic abnormalities, such as insulin resistance, type-2 diabetes, abdominal obesity and dyslipidemia.

25

Disease susceptibility gene identification: can one afford to exclude patients from the analysis?

F. Clerget-Darpoux, P. Margaritte-Jeannin
INSERM U535, Hôpital Paul Brousse, Villejuif, FRANCE

A two-step strategy is often applied in the identification of genetic risk factors for a multifactorial disease. The first step consists in performing a linkage genome scan on affected sib pairs in order to detect candidate regions within which a risk factor is located. The second step aims at searching for association with intragenic polymorphisms within a candidate region in order to identify the genetic risk factor. The same sib pair sample used in the linkage analysis may also be used for the association studies, selecting only one of the two sibs (index case). Many recently published studies further restricted this association step to the index of sib pairs sharing two parental alleles identical by descent (IBD=2). The underlying argument is gain of power in presence of genetic heterogeneity. This belief is false, as illustrated by the following example. Consider a sample of 200 sib pairs in which a disease allele of frequency 0.35 plays a dominant role in 30% of the sib pairs and does not play any role in the remaining 70%. The association test on all 200 index of the affected sibs is highly significant (χ^2=5.66; p=0.02) while not significant if only applied to the 55 index of the IBD=2 affected sib pairs (χ^2=2.17; p=0.14). More generally, we show analytically that, whatever the degree of heterogeneity and whatever the model underlying the effect of the risk factor in the disease, restricting the association study to the sub-sample of sib pairs IBD=2 always causes a loss of power.

26

Genotypes, Haplotypes, and SNP Selection in Candidate Gene Regions

DV Conti, WJ Gauderman
Department of Preventive Medicine, University of Southern California, USA.

Modern molecular techniques make discovery of numerous single nucleotide polymorphims (SNPs) in candidate gene regions feasible. Conventional analysis relies on either independent tests with each variant or the use of haplotypes in association analysis. The first technique ignores the dependencies between SNPs, while the second, though it may increase power, often introduces uncertainty by estimating haplotypes from population data. Additionally, as the number of loci expands, ambiguity in haplotype estimation increases and the specific causal variant may go undetected. Here, we present a genotype-level analysis to jointly model the SNPs and we introduce a modified SNP $\times$ SNP interaction term to capture the underlying haplotype structure. This analysis estimates both the risk associated with each variant and the importance of phase between pairwise combinations of SNPs. The method readily expands to incorporate uncertainty in phase and allows a correction for bias due to over-sampling of cases in case-control data. To avoid unstable estimation due to sparse data, we propose a Bayes model averaging procedure, which highlights key SNPs and phase terms while incorporating uncertainty in model selection. Prior distributions are modified with genetic information such as haplotype tagging of block regions. Coalescent simulations show that this method correctly identifies causal variants while distinguishing crucial SNPs and underlying haplotype structures influencing disease. We demonstrate the performance of this method under various genetic scenarios using simulations and discuss an application to real data.

27

Modeling Extant Temporal Trends Can Increase the Power of Linkage Analysis

J. Corbett, M.A. Province, D.C. Rao
Division of Biostatistics, Washington University, St. Louis, Missouri, USA

Heritability for some complex traits like BMI and blood pressure has been shown to vary in some cases dramatically, with age. Intuitively, allowing for varying heritability, when it is present in sufficient magnitude, should increase the power to detect linkage. In this simulation study, we examine a quantitative trait with temporal trends in heritability using two distinct values of heritability, one low (10%) and one high (50%) heritability groups. Heritability was assumed to derive from two unlinked genes, each acting additively and accounting for half of the heritability in each age group. Markers were assumed to be fully informative and completely linked to the trait loci. Simulations were performed with total sample sizes of

500, 750, and 1,000 sib-pairs, with a 1:2 ratio of high to low heritability sib-pairs and were replicated 1,000 times. Two linkage analyses were performed on each replication. The first assumed there were no trends in heritability, while the second allowed for trends by letting heritability vary between the two groups. As there are two degrees of freedom (d.f.) in the latter analysis, the distribution of the test statistic under the null hypothesis is a $\frac{1}{4}:\frac{1}{2}:\frac{1}{4}$ mixture of a chi-square with 2 d.f. a chi-square with 1 d.f. and a point mass at zero. After adjusting to a 1 d.f. equivalent, the trends analysis showed, on average, about 50% higher LOD scores than the trend-free analysis. Clearly, the increased power due to allowing for trends will depend upon the magnitude of trends present, marker informativeness, etc. but this simulation demonstrates that it is very promising.

28

Effect of recombination, ascertainment and multiple affected offspring on case/pseudo-control analysis

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

The case/pseudo-control approach is a general framework for family-based association analysis, incorporating several previously proposed methods such as the transmission/disequilibrium test and log-linear modeling of parent-of-origin effects. Here I examine the properties of methods based on a case/pseudo-control approach when applied to a linked marker rather than (or in addition to) the true disease locus or loci, and when applied to sibships that have been ascertained on, or that may simply contain multiple affected sibs. Through simulations and analytical calculations I show that the expected values of the observed relative risk parameters (estimating quantities such as effects due to child's own genotype, maternal genotype and parent-of-origin) depend crucially on the ascertainment scheme used, as well as on whether there is non-negligible recombination between the true disease locus and the locus under study. In the presence of either recombination or ascertainment on multiple affected offspring, methods based on conditioning on parental genotypes are shown to give unbiased genotype relative risk estimates at the true disease locus (or loci) but biased estimates of population genotype relative risks at a linked marker, suggesting that the resulting estimates may be misleading when used to predict the power of future studies. Methods that allow for exchangeability of parental genotypes are shown (in the presence of either recombination or ascertainment on multiple affected offspring) to produce false positive evidence of maternal genotype effects when there are true parent-of-origin or mother-child interaction effects, even when analyzing at the true locus. These results suggest that care should be taken in both the interpretation and application of parameter estimates obtained from family-based genetic association studies.

29

A comparison of scores for MCMC linkage analysis

E. W. Daw, J. Ma
Dept. of Epidemiology, UT M.D. Anderson Cancer Center, Houston, TX.

The Bayesian Monte Carlo Markov chain (MCMC) techniques implemented in the program Loki (Heath, 1997) have shown the ability to localize genes for complex traits in both real and simulated data sets. These methods estimate the posterior probability over a complex model space that includes the number, position, and effect of quantitative trait loci (QTL). Unfortunately interpretation of this posterior probability surface has been difficult. Here, we examine and compare several summary scores for assessing the significance of linkage findings. These scores include: (1) the Log Of the Posterior placement probability ratio (LOP), which is the log of the posterior probability of linkage to the real chromosome divided by the posterior probability of linkage to a randomly generated unlinked pseudo chromosome with marker information similar to the marker data on the real chromosomes (Daw et al. 2003); (2) The ''L-score'' produced by Loki which is the marginal posterior probability of linkage over a 1cM interval divided by the prior; (3)The marginal posterior probability over chromosome position and locus variance contribution. All scores must be estimated, and in the case of the LOP, several estimates are considered. We estimate scores on a simulated 350-member, 5-generation family with several missing data patterns and on nuclear families drawn from the extended pedigree. We analyze several simulated traits with different genetic variance contributions. We find that empirical significance levels for the three scores are similar.

30

Strategies for excluding genomic regions from multivariate linkage analysis

M de Andrade (1), C Olswold (1), SLR Kardia (2), E Boerwinkle (3), ST Turner (4)
(1) Health Sciences Research and (4) Hypertension, Mayo Clinic, USA; (2) Human Genetics, Univ. of Michigan, USA, (3) Human Genetics Center, UT Health Science Center, USA

Although biometrical analyses consistently indicate that genetic variation accounts for a substantial percentage of inter-individual variation in BP (30–50%), linkage and association studies to identify responsible genes have provided inconsistent findings. Despite application of state-of-the-art clinical and laboratory methodologies, no single candidate gene or chromosomal region has been consistently demonstrated to influence inter-individual differences in BP levels. However, none of these studies considered the combined effects on multivariate phenotypes such as systolic blood pressure (SBP), diastolic blood pressure (DBP) and another correlated trait (e.g. body mass index). Furthermore, the computational time required for a genome wide

scan using multivariate linkage analysis increases exponentially with the number of traits analyzed. Thus, it is more efficient to first exclude genomic regions that show no potential for linkage prior to conducting a multivariate linkage analysis. In this study using the Rochester Family Heart Study data we present strategies that can be used to exclude specific regions from consideration in a multivariate linkage analysis of a set of correlated traits.

31

Effect estimates from inferred haplotypes: Properties and Utility

S. Demissie, Q. Yang, J. Dupuis, L.A. Cupples
Department of Biostatistics, Boston University School of Public Health, USA

In trying to elucidate the genetic nature of quantitative traits, haplotype analysis may offer a more powerful approach than the analysis of multiple single nucleotide polymorphisms (SNPs), ignoring the dependence due to linkage disequilibrium (LD). While score tests can be used to test for association between inferred haplotypes and a quantitative trait of interest, no estimates of effect sizes are provided by such approaches. Using inferred haplotypes and standard regression analysis we evaluated parameter estimates from three approaches: 1) most probable haplotypes; 2) multiple observations per subject, one for each of H possible haplotypes, weighted by the posterior probability; and 3) single observation per subject, where each possible haplotype is represented by a variable with value equal to its posterior probability, and H-1 variables used as predictors. Parameters were estimated from a "full" model (all haplotypes evaluated simultaneously) and a "haplotype specific" model (each haplotype evaluated against all others combined). Data were simulated for N=2000 unrelated subjects, a single diallelic disease susceptibility marker and three diallelic markers in LD with the disease marker, and a range of additive effects. Preliminary results indicate that in the full models biases from multiple observations were similar to the analysis of known haplotypes. Biases from the single observation per subject were comparable for inferred haplotypes and known haplotypes. Estimates from the multiple observations and most probable haplotypes analyses had larger biases for inferred haplotypes. Multiple observations analysis appeared to underestimate standard errors. Simulations are underway to evaluate small samples and to better understand the utility of haplotype effect estimates.

32

QTL linkage and association studies of bone mineral density in selected osteoporosis pedigrees

M. Devoto(1), K. Sol-Church(1), H. Li(1), K. Wang(1), D.L. Stabley(1), G.N. Picerno(1), C. McKay(1), J. Korkko(2), D. Prockop(2), A. Tenenhouse(3), L.D. Spotila(4)
1) Nemours Children's Clinic, Wilmington DE; 2) Tulane University, New Orleans LA, USA; 3) McGill University, Montreal Canada; 4) Drexel University, Philadelphia PA, USA.

Osteoporosis is a complex trait characterized by reduced skeletal strength and increased susceptibility to fracture. The most important risk factor for osteoporosis is low bone mineral density (BMD). Several studies show that genetic factors play an important role in determining an individual's BMD. We have ascertained 42 multiplex Caucasian families through a pro-band with osteoporosis, comprising a total of 254 individuals. All available individuals have had determination of spinal and femoral neck BMD, and were typed for microsatellite markers spaced at an average 10 cM density in a whole genome scan. Linkage to BMD was analyzed by means of variance component analysis using the SOLAR package, and association was tested by means of quantitative transmission disequilibrium tests using the QTDT software. The highest lod-scores were observed between markers located in 1pter-p36.2 and femoral neck BMD (Zmax=2.85 at D1S214). QTDT analysis indicated linkage disequilibrium with two microsatellite markers, D1S489 (p=0.0012) and D1S2660 (p=0.0098), in the same region. Other regions showing support for linkage (Zmax > 1) were on chromosomes 2, 3, 6, 12, 16, 18 and 22. The positive linkage on chromosomes 2 and 6 was observed for both spine and femoral neck BMD while the other chromosomes were positive for either one or the other of the two traits.

33

Genome Screen for Allergic Rhinitis in the French EGEA study

M.H. Dizier, E. Bouzigon, M. Guilloud-Bataille, D. Torchard, C. Bétard, J. Bousquet, F. Gormand, J. Just, N. Le Moual, R. Matran, F. Neukirch, M.P. Oryszczyn, E. Paty, I. Pin, D. Vervloet, J. Feingold, F. Kauffmann, M. Lathrop, F. Demenais, I. Annesi-Maesano.
INSERM U535, Villejuif, France.

In the sample of 295 French EGEA families with at least one asthmatic subject, a genome-wide search of asthma and atopy-associated phenotypes was performed using a panel of 396 markers (Bouzigon et al, 2003). Using the same sample and set of markers, we then conducted a genome scan for Allergic Rhinitis (AR). Our aim was to search genetic factors specific to AR as well as those shared by AR and asthma. Two AR phenotypes were considered on the basis of self reporting according to a standardized questionnaire: (1) diagnosis (AR1) and (2) symptoms (AR2). Linkage analyses were conducted using the Maximum Likelihood Binomial method that can be applied to whole sibships of affected. Our sample included 67 and 82 sibships of affected with AR1 and AR2 respectively. These analyses provided indication of linkage (p value ≤ 0.001) for AR and/or asthma to three regions: 1p31–32 detected for both AR2 (p=0.008) and asthma (p=0.005) and even more significantly for the

phenotype defined by asthma plus AR2 (p=0.0004), 2q32 detected for AR2 alone (p=0.0002) and 9q22-34 detected for AR1 alone (p=0.001) but only in the subset of families with no more than one asthmatic sib (185 families). No region showed indication of linkage to asthma without being also linked to AR. These results suggest that 1p31-32 may contain a genetic factor common to asthma and AR, while 2q32 and 9q22–34 are more likely to harbor genetic factors specific to AR.

34

Power to Detect Linkage Using Covariates

B.Q. Doan(1, 2), A.J.M. Sorant(2), J.E. Bailey-Wilson(2), Y. Yao(1)
(1) Dept of Epi, Johns Hopkins School of Public Health(2) Inherited Disease Research Branch, NHGRI, NIH

The inclusion of covariates into linkage studies is believed to increase the power to detect linkage signals. Recently several genome wide linkage analyses, such as those published for prostate cancer and Alzheimer's disease, that incorporate covariates parameterized by a conditional logistic regression have been successful in increasing previously identified linkage signals as well as reducing false positive signals. Though some work has been presented that examines the power of this technique, little work has been done to study the best method of incorporating covariates, especially when many possible covariates are thought to affect the phenotype of interest. Because each additional covariate will increase the degrees of freedom, the use of many covariates could actually limit any potential gain in power. Consequently, the use of a propensity score, which is a method used in causal inference that combines multiple variables into one score, will be examined as a possible method of overcoming this problem. In this study, we simulate a model involving several covariates and a genetic locus that affect a phenotype and perform a linkage analysis on affected sibling pairs using a conditional logistic regression approach, as currently implement in LODPAL (S.A.G.E. 4.4). We will present the power to detect linkage without the inclusion of the covariate information and with the inclusion of each covariate individually, with all the covariates simultaneously and with the individual single score parameter determined by all the covariates.

35

Power and significance in genomewide association scans

F. Dudbridge (1), B.P.C. Koeleman (2)
(1) MRC Human Genome Mapping Project Resource Centre, Cambridge, UK(2) Dept Medical Genetics, University Medical Centre Utrecht, Utrecht, NL

We discuss some issues regarding statistical power and significance in forthcoming genome-wide association scans. Weak control of the family-wise error, though less important than in candidate gene studies, will still be required and imposes a limit on current approaches to identifying follow-up loci. We describe a new test, the rank truncated product, which achieves this control with higher power than current methods. We show that, conditional on this control, standard false positive and false discovery rate controlling procedures can be applied with more lenient thresholds. We examine the false discovery rate in the expected situation that the number of true associations is small. We illustrate that the false discovery proportion has high coefficient of variation and that the expected rate itself is variable when conditioned on the actual number of hypotheses rejected. For these reasons, false discovery rate procedures may be of limited usefulness to individual investigators. In contrast controlling the absolute number of false positives allows us to access most of the power of the rank truncated product, with a better-characterized error rate, and this approach may be preferable when the number of true effects is small.

36

Genetic Contribution to Intraocular Pressure: Beaver Dam Eye Study

P.Duggal(1), A.P.Klein(1), K.E.Lee(2), R.Klein(2), J.E.Bailey-Wilson(1), B.E.K.Klein(2)
(1)IDRB/NHGRI/NIH, Baltimore, MD (2)Dept of Ophthalmology, Univ.of Wisconsin Med. School

Primary open-angle glaucoma is a leading cause of blindness in the world. In the US, 3 to 6 million people have elevated intraocular pressure (IOP) and are at an increased risk for developing glaucoma. To investigate a potential genetic contribution to IOP, we performed a complex segregation analysis on 2,337 individuals in 620 extended pedigrees in the Beaver Dam Eye Study. Detailed medical histories and eye exams were performed on all participants. IOP measurement was used as a continuous trait and treatment with drops and age were covariates. Analyses were conducted in S.A.G.E using regressive class D models in REGC. The models for a single major Mendelian locus for IOP were rejected. The most parsimonious model was an environmental mixed model with the taus equal to the allele frequency. This model was not rejected (p=0.39) when compared to a totally unrestricted (general) model. A semi-general model in which the tau of the A allele for AA and BB genotypes are fixed to 1.0 and 0, and the remaining parameters are unrestricted was not rejected (p=0.13) and yielded an estimate of the tauAB genotype of 0.66, close to its Mendelian value of 0.5. Since we could not reject the environmental mixed model with equal taus it is plausible that environmental factors influence IOP. When the Mendelian dominant mixed model was compared to the semi-general model, the dominant model was marginally non-significant (p=0.056). This is consistent with a multifactorial model in which multiple genes and environmental factors may contribute to IOP.

37

Disequilibria of Types

G. Dunn
Dept. of Psychiatry, Washington Univ. USA

We develop a framework for allelic associations based on polynomial algebras in several variables. This is a good

choice conceptually in that such algebras have been studied extensively and are reasonably well understood. From a computational viewpoint, there are several programs (e.g. Maple) that perform polynomial calculations, including computing Grobner bases, which arise in association studies. Within this algebraic setting we define relative haplotypes and genotypes; in the haplotype case, this consists of a haplotype together with a list of subtypes contained in it. When the list of subtypes is empty this is just a haplotype in the usual sense. We then describe generalized disequilibrium coefficients of relative types, which in the case of an ordinary haplotype or genotype reduce to the usual disequilibrium coefficient. The use of relative disequilibria can help determine whether there are associations between some of the alleles of a type and a disjoint subtype. Our model also includes the effects of recombination for which the notion of relative type is well suited. We also give a parallel development of some related statistical concepts, which are used to give simple conceptual expressions for disequilibrium coefficients of haplotypes and genotypes. For example, we define higher order co-variances and show that a disequilibrium coefficient can be expressed entirely in terms of such co-variances.

38

Identification of polymorphisms that explain a linkage peak: Conditioning on parental genotypes

J. Dupuis (1) and P. Van Eerdewegh (2)
(1) Dept. of Biostatistics, Boston University SPH, Boston, MA; (2) Dept. of Human Genetics, Genome Therapeutics Corp. Waltham, MA

Genome scans have been performed for many complex traits and a number of chromosomal regions with a significant linkage peak have been identified. While only experimental studies can validate putative functional polymorphisms, identifying polymorphisms that best explain a linkage peak can help prioritize variants for the development of functional assays. In the context of qualitative traits and nuclear family designs, methods conditioning on the pro-band genotypes or on the sibling genotypes have been proposed to test the null hypothesis that a single nucleotide polymorphism (SNP) is the sole causative site in a region. These methods are dependent on the allele frequency of the causal variant and can be anti-conservative when the allele frequency is overestimated. In contrast, we propose three alternative test statistics, conditional on the parental genotypes, which do not rely on allele frequency estimates. These statistics exploit the fact that parents homozygote at a causative SNP should contribute little to the linkage evidence. Using a simulation study, we compare the three statistical tests conditional on the parental genotypes to previously proposed statistics conditional on the pro-band or sibling genotypes. Preliminary results indicate that while all statistics are very powerful for rejecting SNPs not in linkage disequilibrium (LD) with the functional variant, when moderate LD is present a test statistic separating the sharing from homozygote and heterozygote parents is most powerful.

39

Estimating Haplotype Frequencies in Pooled Data when there is Genotyping Error

S.R. Edwards, R.C. Elston, K.A.B. Goddard
Dept. of Epi & Biost, Case Western Reserve Univ.

In association studies, haplotype analyses are becoming important when trying to identify chromosomal regions that may contain disease genes. Due to the large number of genotypes needed, it is important to consider study designs that incorporate pooling DNA. The expectation maximization (EM) algorithm has been implemented to obtain haplotype frequency estimates from individually genotyped samples, and from pooled DNA under the assumption of no genotyping error. A more realistic scenario is that genotyping error does occur, and therefore we implement a haplotyping method for pooled DNA, based on the EM algorithm, that incorporates genotyping error into the analysis. Data were simulated both with and without genotyping error under numerous genetic models that differ in the allele frequencies and the strength of linkage disequilibrium between markers. We consider pool sizes of 1, 2, 5, and 10 individuals, and evaluate the performance of the haplotyping algorithm, both allowing for genotyping error and under the assumption of no genotyping error. Several measures of error were used to compare the estimated haplotype frequencies to the true, simulated haplotype frequencies. As expected, the performance of the haplotyping methods is worse when the data are simulated with genotyping error regardless of whether or not the genotyping error is taken into account in the analysis. The accuracy of the haplotype frequency estimates depends on the level of genotyping error allowed in the analysis. We have investigated the effects of allowing for genotyping error to estimate haplotype frequencies in pooled DNA.

40

Multipoint linkage analysis of quantitative traits on sex-chromosomes

C.T. Ekstrøm
Dept. of Mathematics and Physics, Veterinary and Agricultural Univ. Denmark

Variance component models have proved a powerful and flexible tool for multipoint linkage analysis of quantitative traits. The variance component models require estimates of genetic similarity to detect linkage and to locate genes and two methods are commonly used to estimate multipoint identity-by-descent (IBD) estimates for autosomal loci: a multiple regression approach as implemented in SOLAR or a hidden Markov model approach as implemented in Genehunter. I present an extension of the variance component model for linkage analysis of quantitative traits located on sex chromosomes and extend the two multipoint IBD estimation methods to

sex-linked loci. Simulation studies are used to assess the power and precision of the variance component model to detect quantitative trait loci (QTLs) located on the sex chromosome. The two multipoint IBD estimation methods are compared and are shown to have the same accuracy to identify QTL position but the hidden Markov model yields a larger average maximum LOD score to detect linkage than the regression model. The extension of the multipoint IBD estimation methods and the variance component model to the X chromosome demonstrate the versatility of the variance component model for linkage analysis of quantitative traits on both autosomes and sex chromosomes.

41

Inference of Haplotype Effects in Case-Control Studies Using Unphased Genotype Data

M.P. Epstein (1), G.A. Satten (2)
(1)Dept. of Human Genetics, Emory University, USA, (2)Centers for Disease Control, USA

Haplotype-based association methods are powerful procedures for identifying genes that influence complex disease. For a case-control study, a variety of statistical methods exist that detect haplotype-disease association by comparing haplotype frequencies among sampled cases and controls. As many study samples often consist of unphased genotype data (resulting in haplotype ambiguity), many of these methods apply the Expectation-Maximization (EM) algorithm for proper haplotype inference. However, the majority of such methods fail to perform inference on the effect of particular haplotypes or haplotype features on disease. As such inference is valuable, we develop a retrospective likelihood for estimating and testing the effects of specific features of SNP-based haplotypes on disease in a case-control study assuming unphased genotype data. Our proposed method has a flexible structure that allows, among other choices, modeling of multiplicative, dominant, and recessive effects of specific haplotype or haplotype features on disease. To account for ambiguous haplotype information in genotype data, we apply a variant of the EM algorithm called the Expectation-Conditional-Maximization (ECM) algorithm for inference. Our method additionally relaxes the requirement of Hardy-Weinberg Equilibrium (HWE) of the sample haplotype frequencies, which is typically required of EM-based haplotype methods. Using simulation studies our results suggest that our method returns unbiased estimates of haplotype effect size and has excellent power to detect such effects.

42

A Genome-wide Scan for NIDDM Susceptibility in the Saguenay-Lac-Saint-Jean Region of Quebec

J.M. Faith(1), M. Lemire(1), A. Verner(1), C. Darmond-Zwaig(1), J. Platko(2), J. Rioux(2), K. Morgan(1, 3), T.J. Hudson(1, 3), D. Gaudet(4), J.C. Engert(3)
(1)McGill University and Genome Quebec Innovation Centre, Montréal, Québec, Canada; (2)Whitehead Institute for Biomedical Research, MIT, Cambridge, MA; (3)Departments of Human Genetics and Medicine, McGill University, Montréal, Québec, Canada; (4)Dyslipidemia, Diabetes and Atherosclerosis Group and Community Genomics Research Center, Université de Montréal and Complexe hospitalier de la Sagamie, Chicoutimi, Québec, Canada.

One of the leading health care costs in the western world is Non-Insulin Dependent Diabetes Mellitus (NIDDM or Type II Diabetes). This disease has long been known to exhibit familial aggregation but the vast majority of NIDDM is thought to have a multifactoral origin. There have been several genome scans performed to date to identify loci that confer susceptibility to NIDDM. However, there exist currently undefined contributors to the genetic risk for this disease. The French Canadian population of the Saguenay-Lac Saint-Jean region (SLSJ) of Quebec, Canada is a 300–400 year old founder population and may provide increased power to detect genes contributing to complex traits. In order to identify loci that contribute to NIDDM susceptibility, we performed a whole genome scan on 240 individuals from 42 NIDDM families (an average of 5.74 sibs/ family) from the SLSJ. By using 382 evenly spaced polymorphic microsatellite markers, an approximately 9.2 cM resolution was achieved. Markers had an average heterozygosity of 74%. Multipoint nonparametric linkage analysis using the GENEHUNTER and GENEHUNTER-PLUS software packages identified one chromosomal region with a nonparametric linkage (NPL) score greater than 3.0. The highest single point score, 1.75, was located at D14S611, and the maximum multipoint score, 3.24, was located at the same marker. We are currently fine mapping this region.

43

Pedigree Joint Linkage Disequilibrium and Linkage Mapping of Quantitative Trait Loci

R Fan and C Spinka
Department of Statistics, Texas A&M University, USA

To map quantitative trait loci (QTL) for complex diseases, linkage disequilibrium (LD or association) regression analysis using population data can be performed. However, population substructure may affect the results and produce false positives. To remedy this, variance component models are proposed to perform joint LD and linkage mapping of QTL using both population and pedigree data. Previously, either nuclear families or sib-ships were used to construct variance component models. In this paper, we propose to use large pedigrees in joint LD and linkage mapping of QTL by variance component models, which simultaneously model both linkage and LD parameters. The LD information and covariate are modeled by the mean, and linkage information is modeled by the variance-covariance matrix. The association is decomposed into additive and dominant components. Analytical formulas are provided to calculate the regression coefficients and non-centrality parameters of test statistics of LD in the presence of linkage. Power, sample sizes and type I error rates are explored by both the

analytical formulas and simulations. Comparison with Fulker and Abecasis et al. ''AbAw'' approach is investigated by both simulation and theoretical analysis.

44

Are there better quantitative traits in the search for obesity-related genes than BMI?

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

Obesity (Ob) is a complex disorder with multiple genes likely to be involved. Several genome scans of Body Mass Index (BMI) have been performed but linkage results are inconsistent across studies. Little work has been done to identify better Ob-related quantitative traits (QT) prior linkage analysis. Physiological and statistical relationships between Ob and hypertension (HT) are well established, and genetic correlation has been suggested. In this study, we used 97 HT families (n=1180) from a French-Canadian relative isolate with the purpose of identifying the most promising QT for genome scans of Ob-related traits in HT. More than 600 phenotypes have been collected. Since body water (BW) regulation has been suggested to be involved in Ob-related HT, 2 BW measures in addition to 5 anthropometric measures have been analyzed in n=242–792 subjects. Using oligogenic segregation analysis based on Bayesian Markov chain Monte Carlo methods, our analyses suggest at least 1 locus with an individual contribution $>10\%$ to each QT variance (V). Contributions of the largest loci to V of anthropometric QT, including BMI, are ~41–64% and ~11–30% for BW QT. Among the QT studied, subscapular skin fold is the best candidate because it has the highest heritability (~83%) and a small number of loci implicated (mean 3.9), with the 3 largest loci contributing ~55, 20 and 7%, respectively, to V. A sex effect of 21–42% on V was estimated for BW QT, whereas it was only ~2% for most anthropometric QT. In this data set, BMI is a good QT but perhaps, not the best one. Careful analyses of QT related to weight and BW regulation prior linkage analyses, as well as covariate adjustments, will be useful in mapping genes implicated in Ob-related HT. Analysis of more Ob-related QT is under way.

45

Comparison of quantitative and qualitative linkage analysis by re-analysis of an affected sib pair genome scan for obesity

F. Geller (1)+, A. Dempfle (1)+, T. Görg (1), J. Hebebrand (2)
1 - Institute of Medical Biometry and Epidemiology, Philipps-University, Marburg, Germany, 2 - Clinical Research Group, Philipps-University, Marburg, Germany, +- contributed equally

Linkage studies of quantitative traits can be analyzed with variance components methods (e.g. Merlin VC) or regression based methods (e.g. Merlin Regress). Alternatively, the phenotype can be dichotomized based on a threshold and analyzed as a qualitative trait (e.g. MLBGH). In order to compare these 3 methods, we re-analyzed a subset of 70 families from a genome scan with obese sib pairs. The aim of our study was to investigate whether Merlin Regress is able to identify regions of high IBD-sharing in affected sib pairs based on the BMI values. The results from MLBGH and Merlin Regress were close to the ones from the original analysis. The nine formerly reported regions with LOD scores greater 0.7 displayed LOD scores of comparable magnitude in the new analyses. The LOD scores from Merlin VC in the nine regions surpassed 0.7 only on chromosome 8, stressing the power problems of variance components methods in selected samples. Neither Merlin Regress nor Merlin VC showed additional—potentially false positive—peaks. In conclusion, we were able to show that the regression-based quantitative linkage analysis with Merlin Regress is able to identify regions of increased IBD-sharing of affected sib pairs even in this highly selected sample, whereas the variance components method failed to reproduce the results from the qualitative analysis.

46

Quantification of the bias in the odds-ratio of statistically-inferred haplotypes in case-control data

E. Génin, P. Margaritte-Jeannin
INSERM U535, Hopital Paul Brousse, Villejuif, France

To test for the association of a candidate gene with a complex disease, haplotypic methods may be more powerful than single-locus methods. Indeed, in a given genomic region, disease susceptibility may be due to the combined effects of multiple sequence variants and studying these variants together in haplotypes could then be a better strategy than looking at each of them individually. However, most molecular techniques only provide single-locus genotypes and haplotypes are thus not directly available. Statistical methods have been developed to estimate haplotype frequencies from genotypic data. These estimations are obtained under the hypothesis of random mating in the population and require Hardy-Weinberg proportions (HWP) in the sample. In a genomic region known to harbor disease susceptibility loci, deviations from HWP may be observed in the patient sample and this could lead to bias in haplotype frequency estimation. In this study, we have performed simulations under different models of disease susceptibility and quantified the bias on the odds-ratio estimate due to the fact that haplotypes are not known but statistically inferred. We found important bias, especially when linkage disequilibrium between loci is small and rare haplotypes are present. The disease susceptibility model only seems to have a marginal effect. An illustration on Multiple Sclerosis family data and DRB1-TAP2 haplotypes is provided where the OR for a given haplotype is increased from 3.9 (95%CI[2.7, 5.8]) when haplotypes are known to 6.5 (95%CI[4.0,10.4]) when haplotypes are statistically-inferred.

47

Genetic Dissection Of A Multivariate Phenotype Using Reverse Regression

S. Ghosh(1), P.P. Majumder(1), T. Reich(2)
(1) Indian Statistical Institute, Kolkata, India, (2)Washington University School of Medicine, St. Louis, USA

A complex trait is usually a function of a multivariate phenotype comprising correlated quantitative variables. 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, we find that the proposed method is robust to violations in distributional assumptions like normality. We show that there is a 8–15% gain in power using this method on the multivariate phenotype, when compared to Haseman-Elston regression or the reverse regression procedures based on the first principal component. 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.

48

Anticipation in families with lymphoproliferative tumors: an artifact of ascertainment?

L.R. Goldin(1), R.M. Pfeiffer(2)
(1)Genetic Epidemiology Branch, DCEG/NCI, Bethesda, MD, (2)Biostatistics Branch, DCEG/NCI, Bethesda, MD.

Families with Hodgkin lymphoma (HL), non-Hodgkin lymphoma (NHL) and chronic lymphocytic leukemia (CLL) typically show anticipation where the offspring have an earlier age at diagnosis than the parents. However, ascertainment and truncation of the age at observation of offspring can bias the assessment of anticipation in family data. We have tested for anticipation in HL, NHL, and CLL using data on families ascertained from population-based samples in Sweden and Denmark Large family cancer databases were created by linking population registries (containing parent-offspring links) to cancer registries. We analyzed first-degree relatives of more than 7000 HL cases, 5900 CLL cases and 25,000 NHL cases. Even with large sample sizes these tumors are uncommon and the numbers of parent-offspring pairs is small, especially for HL and CLL. Using life table analysis, we tested for differences between age of diagnosis in parents and offspring. The mean age of diagnosis in the offspring was substantially lower than that in the parents for all three tumors. In the case of CLL, in 18/20 parent-offspring pairs, the age of diagnosis of the offspring was lower than the parent. However, for both CLL and HL, there were no differences between the survival curves of parents and offspring. For NHL, there was a significant parent-offspring difference that could be explained by the strong secular increase in NHL incidence rates. We conclude that observed anticipation in families with lymphoproliferative tumors is most likely a result of ascertainment bias.

49

Prospective risk of cancer in CDKN2A germline mutation carriers

AM Goldstein(1), JP Struewing(2), MC Fraser(1), MW Smith(3), MA Tucker(1)
(1)Genet Epidemiol Branch, DCEG; (2)Lab of Pop'n Genet, CCR; (3)Lab of Genomic Diversity, CCR and BRP, SAIC Frederick; National Cancer Inst, NIH, DHHS, USA

The CDKN2A gene is the major known high-risk melanoma susceptibility gene. Germline mutations have been detected in 20 percent of melanoma-prone families. Susceptibility to other cancers has also been suggested. In particular, there is a significantly increased risk of pancreatic cancer in a subset of families with CDKN2A mutations. However, most studies examining risks of other cancers classified individuals according to the family's CDKN2A mutation rather than determining individual mutation status. Risks could also be biased because of cancer occurrence prior to family ascertainment. We examined the risk of nonmelanoma cancer in CDKN2A melanoma-prone families restricting the analysis to the period after ascertainment and using individual mutation data. Analyzing 117 mutation-positive (of whom 64 had melanoma) and 136 mutation-negative subjects yielded no significant cancer associations for mutation-negative subjects. In contrast, mutation-positive subjects had a significantly increased risk for all cancers combined [SIR=2.17, 95% CI 1.12-3.80] primarily because of pancreatic cancer. No other tumors showed significantly increased risks. Differences in CDKN2A-nonmelanoma cancer associations across studies may result from variation in genetic backgrounds, insufficient follow-up, misclassification of mutation carriers, or influence of non-CDKN2A relatives. Larger sample sizes, prospective follow-up, and individual mutation data will be required to understand these differences.

50

Evidence for Interaction Between Sulfotransferase Genotypes and Haplotypes with Cigarette Smoking in a Colorectal Polyp Case-Control Study

E.L. Goode(1,2), J.D. Potter(1,2), and J. Bigler(1)
(1)Fred Hutchinson Cancer Research Center, Seattle WA, USA; (2)University of Washington, Seattle WA, USA

Epidemiologic evidence suggests that hyperplastic and adenomatous polyps lead to colorectal cancer via separate

pathways, and that tobacco-related carcinogens are risk factors for hyperplastic but not adenomatous polyps. Because sulfotransferases are involved in elimination of tobacco carcinogens, we hypothesized that polymorphisms in SULT1A1 or SULT1A2 may interact with cigarette use in development of hyperplastic polyps. We examined the role of these genes in a clinic-based study of 298 cases with hyperplastic polyps (with or without additional adenomatous polyps), 382 cases with adenomatous polyps only, and 592 polyp-free controls. Overall, genotype at SULT1A1 R213H was not significantly associated with risk of either type of polyp. However, among current smokers (n=185), individuals with RH genotype had a 2.3-fold increased risk of hyperplastic polyps (95% CI: 1.1–4.6) and those with HH genotype had a 2.9-fold increased risk (1.2–7.2). Among never- or former-smokers (n=705), no increase in risk was seen with RH or HH genotype (OR 1.3 (0.9–1.8) and 0.9 (0.5–1.5)). Similar results were seen for SULT1A2 N235T. Haplotype analysis based on SULT1A1 R213H, SULT1A2 P19L, and SULT1A2 N235T (mean $|D'|=0.97$) also showed this interaction and suggested that chromosomes with H at SULT1A1 R213H may be driving the results. No difference in adenomatous polyp risk with SULT1A1 or SULT1A2 was seen by smoking status, consistent with differing etiologies for the two polyp types.

51

Statistical Identification of Causal Genetic Variants in Quantitative Trait Loci

HHH Göring, JT Williams, L Almasy, J Blangero
Dept. of Genetics, Southwest Foundation for Biomedical Research, San Antonio, TX, USA

The ultimate goals of a gene mapping study are to identify the gene(s) influencing a trait of interest and to determine the allele(s) with functional effect. While genome-wide linkage analysis in pedigrees has proven to be an invaluable tool for initial gene localization, identification of the causal gene(s) and functional variant(s) within these positional candidate regions remains a daunting task for all but the simplest ''Mendelian'' traits. To overcome this impasse, we have developed a statistical approach designed to answer the questions, Is this gene involved in the etiology of the trait?, and "What variants within the gene are functional with respect to the phenotype(s) of interest?". Our approach, Bayesian Quantitative Trait Nucleotide analysis, relies on sequence data from a candidate gene, ideally including its regulatory regions, in a pool of individuals, such that all genetic variants can be identified probabilistically. The basic idea is to compute the likelihood of the observed phenotypic data (which may be quantitative or qualitative) for all possible combinations of SNPs and to use Bayesian model averaging and selection to determine whether the gene is involved in trait etiology and to identify the functional polymorphisms. For simplicity, we have so far focused on additivity of alleles within and between SNPs, neglecting potential interactions between the alleles of a polymorphism (dominance) and between polymorphisms (epistasis). We have used simulation, based on actual haplotype data, to assess the behavior of the approach under the null and alternative hypotheses with respect to involvement of the gene as a whole and the individual variants within the gene. Overall, the approach appears to be a very promising tool for determining whether a gene is involved in the etiology of a complex trait, and which variants within the gene are functional. The estimated posterior probabilities of functionality may be used to prioritize the polymorphic sites for further laboratory investigation.

52

The Heritability of Schizophrenia as a Function of Age in the Presence of Strong Cohort Effects

C.M.T. Greenwood(1), J.A. Husted(2), A.S. Bassett(3)
(1) Genetics and Genomic Biology, Hospital for Sick Children, Toronto, Canada, (2) Dept. of Health Sciences and Gerontology, University of Waterloo, Canada, (3) Dept. of Psychiatry, University of Toronto

Many mental illnesses show heritability and appear to cluster in families. In the mid 1940's, Dr. Lionel Penrose collected information from all mental institutions in Ontario on relatives hospitalized for mental illness (or suicide) between 1880 and 1945. We were interested in investigating, in this extremely large data set whether the heritability of schizophrenia varied with age of onset. No information was recorded on unaffected individuals in these families. We examined 1,650 sibships with at least two hospitalized siblings, and used extensions of the Cox proportional hazards model to examine the hazards of hospitalization with schizophrenia affective disorders, and other conditions as a function of age. These four diagnostic groupings were treated as competing risks. Year of birth was included in all models to adjust for the very strong time trends in hospitalization rates. Correlation within sibships was accounted for either by using a GEE correction to variances, or by modeling each sib conditionally on all older sibs (the regressive models). The heritability of age of onset was investigated by asking whether disease risk was influenced by onset age in other siblings. For both analysis methods, we identified a trend towards increased hazard of schizophrenia hospitalization when siblings had either very young (<20) or very old (>50) ages of onset (p=0.045). The results suggest increased heritability of schizophrenia with younger age at onset.

53

EM Algorithm Based Genetic Tests of Association and/or Linkage with Missing Parental Data

C. Guo (1), K. Lunetta (2), A. destefano (1), J. Dupuis (1), L. A. Cupples (1)
1. Department of Biostatistics, Boston University, USA
2. Genome Therapeutics, MA, USA

In family-based association tests, both diseased cases and their parents' genotypes are collected for testing linkage disequilibrium (LD). However, the genotypes of both parents are not always available. Under such scenarios, Sun et al (*1999*, Am J. Epi) proposed the 1-TDT to detect linkage between a candidate locus and a disease locus using genotypes of the affected individuals and the one available parent, as well as a statistic, which combined

these subsets with affected individuals with data from both parents. In this paper, we apply the Expectation-Maximization algorithm to deal with one parent only data, a method that can provide unbiased information about transmitted and non-transmitted markers. We then combine these one-parent families with two-parent families to formulate the EM-TDT & EM-HRR statistics, and additionally subjects with no parental data for the EM-HRR. According to the simulations, the EM-TDT and 1-TDT perform similarly in detecting LD. Due to the haplotype relative risk's data structure, the transmitted markers are always present regardless of missing one or two parents. As a result of having larger sample size, the EM-HRR is uniformly more powerful in detecting linkage disequilibrium than both EM-TDT and 1-TDT, when the population is under HWE. When the admixture effect is not severe or strong association with severe admixture, EM-HRR is still the most powerful test.

54

The effects of sex-specific recombination on tests of imprinting using affected sibling pairs

P. Holmans
MRC Biostatistics Unit, Cambridge, UK

Imprinting (parent-of-origin effect) occurs when the expression of a gene is dependent on the parent from which it was inherited. In an affected sib-pair study, imprinting will be manifested as an increase in the probability of an affected pair sharing either their paternal or maternal allele identical by descent (IBD). This can be tested by maximizing the likelihood of the marker data with respect to paternal and maternal IBD probabilities separately, and comparing to that obtained when they are assumed to be equal. Differences in paternal and maternal IBD probabilities can also be caused by sex-specific recombination fractions between the disease locus and the marker(s). For a single marker, the effect of this on the Type I error rate of a likelihood-ratio test of imprinting depends on the size of genetic effect, distance between disease and marker loci, and sample size. Region-wide Type I error rates for a number of multipoint test statistics for imprinting were assessed by simulation. Use of the correct statistic gave approximately correct Type I error rates, except when the effect of the disease locus and the ratio of sex-specific genetic distances were both large, and the marker grid was wide (20cM). Powers of the test statistics to detect imprinting were investigated under a variety of disease models.

55

Using Random Forests for the Combined Analysis of Microarray and Genetic Marker Data

S. Horvath (1, 2), P. Kraft (2)
(1) Depts of Human Genetics and Biostatistics, University of California, LA (2) Depts of Epid. and Biostatistics, Harvard School of Public Health, Boston, MA, USA

Microarray data generated by functional genomics, gene expression, DNA methylation, and proteomics experiments are increasingly used as covariates in population based allelic association studies. A rationale of such studies is that some microarray covariates may confound the relationship between the genotypes and the affectation status. Since there may be thousands of microarray covariates but only dozens of observations many classical statistical methods such as multivariable logistic regression are inadequate. Here we propose a method for summarizing the microarray data into a single scalar M that can be used as covariate in a logistic regression model. The method is based on random forest predictors introduced by L. Breiman, which is a state-of-the-art supervised learning method that is well-suited for data with many covariates but relatively few observations. Specifically we will use random forest predictors to arrive at an out-of-bag estimates of the probability of being affected M. We explain how to determine the significance level of M when it is incorporated in a logistic regression model and argue that one should compute 1-sided p-values for the corresponding Wald test statistic. This procedure is similar in spirit to the pre-validation procedure studied by Tibshirani and Efron but we show with simulations that it avoids a ''leakage'' of degrees of freedom. To understand how M depends on important covariates we use regression trees and partial dependence plots.

56

A score statistic to test for familial aggregation in a proband-family design

J.J. Houwing-Duistermaat(1), R. el Galta(1), J.C. van Houwelingen(2), CM van Duijn(1)
(1)Dept of Epi & Biostat, Erasmus MC, Rotterdam, The Netherlands (2)Dept of Med Stat, LUMC, Leiden, The Netherlands

Before genome scans are performed in a sample of families, it is important to know if the trait aggregates within families. To test for clustering, a robust score statistic may be used which allows for adjustments of covariates, for example candidate genes. However, often the families are selected via affected probands in the case of binary traits and via probands with extreme values in the case of continuous traits. For such a proband-family design, we derived the score statistic by using the conditional log-likelihood given the outcomes of the probands. To adjust for covariates, we proposed to use estimates of parameters obtained from large-scale epidemiological studies, because valid estimates of these parameters cannot be obtained from the selected families. Alternatively, a weighted sum of estimates from a set of studies may be used. The statistic appeared to be the sum of a linear term, which measures the similarity between probands and relatives and a quadratic term, which measures the similarity among the relatives. We used a scaled chi-square distribution to approximate the distribution of the statistic under the null hypothesis. This distribution appeared to perform well. As illustration the score statistic was applied to a sample of 40 families with diabetes type II varying in size from 3 to 12 subjects and selected via 42 probands. Age and sex specific

prevalence of diabetes were obtained from the Rotterdam Study (n=8000). Clustering of diabetes was highly significant.

57

Modeling fetal viability loss in the Maternal-Fetal Genotype Incompatibility (MFG) test

H Hsieh(1), CGS Palmer(2, 3), EF Reed(4), JA Woodward(3, 5), J Lönnqvist(9), L Peltonen(6, 8), JS Sinsheimer(1, 6, 7)
(1) Biostat (2) Psychiat (3) Stat (4) Path (5) Psychol (6) Hum Gen (7) Biomath, UCLA; and (8) Mol Med (9) Ment Hlth & Alc Res, Nat Pub Hlth Inst, Helsinki

Genotype incompatibility between a mother and fetus can create adverse prenatal conditions, either because maternal-fetal genotypes differ (mismatch) or are the same (match). Both scenarios potentially increase disease susceptibility. However, maternal-fetal genotype incompatibility may also affect ascertainment through fetal viability loss that is directly related to the locus under investigation. Sinsheimer et al (Gen Epi, 24: 1–13, 2003) proposed the MFG test using case-parent trios to examine incompatibility. Unfortunately, the incompatibility effect estimated from their method cannot distinguish between a fetal viability effect and an incompatibility effect and so can result in an inaccurate incompatibility estimate. We developed a new method that extends the MFG test using the multiple sibling version of the conditional likelihood model of Kraft et al.(submitted). Our method models both incompatibility and fetal viability in the joint distribution of the children and parents' genotype conditional on the children's affection status. Through simulation, we show that moderate sample sizes can provide significant power to detect fetal viability loss and allow accurate estimation of incompatibility. An empirical data analysis of maternal-fetal genotype incompatibility at HLA loci for risk of schizophrenia in Finnish family material, incorporating the fetal viability parameter, will also be discussed.

58

Semiparametric estimation of marginal hazard function from case-control family studies

L. Hsu(1), L. Chen(2), M. Gorfine(3), K. Malone(1)
(1)Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, USA, (2)Department of Biostatistics, University of Washington, USA, (3)Bar Ilam University, Israel

Estimating marginal hazard function from the correlated failure time data arising from case-control family studies is complicated by non-cohort study design and risk heterogeneity due to unmeasured, shared risk factors among the family members. Accounting for both factors in this paper, we propose a two-stage estimation procedure. At the first stage, we estimate the dependence parameter in the distribution for the risk heterogeneity without obtaining the marginal distribution first or simultaneously. Assuming that the dependence parameter is known, at the second stage we estimate the marginal hazard function by iterating between estimation of the risk heterogeneity (frailty) for each family and maximization of the partial likelihood function with an offset to account for the risk heterogeneity. The simulation study shows that the proposed method performs well under finite sample sizes. We illustrate the method with a case-control family study of early onset breast cancer.

59

Genome-wide Linkage of Familial Myopia

G.P. Ibay(1), B. Doan(1), L. Reider(2), D. Dana(2), M. Schlifka(2), H. Hu(1), T. Holmes(1), J. O'Neill(1), J.E. Bailey-Wilson(1), D. Stambolian(2)
(1) Inherited Disease Research Branch, NHGRI/NIH, (2) University of Pennsylvania, USA.

Purpose. To identify regions of the human genome containing genes responsible for nonsyndromic myopia using pedigrees from four ethnic groups of the Myopia Family Study—Ashkenazi Jewish, Amish, African- and Chinese-American families. Methods. Cycloplegic and manifest refraction were performed on 46 Jewish and 66 Amish families. Individuals with –1.00 D in each meridian of both eyes were classified as myopic. A genome-wide scan using 390 microsatellite markers was recently completed at the Center for Inherited Disease Research (CIDR). Parametric and nonparametric linkage analyses are being conducted to determine what loci are important in families with less severe, clinical forms of myopia. Results. The results of a preliminary study did not indicate any strong evidence of linkage of myopia in a subset of families to the candidate regions on chromosomes 12 and 18. A GWS will allow us to determine if other loci may play a role in myopia susceptibility in the total sample of families. Results for the GWS will be presented.

60

The "diplotypic test": a powerful strategy to detect association

A-S Jannot (1, 2), L. Essioux (2), F. Clerget-Darpoux (1)
(1) Unité INSERM 535, Villejuif, France, (2) ValiGen, La Défense, France

When performing an association test to detect the role of a candidate gene using intragenic SNPs, one can either use a haplotypic test or a diplotypic test (considering the phased pair of haplotypes). Until now, the power of both tests has only been compared for simple situations using a multi-allelic TDT test versus a multi-genotypic TDT test, showing the advantage of haplotypic test when there are more than two alleles. One hypothesis when many SNPs are available is that the number of haplotypes and genotypes is large compared to the sample size. Many genotypic and haplotypic categories have few individuals, each of these categories adding one degree of freedom while providing little information and this has never been taken into account. To circumvent this issue, we propose and compare three grouping strategies to apply before performing an association test. These strategies are based i) on a single baseline group pooling all rare categories, ii) a measure of dissimilarity or iii) on SNP selection. We compare haplotypic and diplotypic tests using these

strategies for different situations and more particularly, we consider the case of haplotypic interaction. We show that all grouping strategies improve the power of both tests and that one of them, the one based on a measure of dissimilarity, increase the most the power. We also show that for strong haplotypic interaction diplotypic tests are powerful while haplotypic tests have no power and that the loss of power for diplotypic tests versus haplotypic tests is not strong in case of no haplotypic interaction.

61

Estimation of Heritability Attributable to Single-locus Effects with a Regression of Offspring on Mid-parent Method for Cardiovascular Risk Factors

SH Jee(1, 2), M-H Roy-Gagnon(3, 4), YS Jang(2), TH Beaty(4), AF Wilson(3)
(1)Graduate School of Public Health, Yonsei University, (2)Cardiovascular Genome Center, Yonsei University Medical Center, Seoul, Korea, (3)Inherited Disease Research Branch, NHGRI, NIH, Baltimore, MD (4)Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

The objective of this study was to estimate the heritability attributable to single-locus effects with a regression of offspring on mid-parent (ROMP) method for cardiovascular risk factors. In this ROMP method, regression coefficients of offspring phenotype on the mid-parent value is estimated with and without including a single marker effect, and the difference between these two regression coefficients provides an estimate of the heritability attributable to the marker locus. The study population used here included 1, 550 family members of 295 patients. Estimated heritability was 35 to 46% ($p<0.0001$) for total cholesterol with 6.2% ($p=0.0088$) attributable to the S128R polymorphism. For triglyceride, the estimated heritability was 47.6% ($p<0.0001$) with 2% ($p=0.0182$) attributable to the G-217A polymorphism. The heritability was 36–46% for LDL-cholesterol. For LDL cholesterol the S128R marker effect was 8.7% ($p=0.0049$). Estimated heritability was 62.2% ($p<0.0001$) for apoAI with 3.2% ($p=0.0338$) attributable to the polymorphism G-217A and 58 to 75% ($p<0.0001$) for apoB with 4.8% attributable to polymorphism S128R. These results highlight the importance of considering multiple genetic factors in studies of cardiovascular risk factors.

62

Case-control-family designs and polymorphisms: application to breast cancer risk

M.A. Jenkins(1), R.L. Milne(1), M.C. Southey(1), A.B. Spurdle(2), D. Nyholt(2), J-H. Chang(1), G. Chenevix-Trench(2), J.L. Hopper(1)
1) Centre for Genetic Epidemiology, The University of Melbourne, Australia; 2) Queensland Institute of Medical Research, Brisbane, Australia

Common polymorphisms may be non-spuriously associated with disease, because the variant is functional or in linkage disequilibrium with a causative variant. Associations have been estimated using individuals (e.g. case-control studies), or family-based approaches limited to case-relative control or case-parent triads (popular due to minimal demands on sampling relatives). The population-based case-control-family design obtains biological material and information on risk factors, including personal and family cancer history, from cases, controls and relatives, and can be used for the above approaches, allowing results from between-family and within-family analyses to be compared. New methods can assess associations even if there are missing genotypes. We have conducted a case-control-family study of breast cancer in which polymorphisms in candidate genes have been measured on 1300 cases, 1800 of their relatives and 700 controls. We conducted case-control analyses and TDTs. We also conducted within-case-family analyses using a modified segregation analysis (using MENDEL) that modeled background familial effects and estimated variant effects in terms of hazard ratios (comparable to odds ratios from case-control analyses) using population incidence. Population controls provide estimates of allele frequency and risk factor prevalence. We present examples for polymorphisms in estrogen metabolism and DNA repair genes.

63

Whole Genome Scan in a Complex Disease using 11245 SNPs

S John (1), N. Shephard (1), M. Cao (3), J. Che (3), N. Vasavda (2), G. Liu (3), T, N. Gibson (2), K. Jones (3), J. Worthington (1), G. Kennedy (3)
(1)University of Manchester, Manchester, UK; (2) Astrazeneca, Alderley Park, UK(3) Affymetrix, Inc. Santa Clara, CA, USA.

Genome scans comprising ~400 microsatellite markers have been applied successfully to linkage studies; however genotyping micro-satellites is resource intensive. We undertook a paradigm study to determine whether high density SNP markers could be used successfully in a genome scan for a complex disease. We used a recently described array-based genotyping technology called whole genome sampling analysis (WGSA), which uses one generic primer to amplify >10,000 SNPs in a single reaction and makes automated genotype calls at >99% accuracy. We analyzed 655 individuals from the UK cohort of rheumatoid arthritis families (252 affected sibling pairs) for 11, 245 genome-wide SNPs and obtained >6.6 million genotype calls. The median spacing of the SNPs is 115 kb and the median heterozygosity is 0.40. Non-parametric multi-point linkage analysis was performed using Merlin. In addition to confirming linkage to HLA, we obtained evidence for linkage to six other regions, two were detected in a microsatellite whole genome scans and four were not. Three loci detected in the microsatellite genome scans were not detected in the current SNP study. Several factors could account for these results; differences in information content and marker density between micro-satellites and SNPs, and possible linkage disequilibrium between SNPs. We calculated multipoint polymorphic information content and found it to be uniformly high (85–97%) across the genome.

64

Gene-environment interaction estimation methods and strategies: a review

C. A. Peterson Jones (1, 2), C. Aragaki (1)
(1) Department of Epidemiology and Biostatistics, Dallas Regional Campus, School of Public Health, University of Texas Health Science Center at Houston, USA (2) Department of Epidemiology, School of Public Health, University of North Texas, USA

To assess the current state of applied human gene-environment interaction (GEI) estimation, all articles with GEI were obtained from 1992–2002. 935 articles were identified as entered into PubMed as of April 2003 and 934 (99.9%) were obtained and assessed. 787 (84.2%) out of 934 did not present any results. These articles were not about human subjects or human cell lines (42.5%), discussed GEI without results (22.7%), contained new statistical methods for GEI estimation but did not present results (13.1%), presented results for a genetic effect but no specified environmental risk factor (15.2%), or presented results for an environmental effect but specified no genetic factor (5.1%). Of the 147 articles that presented data, we found that 74 articles did not analyze the data (50.3%), instead presented frequency counts or means for patients within each genetic and environmental risk factor combination. Using logistic regression to predict gene-environment interaction excess odds ratios for 2 genetic and 2 environmental categories was the most common presentation method. This indicates that even though there is a large variety of analytic techniques available to investigators to estimate gene-environment interactions, there needs to be either more application of developed GEI estimation methods or development of GEI estimation methods that are easily accessible and interpretable.

65

Segregation Analysis of Nuclear Sclerosis Incorporating Smoking as a Covariate

A.P. Klein(1), K. E. Lee(2), J.A. O'Neill (1), R. Klein (2), J.E. Bailey-Wilson(1), B.E.K. Klein(2)
(1) NIH/NHGRI/IDRB, Statistical Genetics Section, Baltimore, MD; (2) Department of Ophthalmology and Visual Sciences, University of Wisconsin-Madison Medical School, Madison WI.

Cataract is the most common cause of blindness worldwide. Nuclear sclerosis is the degree of progressive opalescence of the human lens. Severe nuclear sclerosis is considered to be nuclear cataract, the most common form of age-related cataracts. Cigarette smoking has been shown to be associated with more severe nuclear sclerosis. We assessed the familial aggregation of nuclear sclerosis within data collected as part of the Beaver Dam Eye Study. Previous segregation analysis of nuclear sclerosis with these data not controlling for cigarette smoking provided evidence for the inheritance of a major gene accounting for 35% of the variation in adjusted nuclear sclerosis. Segregation analysis was performed using REGC and REGCHUNT. Analyses were based upon 2,089 individuals in 620 extended pedigrees with complete data on age, sex, pack years cigarette smoking exposure and degree of nuclear sclerosis. Cigarette smoking was shown to be a significant covariate in the segregation analysis. However, unlike our previous analysis, which did not take smoking into account we did not find clear evidence supporting the involvement of a single major gene influencing severity of nuclear sclerosis. All mendelian and environmental models were rejected when compared to a the general transmission model but the tau-AB free model was not rejected. These results highlight the complex etiology of nuclear sclerosis.

66

Heritability of dyslexia and related phenotypes—first results from the German bi-center study

I. R. König(1), W. Deimel(2), E. Plume(3), J. Bartling(2), A. Warnke(3), H. Remschmidt(2), A. Ziegler(1), G. Schulte-Körne(2)
(1) Institute of Medical Biometry and Statistics, University Lübeck, Germany(2)Department of Child and Adolescent Psychiatry, University Marburg, Germany(3)Department of Child and Adolescent Psychiatry and Psychotherapy, University Würzburg, Germany

Dyslexia is a specific disorder in learning to read and spell in spite of adequate education, normal intelligence, no sensory deficits, and adequate sociocultural opportunity. Dyslexia is known to be a hereditary disorder affecting about 5% of school-aged children, making it the most common of childhood learning disorders (Schulte-Körne, 2001, J Child Psychol Psychiatry, 42, 985–997). Etiological research has focused on basic perceptual deficits yielding conflicting results. It has been demonstrated that dyslexics are handicapped in processing of speech and visual stimuli. The heritability of each separate component has remained largely unclear. We perform a German bi-center study to analyze the genetic background of dyslexia and its underlying processing disorders as related phenotypes. Employing a single proband sib pair design as recommended (Ziegler, 1999, European Child & Adolescent Psychiatry, 8, 35–9), we include 250 dyslectic children with at least one sibling and its parents. Recruitment began in 8/2001 and will be finished in 8/2003. As first results, we present the relationship between related phenotypes. In addition relative risk ratios for information processing disorders will be determined. This will form the base for comprehensive linkage analyses of all involved phenotypes.

67

Racial Differences in the Familial Aggregation of Breast and Other Female Cancers in the Women's Contraceptive and Reproductive Experiences (CARE) Study

J.F. Korczak, M.S. Simon, C.L. Yee, A.G. Schwartz
Department of Internal Medicine, Wayne State University and Population Studies and Prevention Program, Karmanos Cancer Institute, Detroit, MI, USA

The familial aggregation of breast and other cancers has been well documented in Caucasians, but less data are available from studies involving African Americans. We

assessed the familial aggregation of breast, ovarian, uterine, and cervical cancers in the first-degree relatives of 5, 863 Caucasian and 3,177 African American participants in the National Institute of Child Health and Human Development Women's CARE Study. The cancer status of each member of the cohort of relatives comprised the outcome in unconditional logistic regression, adjusting for correlated data with generalized estimating equations and defining a family history indicator according to whether an individual was related to a case or a control. Race was a significant factor for each cancer site. For breast cancer, the estimated relative risk and 95% confidence interval (RR; 95% CI) were somewhat higher in Caucasians (2.00; 1.72–2.32) than African Americans (1.76; 1.40–2.22). Racial differences were also obtained for ovarian and cervical cancer. In Caucasians, relatives were at increased risk of ovarian cancer (1.56; 1.06–2.29) but not cervical cancer (0.76; 0.53–1.09), whereas in African Americans, relatives were more likely to be diagnosed with cervical cancer (1.81; 1.05–3.10) but not ovarian cancer (0.84; 0.50–1.40). For both races, relatives were not at increased risk of uterine cancer. The differences seen in cancer aggregation patterns between Caucasians and African Americans reflect potentially different genetic effects, environmental exposures and/or cultural factors that may impact cancer risk.

68

Clustering family-based gene expression data

P. Kraft(1), S. Horvath(2)
(1) Depts. of Epidemiology and Biostatistics, Harvard School of Public Health, (2) Depts. of Human Genetics and Biostatistics, University of California, Los Angeles

Comprehensive gene-expression data from related individuals can be used to dissect complex disease genetics. For example, gene-expression data can be combined with marker data to map loci that regulate expression. Here we consider using gene-expression data from sibships to cluster genes in terms of their correlation across samples; tightly correlated genes are likely to be co-regulated or have similar function. Standard un-stratified measures of correlation such as Pearson's correlation can produce spurious results when applied to family data. We present a sibship-stratified measure of correlation, the FEXAT correlation. We show via simulation that the FEXAT recovers information obscured by Pearson's correlation. We then compare the performance of several clustering algorithms using the standard Pearson's and FEXAT correlations when applied to gene expression data from fifteen CEPH sibships. We find moderate agreement between the clusters based on the two correlations.

69

Random Coefficient Models in Genome-Wide Pharmacogenetic Searches for Drug Response QTLs in Vitro

AT Kraja, JW Watters, HL McLeod and MA Province
Washington U St. Louis

CEPH cell lines are a powerful resource for conducting searches for pharmacodynamic or pharmacokinetic QTLs. They have large kindreds that are highly informative for linkage, with publicly available genome scan marker data (Jean Dausset Foundation). We obtained N=692 cell lines (Corriell) out of which 31 CEPH pedigrees were treated with two drugs commonly used in cancer chemotherapy-docetaxel (DO) and 5-FU (FU). Each medication was administered to all lines in 12 replications at each of 9 different doses. We characterize the individual drug response toxicity using 3 classes of phenotypes: 1) IC50 (dose which inhibits 50% of growth) 2) fixed dose viability (VIAB % cell survivability relative to untreated) and 3) derived parameters for the entire dose-response curve using a random coefficient model (RCM). For the later, we transform the logistic dose-response to a linear scale and fit a linear mixed model via REML to estimate the individual subject responses. Subject specific RCM derived intercepts (RCMa) are related to their IC50s, while their slope (RCMb) indexes the rate of individualized dose-response changes. We use a variance-components linkage model (SEGPATH) for identifying chemotherapy toxicity/tolerance QTLs, This gives overall heritabilities for IC50 52% (DO) 21% (FU); VIAB 72% (DO) 43% (FU); RCMa 83% (DO) 55% (FU), RCMb 72% (DO) 50% (FU) with the maximum signal seen in the preliminary data so far being LOD=3.4 on chromosome 9. Since the RCM characterizes the entire dose-response curve for each subjectit provides an efficient way to use all of the data in pharmacogenetic genome scans.

70

Computational Sequence Comparison of Orthopoxvirus and Human Complement Regulators

J. Krushkal (1), D. Konz (2), A. Emery (2), E. Ciulla (2), R. Adkins (1), and I. Gigli (3)
(1) University of Tennessee Health Science Center, Memphis, TN; (2) Worcester Polytechnic Institute, Worcester, MA; (3) Institute of Molecular Medicine, University of Texas – Houston, Houston, TX

Many orthopoxviruses are important infectious agents in public health and disease. They include variola major virus (the cause of human smallpox). Their genomes contain genes similar to human regulators of complement activation (RCA). This allows orthopoxviruses to down regulate the human complement and to evade the host defense. Both poxviral and human complement regulators contain tandem short consensus repeats (SCRs).We evaluated the relationships among orthopoxvirus and human RCA proteins using phylogenetic sequence comparisons. Two groups of viral RCA proteins were studied. The first group included complement regulators in variola major virus (smallpox inhibitor of complement enzymes, or SPICE), vaccinia virus (vaccinia virus complement control protein, or VCP), and cowpox virus (inflammation modulatory protein, or IMP). We summarize relationships between domains of poxviral and human complement regulators. Computational prediction of the possible functional role of the single nucleotide polymorphisms in human and poxviral

complement regulators is currently underway. The second group of poxviral proteins analyzed included plaque size-host range proteins B7R in variola virus, B5R in vaccinia virus, B4R in cowpox virus, and C1R in ectromelia virus. These proteins, which have a function other than complement regulation, were found to have a different SCR organization than active RCA proteins.

71

Analysis of Longitudinal Phenotypes using SNPs

B. Kulle, H. Bickeböller
Dept. of Genetic Epidemiology, University of Göttingen, Germany

Only few methods are available for the analysis of quantitative traits in longitudinal genetic epidemiological association studies. For independent subjects we introduce a nonparametric factorial design for longitudinal data, modeling the phenotypic value as the dependent variable. Factors are the genotype at the considered SNP and the time categories corresponding to the measurements. To identify an associated marker a rank statistic tests if there is a difference in the time course across the genotypes. If there are no changes over time one can also test for the main effect of the genotype. No assumptions are made on normality or variances of the dependent variable. The dependent variable can also be ordinarily scaled. This approach is compared with the ANOVA-approach, which is often used in such cases.

72

The WECARE Study to assess genetic susceptibility to radiation-induced breast cancer: A counter-matched case-control study

B Langholz (1), DC Thomas (1), WD Thompson (2), J Bernstein (3)
(1) Dept of Prev Med, U of Southern California, (2) Dept of Applied Med Sciences, U of Southern Maine, (3) Dept of Community and Prev Med, Mt. Sinai School of Med.

The source population for the WECARE Study is 30,000+ breast cancer patients ascertained through an international consortium of five population-based cancer registries, all of which routinely record (perhaps imperfectly) whether radiation treatment was given as part of the cancer treatment regimen (registry radiation treatment, RRT+/RRT−). The outcome of interest is asynchronous bilateral breast cancer. Cost constraints dictated that we could draw 2 controls for each of the 700 cases in the cohort. In this sample we are assessing: 1) cumulative amount of scatter radiation dose received by the contra-lateral breast for each woman who received radiation treatment; and 2) ATM, BRCA1, and BRCA2 mutation carrier status. These data will be used to assess the role of these genes in determining the risk of (second) breast cancer as a function radiation dose, a $G \times E$ interaction. In designing the study we considered different ways of sampling 2 controls for each case, including RRT counter-matched designs. It was found that, for plausible gene carrier prevalence and RRT/true RT status correlations, counter-matching with 1 triplet member RRT- and 2 RRT+ was the most statistically powerful. As one of the first studies that implements the counter-matched design, we discuss the power relative to alternative designs, analysis of the data and practical issues that have arisen in the implementation of the design in the WECARE Study.

73

A Genome Scan of Serum Leptin Levels in Families with Sleep Apnea

E.K. Larkin(1, 2), R.C. Elston(1), S.R. Patel(3), L.J. Palmer(4), P.V. Tishler(3), S. Redline(2)
(1)Dept. of Epidemiology & Biostatistics, Case Western Reserve Univ.USA, (2)Dept.of Pediatrics, Rainbow Babies and Children's Hospital, USA (3)Channing Laboratory, Brigham & Women's Hospital, USA (4) Western Australian Institute for Medical Research, Australia.

Obstructive sleep apnea (OSA) is a prevalent complex disorder with substantial co-morbidity. The cause of OSA remains unknown. One leading hypothesis supposes that abnormalities in leptin secretion or function impacts OSA by influencing obesity, the major risk factor for OSA. Leptin, a plasma-circulating hormone secreted by adipocytes, affects hypothalamic regulation of satiety and hunger. A genome scan with 402 microsatellite markers of 156 sib-pairs representing 23 white and 36 African American families from the Cleveland Family Study was undertaken to identify quantitative trait loci that may be important in the regulation of serum leptin levels. Race-specific model-based and model-free analyses of a direct measure of sleep apnea, the apnea hypopnea index, were conducted initially because that phenotype appears markedly different between whites and African-Americans. However, by using an intermediate phenotype such as leptin, and after adjusting for covariates (including sex, race, and body mass index), we here combined data from the two races and unify some of the linkage results found in the race-stratified analyses of OSA. In particular suggestive linkage was found for leptin that both overlapped and differed from areas of linkage found in race-specific analyses of the apnea hypopnea index.

74

Optimal Selection Procedures for Linkage Analysis of Complex Traits

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

We propose a general strategy for selecting informative samples based on trait values. Our criterion uses the classical statistical concept of Fisher's information. In the context of quantitative traits, the information is calculated for an inverse variance components model i.e. the proportion of alleles shared identical by descent are considered conditional on trait values. The criterion is closely related to the score test [1] and in the context of sib-pairs designs to the optimal Haseman-Elston regression [2].

An advantage of the present approach is that simulations are not required, closed forms are obtained for the information and give useful insight on alternative designs. This continuous setting can easily be extended to dichotomous traits assuming an underlying continuous liability threshold model and it is shown that for common traits, discordant pairs provide valuable information relative to affected sib-pairs. The score test provides a valid method of analysis for selected samples thus obtained as described in another abstract in this meeting [3]. This sample selection strategy is currently being used for the design of linkage studies in the GenomEUtwin project [4]. [1] Putter et al. Genet. Epidemiol. 22, 2002, pp. 345–355 [2] Sham et al. Am. J. Hum. Genet. 68, 2001, pp. 1527–1532 [3] Putter et al. Abstract, IGES conference, 2003 [4] www.genomeutwin.org

75

Retinal Vessel Measurements Among Family Members in the Beaver Dam Eye Study

K.E. Lee (1), B.E.K. Klein (1), R. Klein (1), M.D. Knudtson (1), P. Duggal (2)
1)Department of Ophthalmology and Visual Sciences, University of Wisconsin, Madison, WI. 2)Statistical Genetics Section, NIH/NHGRI/IDRB, Baltimore, MD, USA

A narrowing of the retinal arterioles has been shown to be predictive of cardiovascular disease, diabetes and stroke. Retinal vessels are easily imaged and caliber quantified. These analyses aim to investigate a potential genetic contribution to the vessel caliber among individuals from the Beaver Dam Eye Study. All persons aged 43–84 years and living in Beaver Dam, WI in 1988 are eligible for the study. Within this population, 56% (N=2,783) are identified as belonging to one of 602 extended pedigrees. An additional 1,656 have a spouse in the study. A standardized exam and interview was administered from 1988–1990 during which photographs of the retina were obtained (including ETDRS Field 1, centered on the optic nerve head). Computer-assisted grading of the retinal vessel diameters from the digitized photos allows estimation of the central retinal arteriole (CRAE) and central retinal venule (CRVE). These are then used to calculate the arteriole-to-venule ratio (AVR). Familial correlations are calculated using FCOR in the SAGE package (v4.3). Analyses are done on the right eye measures for 3,460 individuals, after adjustment for age, gender, blood pressure and smoking. Sibling correlations (from 869 pairs) are 0.23, 0.18 and 0.10 for the CRVE, CRAE and AVR respectively. Parent-child correlations (343 pairs) are 0.24 (CRVE), 0.25 (CRAE) and 0.18 (AVR). Avuncular correlations (547 pairs) are 0.17 (CRVE), 0.11 (CRAE) and 0.10 (AVR). Cousin correlations (1359 pairs) are 0.09 (CRVE), 0.07 (CRAE) and 0.03 (AVR). Spousal correlations (893 pairs) are 0.03 (CRVE), 0.04 (CRAE) and 0.01 (AVR). These correlations are consistent with the possibility of a genetic determinant of vessel caliber. Commingling and segregation analyses are currently underway and will be presented.

76

The bias in allele sharing methods caused by transmission ratio distortion

M. Lemire(1), N.M. Roslin(1), C. Laprise(2), T.J. Hudson(3)K. Morgan(1, 4)
(1) Research Inst. of the McGill Univ. Health Centre (2) Centre de médecine génique communautaire, Univ. du Québec à Chicoutimi(3) McGill Univ. and Genome Quebec Innovation Centre, (4) Depts. Of Human Genetics and Medicine, McGill University, Canada

Transmission ratio distortion (TRD) is a departure from Mendel's rules of inheritance. It is a mechanism that could be observed for example, in the region of a gene under selection, causing post-fertilization loss. It is a confounding factor for tests of linkage that are based on the evaluation of excess sharing of alleles among affected relatives, when the TRD is independent of the trait or disease under study. Reported regions of TRD are sparse, and as a consequence little emphasis has been placed on the control of this confounding factor. We study the bias caused by disease-independent TRD in tests of linkage based on the identity by descent of alleles in affected relative pairs. The bias depends on the relationship between the affected individuals in a pedigree, and can either inflate or reduce the true type I error of the test. As an illustration we use a sample of pedigrees ascertained for asthma for which a modest linkage signal was obtained in the 6q26 region (multipoint score of 2.23, p=0.013, based on the exponential model of Kong and Cox (1997) [Am J Hum Genet 61:1179] and the Spairs statistic of Whittemore and Halpern (1994) [Biometrics 50:118]). TRD has been observed for markers in this chromosome region in three-generation CEPH families unselected for disease.

77

Haplotype Association Analysis for Late Onset Diseases Using Nuclear Family Data

C. Li(1), M. Boehnke(2)
(1)Program in Human Genetics, Vanderbilt Univ, USA (2) Dept. of Biostatistics, Univ of Michigan, USA

In haplotype-based association studies for late onset diseases, one attractive design is to use available unaffected spouses as controls (Valle et al. 1998). Given cases and spouses only, the standard expectation-maximization (EM) algorithm (Dempster et al. 1977) for case-control data can be used to estimate haplotype frequencies. But often we will have offspring for at least some of the spouse pairs and their genotypes provide additional information about the haplotypes of the parents. Existing methods may either ignore the offspring information, or reconstruct haplotypes for the subjects using offspring information and discard data from those whose haplotypes cannot be reconstructed with high confidence. Neither of these approaches is efficient, and the latter may also be biased. For case-spouse control data with offspring genotypes available for some spouse

pairs, we propose a unified, likelihood-based method of haplotype inference. The method makes use of available offspring information to apportion ambiguous haplotypes for the subjects. For subjects without offspring information, haplotypes are apportioned as in the standard EM algorithm for case-control data. The method enables efficient haplotype frequency estimation using an EM algorithm and supports probabilistic haplotype reconstruction with the probability calculated based on the whole sample. We describe likelihood-ratio and permutation tests to test for disease-haplotype association and describe three test statistics that are potentially useful for detecting disease-haplotype association.

78

Relation of Insulin-Like Growth Factor (IGF)-I and IGF Binding Protein-3 Gene Polymorphisms to Prostate Cancer Risk: A Sibling-Matched Case-Control Study

L. Li (1), M. Cicek (2), F. Schumacher (1), G. Casey (2), J. S. Witte (2)
(1) Department of Epidemiology & Biostatistics, Case Western Reserve University; (2) Department of Cancer Biology, Cleveland Clinic Foundation

Circulating levels of IGF-I, or IGF-I in relation to IGFBP-3, have been related to risk of prostate cancer in a number of epidemiologic studies. Genetic influence is a major determinant of inter-individual variability of IGF-I and IGFBP-3. However, the potential role of IGF-I and IGFBP-3 genetic polymorphisms in the etiology of prostate cancer is essentially unknown. We evaluated the relationship of a (CA)n repeat polymorphism in the IGF-I gene and a single nucleotide polymorphism in the promoter region (−202) of IGFBP-3 gene with prostate cancer risk in a sibling-matched case-control study. Four hundred forty cases and their 479 older sibling controls from 414 discordant families were included. IGFBP-3 gene polymorphism was significantly associated with serum levels of IGFBP-3 (mg/mL): the means were 26.4, 24.0, and 21.6, respectively, for genotype AA, AC and CC (p for trend <0.05). IGF-I (CA)n repeat polymorphism, however, was not related to serum levels of IGF-I. In multivariate conditional logistic models (IGF-I genotype at allele-level as a continuous variable), we found no significant association between IGF-I genotype and risk of prostate cancer: compared to those with 19 (CA)n repeats, the odds ratios were 0.97 (0.67–1.41), and 0.99 (0.78–1.27), respectively, for those with ≤ 18 (CA)n repeats and those with ≥ 19 (CA)n repeats (p for trend >0.5). Similarly, no significant association was found for IGFBP-3 genotypes: compared with those homozygous for CC, the odds ratios were 0.81 (0.51–1.30), 0.76 (0.40–1.44), respectively, for those with AC and AA genotypes (p for trend >0.5). Further adjustment for serum levels of IGF-I or IGFBP-3 did not materially altered the results. Stratified analysis by disease aggressiveness yielded similar results. Our data do not support an important link of these polymorphisms to prostate cancer susceptibility.

79

Genome-wide scan of clamp defined insulin resistance in Mexican American coronary artery disease families

X. Li(1), M.J. Quiñones(2), D. Wang(1), M.F. Saad(2), I. Enriquez(2), X. Jimenez(2), G. Hernandez(2), R. De La Rosa(2), W.A. Hsueh(2), H. Yang(1), J.I. Rotter(1)
(1) Cedars-Sinai, USA, (2) UCLA, USA

Insulin resistance (IR), resulting from both genetic and environmental factors, is a risk factor for coronary artery disease (CAD). We performed a genome-wide linkage analysis on direct measures of IR using data from 101 Mexican American nuclear families ascertained via a parent (proband) with documented CAD. Through the euglycemic clamp, two IR phenotypes, the glucose infusion rate (GINF) and whole body insulin sensitivity (SI, calculated as the ratio of GINF to the change in plasma insulin over basal), were measured in adult offspring and offspring spouses (n=438). 499 adult offspring and their parents were genotyped for 408 microsatellite markers along the genome at ~10cM density. Heritability estimates (h^2) and multipoint linkage analysis were performed using a variance components procedure implemented in SOLAR. H^2 were 0.40 for GINF and 0.57 for SI respectively (all $P<0.001$). Evidence for suggestive linkage (lod score, LOD>=2) was observed on chromosome 21 for GINF with the maximum LOD 2.39 at 25 cM. For SI, the maximum LOD was 3.36 at 90 cM on chromosome 19, and two additional suggestive linkage peaks were observed on chromosomes 1 (114 cM, LOD 2.07) and 14 (76 cM, LOD 2.82). Other linkage signals (LOD>=1.3) were observed on chromosomes 2, 6 and 10 for GINF, and on chromosomes 2, 4, 9, 13 and 15 for SI. These results indicate that there is a strong genetic influence on IR in CAD families and we may have identified one or more susceptibility locus for IR.

80

Evidence for a gene influencing hematocrit on chromosome 6q: a genome-wide scan in the Framingham Study

J.-P. Lin(1), C.J. O'Donnell(2), D. Levy(2), L.A. Cupples(3)
(1)DECA/NHLBI/National Institutes of Health, USA, (2)DECA/NHLBI/Framingham Heart Study, USA, (3)Dept. of Biostat, Boston University School of Public Health, USA

Many studies have shown that hematocrit levels are significantly associated with cerebrovascular, cardiovascular, and peripheral vascular diseases, as well as all-cause mortality. The mortality of subjects with high hematocrit has been shown to be six times higher than expected. Twin studies in humans have shown that hematocrit variation is largely determined by genetic factors with heritability estimated as 40%–65%. So far, no linkage analysis on HCT in humans has been reported. We carried out a 10 cM genome-wide scan in a community-based Caucasian cohort, the Framingham Heart Study. Our study population consisted of 330 families with 1534

individuals being both genotyped and phenotyped, including 1375 sibling pairs, 88 cousin pairs, and 180 avuncular pairs. Using variance-component linkage methods implemented in SOLAR, the heritability was estimated as 41% after age, sex, weight, alcohol intake, smoking, total cholesterol, HDL cholesterol, triglyceride, and diabetes adjustment. The genome-wide linkage analysis demonstrated evidence of significant linkage of hematocrit to chromosome 6q23–24 with a LOD score of 3.4 at location 134 cM. Within this region, a candidate gene, EBP41L2, was newly mapped to (130cM). EBP41L2 is a member of the erythrocyte membrane skeletal protein 4.1 (EPB41) gene family. Mutations in the EPB41 gene cause hereditary elliptocytosis. Another candidate gene, coding for a putative heme-binding protein, HEBP2, was mapped to this region also (138cM). Function studies of EBP41L2 and HEBP2 have not been reported. Only one other region in the genome produced a multipoint LOD score greater than 1.5 (LODs=1.7). Further studies would be worthwhile to determine the relationship between hematocrit and these genes.

81

Haplotype structures and analyses of 21 SNPs in K-ATP genes in a large cohort: do different methods result in similar conclusions?

JA Luan(1), I Barroso(2), RPS Middelberg(1), AH Harding(1), PW Franks(1), RW Jakes(1), M Sandhu(1), S O'Rahilly(3), AJ Schafer(2), NJ Wareham(1)
(1)Inst Pub Health, (3)Dept Clin Biochem & Med, Univ Cambridge, UK, (2)Incyte, CA, USA

Genetic epidemiological studies have demonstrated that SNPs in the pancreatic ATP-sensitive potassium (K-ATP) channel subunits sulfonylurea receptor 1 (ABCC8/SUR1) and the inwardly rectifying K+ channel Kir6.2 (KCNJ11) are associated with Type 2 diabetes. We genotyped 21 diabetes candidate SNPs in both genes in a large population based cohort (N=921). Haplotype analyses were performed using four SNP selection methods for haplotype reconstruction: (a) using all SNPs; (b) LD SNP blocks (strong LD among SNPs within blocks); (c) SNP reduction by haplotype tagging; (d) SNP reduction by general linear model using 2hr glucose as the outcome. The intermediate traits of diabetes (BMI, fasting insulin, fasting glucose and 2hr glucose) were used for testing associations among haplotypes. The three most common haplotypes (freq > 5%) obtained by (a) only accounted for 25% of the total sample thereby losing data. A 12 SNP block and a 7 SNP block were yielded by (b). However, none of the haplotypes from either block was associated with any phenotypes. Haplotype tagging method (c) identified 5 of the 21 SNPs, but no association was detected. Three SNPs were selected by (d) and we found that this haplotype was associated with fasting insulin. Our results show that different haplotype structures lead to different conclusions. Further research into methods of reducing the number of SNPs in haplotype analysis is required.

82

A Comparison of SNP Ranking Methods for Large Scale Association Studies

K.L. Lunetta, A. Bureau, P. Van Eerdewegh
Genome Therapeutics Corp., Waltham, MA, USA

The study of complex diseases is entering a phase where vast amounts of genetic information will need to be analyzed and interpreted. In the context of large-scale association studies, one critical task is to identify the most important predictor variables, such as SNPs and environmental covariates, from among large sets of potential predictors and their interactions. Using simulations, we compare the ranking of SNPs using three analysis methods: 1) Random Forests, a tree-based method in which forests of trees are grown on bootstrap samples of the observations, using a random subset of the potential predictors to grow each tree. The predictive importance of the variables is quantified using the individuals not used in growing each tree. 2) POLYCLASS, a nonparametric adaptive regression technique that fully automates model selection based on cross-validation. Variables in the final model can be ranked using an importance-anova decomposition. 3) The p-values of the Fisher exact test for the tables comparing case and control SNP genotype frequencies. The methods are compared using data sets of 500 cases and 500 controls and 100 SNPs. Four of the SNPs are simulated to interact to increase risk of a binary trait using several models of interaction and a range of allele frequencies. The remaining non-associated SNPs have a range of allele frequencies. The three methods' rankings of the 4 risk SNPs among all the SNPs are similar for many models. The modeling methods, which choose SNPs based on prediction accuracy, are less likely to rank rare risk SNPs highly. However, these methods can give more information about the underlying model.

83

Challenges and Successful Strategies in Recruitment: Experience from Prostate Cancer Study in Southern Louisiana.

D.M. Mandal(1)S.L. Halton(1), T.N. Turley-Stoulig(1), E.M. Gillanders(2), J. Carpten(2), J. Trent(3), J.E. Bailey-Wilson(2), W. Rayford(4)
(1)Dept. of Genetics, (4)Dept. of Urology, LSU Health Sciences Center, New Orleans, LA; (2)NHGRI/NIH, Bethesda, MD; (3)Translational Genomics Research Institute, Phoenix, AZ.

To date, three genes have been successfully cloned for prostate cancer. However, more data are needed to confirm the suggested linkages and to identify and characterize specific mutations in diverse populations. We have initiated a study to ascertain families with history of three or more prostate cancer cases through the prostate cancer-screening program at the LSUHSC Department of Urology, from participating local urologists and collaborative hospitals from Southern Louisiana. Entering participants into the study has been extremely challenging

due to the raised concern of patient confidentiality and privacy issues, but we have been able to build some successful collaboration through innovative approaches. Thus far, 224 individuals have been interviewed. Of those, nine reported a significant family history of prostate cancer. Biological sample collection and pathological confirmation are complete on three of those families, which are in the process of genome screening. Follow-up is ongoing in the remaining six families. A description of the data resource will be presented, and the critical issues involved in recruitment in a genetic study will be discussed, with particular emphasis on the challenges due to the raised concern of patient confidentiality and privacy issues and innovative approaches used in successful recruitment.

84

Penetrance estimation via maximization of retrospective likelihood using genetic test results of families screened for BRCA1/2 mutations

F. Marroni(1, 2), P. Aretini(2), G. Bevilacqua(2), J.E. Bailey-Wilson(3), G. Parmigiani(4), S. Presciuttini(1, 3), and the Italian Consortium for Hereditary Breast and Ovarian Cancer
1) Dept Biomedicine, Univ. Pisa, Italy; 2) Dept Oncology, Univ. Pisa, Italy; 3) NHGRI/NIH, Baltimore, MD, USA; 4) Dept Biostatistics, JHU, Baltimore, MD, USA

We collected 568 families of Caucasian ancestry from five centers participating in the Italian Consortium for Hereditary Breast and Ovarian Cancer; 80 mutations were detected in BRCA1 and 53 in BRCA2. We maximized the likelihood of the genetic test results by perturbing, through a Markov-Chain Monte Carlo method, the genetic model used to predict the presence of a BRCA1/2 germline mutation in a proband. The software BRCAPRO was used as a carrier probability calculation tool. Two different sets of four penetrance functions (breast and ovarian cancer in BRCA1 and BRCA2 carriers) were estimated separately, a ''linear'' and a ''quadratic'' model, respectively. The final log-likelihood was similar for both sets, and the penetrance curves of the two sets were practically indistinguishable, indicating convergence towards the same optimal genetic model. This had a total log-likelihood value about 70 units higher than that of the original model, based on published penetrances. Our data refer to a substantial proportion of the families that currently require genetic counseling in Italy; therefore our new penetrance estimates provide the most accurate genetic model so far available for this population segment and may lead to a mutation-predicting model specifically adapted to this country.

85

Linkage disequilibrium analysis of SNPs and schizophrenia

M Martinez, J Duan, AR Sanders, L Martinolich, EB Carpenter, BJ Mowry, DF Levinson, RR Crowe, JM Silverman, PV Gejman
INSERM, EM106, Evry, France

The 6q13-q26 region has been implicated as harboring a schizophrenia susceptibility gene (SCZD5) in our and other independent family collections. To refine the region and localize the gene(s) of interest we have undertaken a linkage disequilibrium analysis using family-based association tests. Our sample contains 214 families (>800 subjects with DNA). Most families have >1 affected offspring, 20% are extended pedigrees (>2 generations), and not all affected offspring have both parents with DNA. So far, we have saturated the linked region with 51 SNPs of 17 putative candidate genes. Since our association tests have to be conducted under the null hypothesis that there is linkage but no association (H'0), we first performed a classical-TDT analysis using only one affected offspring by family. The number of informative triads (N=informative transmissions to affected offspring) can result too small to ensure accurate statistical significance. Several approaches have been developed to test H'0: FBAT(Horvath et al.2003), PDT (Martin et al.2000) and TRANSMIT (Clayton, 1999). These methods are expected to provide similar outcomes in simple cases. They differ, however, in the way they use and/or reduce the family information: pedigrees are broken down into nuclear families (PDT and TRANSMIT) or not (FBAT); only fully informative triads (both parents genotyped) are used (PDT) or not (FBAT, TRANSMIT). Here, we report association results based on different sampling strategies and family-based association methods, and show that the associated SNPs and significance levels vary substantially from one approach to the other. For instance ACAT2-SNPs were significant with FBAT (p=0.03) and TRANSMIT (p=0.01) but not with PDT (p=0.10).

86

MER2SOL: Translating MERLIN or Loki IBD Data to SOLAR Format

M.B. Miller
Division of Epidemiology, Univ. of Minnesota, USA

The multipoint identity by descent (IBD) approximation implemented in SOLAR does not use phase information. Because of this, the approximation sometimes produces inaccurate multipoint IBD probabilities that can significantly reduce power to detect QTLs. I demonstrate both of these facts using simulated data. Most software programs that compute multipoint IBD probabilities force the user to specify a single set of marker allele frequencies, even with ethnically heterogeneous data where a single set is not appropriate. Few programs for QTL mapping allow the user to store IBD probabilities for later analyses (SOLAR is one exception). Few programs allow for use of different multipoint IBD computational algorithms (e.g. Lander-Green or MCMC) for different families within a single analysis. Finally, few programs can handle MZ twin family data properly in variance components analysis (SOLAR is an exception). I have developed the program 'MER2SOL' to solve all of the problems just described for users of SOLAR. First, multipoint IBD probabilities are computed in MERLIN (Lander-Green) or Loki (MCMC) on individual families, or sets of families, using appropriate allele frequencies. Then MER2SOL can be used to convert

the collection of multipoint IBD output files to a single SOLAR 'mibd' database. An additional pair of scripts is used to prepare MZ twin-family linkage-format files for use with MERLIN or Loki (by removing one of the twins), and to process the IBD output (by adding the twin back in) before converting the files with MER2SOL. MER2SOL, written in C, is freely available under the GNU Public License.

87

Sequential QTL Mapping by Variance Components in Ordered Subsets

M.B. Miller(1), W. Tang(1), M. Province(2), S. Hunt(3), D. Arnett(1)
(1)Division of Epidemiology, Univ. of Minnesota, USA, (2)Division of Biostatistics, Washington University, St. Louis, USA, (3)Cardiology Division, University of Utah, USA

An unpublished sequential approach to QTL linkage analysis orders individuals on a covariate, selects those individuals for linkage analysis who exceed some threshold on the covariate, then changes the threshold monotonically until all individuals are included in the linkage analysis. The method is implemented at a single chromosomal location using Haseman-Elston (H-E) regression in sibling pairs. This produces a series of LOD scores as functions of the covariate threshold and can reveal previously undetected evidence for linkage or can enhance evidence from earlier analyses of all subjects. We propose an extension to this method that allows for use of general pedigrees, whole chromosome scans, and use of variance components methodology instead of H-E regression. The method is applied to data from the NHLBI HyperGEN ECHO study using left ventricular mass as the linkage trait and body mass index as a covariate revealing a maximum LOD score of 3.2 on chr 7. Our software, which prepares data for SOLAR analyses and manipulates the output for presentation, is freely available. Alternative interpretations of our results in terms of GxE interaction, epistasis, heterogeneity with pleiotropy, and false positive error are considered.

88

Finding susceptibility loci in case-control studies of candidate genes in the presence of epistatic interactions

J. Millstein, W.J. Gauderman
University of Southern California, Los Angeles, USA

For multifactorial diseases accounting for possible gene-gene interactions may prove an integral part of any strategy designed to identify susceptibility genes among candidate genes. However, the number of possible interactions can be large and the resulting multiple testing problem cannot be ignored. We developed a search strategy to address this issue. Candidate genes are tested in stages according to the order of interaction, starting with the marginal effects of each gene. We perform likelihood ratio tests based on a logistic model. Each stage considers interactions of a given level and lower order terms. Each model is compared against a null model containing an intercept term and any lower order terms that were significant in reduced models. The number of gene sets to test for possible interactions is reduced in a preliminary step that ranks gene sets according to a chi squared statistic based on observed vs. expected number of subjects with variant alleles at all loci in each gene set. The upper ranked gene sets are selected for analysis of possible epistatic interactions. We compared strategies to find susceptibility genes, not identified by their marginal effects alone, using simulated data. Results: In the presence of interaction single degree-of-freedom tests with Bonferroni corrections and no gene set eliminations were limited in effectiveness. Our approach more than doubled the number of new genes found. Our approach was further improved by eliminating those gene sets least likely to influence disease risk. Conclusions: Accounting for possible interactions using multi degree-of-freedom tests and reducing the number of gene sets considered for interactions can significantly increase the number of susceptibility genes found in the presence of epistasis.

89

Tests for Covariate-Associated Heterogeneity in IBD Allele Sharing of Affected Relatives

L. Mirea(1), L. Briollais(1, 2), S.B. Bull(1, 2)
(1)Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON, Canada, (2)Department of Public Health Sciences, University of Toronto, Toronto, ON, Canada

Conventional linkage tests based on identical-by-descent (IBD) allele sharing among affected relatives do not allow for possible differences among families, such as arise in the case of locus heterogeneity, and thus have reduced ability to detect linkage in the presence of such heterogeneity. We have investigated tests for heterogeneity of non-parametric linkage (NPL) scores with a family-level covariate, which may be associated with different disease mechanisms leading to differences in IBD allele sharing. Likelihood ratio tests for heterogeneity were formulated based on an extension of the linear and exponential likelihood models developed by Kong and Cox (1997 Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188). Alternatively, we examined the asymptotic and permutation distributions of T-tests for differences between mean NPL scores from two covariate-defined family groups, assuming exchangeability. The size and power of heterogeneity tests were evaluated for Sall and Spairs scoring functions using data sets of families with affected sibling and cousin pairs generated under a model of locus heterogeneity. In certain simulation scenarios, the likelihood ratio statistics did not follow the expected asymptotic distributions. The type I error estimates for the T-statistics conformed to nominal 5% and 1% levels in all scenarios considered and corresponding power was comparable to that of the likelihood ratio tests.

90

Fine-Scale Mapping of Diseases via Spatial Clustering Techniques

J. Molitor, P. Marjoram, D.C. Thomas
Department of Preventive Medicine, Keck School of Medicine, University of Southern California, USA

We present a method for performing fine mapping and analyzing haplotype risk. The method uses techniques widely applied in the context of spatial clustering. Haplotypes are assigned to clusters according to a similarity measure based upon IBS shared length around putative functional mutations. Each cluster has an associated risk. The space of possible clusters, functional mutations and model parameters is explored using Markov chain Monte Carlo methods. Our approach uses the information at all markers simultaneously, allows naturally for the existence of missing data, and extends easily into contexts in which there might be more than one functional mutation influencing the phenotype.

91

Connecting the Dots Between Genes, Biochemistry and Disease Susceptibility

J.H. Moore(1), L.W. Hahn(1)
(1)Program in Human Genetics, Vanderbilt University, Nashville, TN, USA

Understanding how DNA sequence variations impact human health through a hierarchy of biochemical and physiological systems is expected to improve the diagnosis, prevention, and treatment of common, complex human diseases. We have previously developed a hierarchical dynamic systems approach based on Petri nets for generating biochemical network models that are consistent with genetic models of disease susceptibility. This modeling approach uses an evolutionary computation approach called grammatical evolution as a search strategy for symbolic manipulation and optimization of Petri net models. We have previously demonstrated that this approach routinely identifies biochemical network models that are consistent with a variety of genetic models in which disease susceptibility is determined by nonlinear interactions between two DNA sequence variations. In the present study, we have successfully extended this approach for the identification of biochemical networks that are consistent with disease susceptibility due to high-order nonlinear interactions. We find that the form of the grammar used by the evolutionary search algorithm for specifying initial Petri nets is critical for successful model discovery. In fact this may be more important than the length and size of the computational search. We speculate that this modeling approach will play an important role in facilitating thought experiments with the goal of generating hypotheses about the nature of biochemical systems models that are consistent with genetic models of disease susceptibility.

92

Confirmation of SLE susceptibility locus, SLER1, at 5p15.3

S Nath, B. Namjou, P. Garriott, S. Frank, P. Joslin, P. Joslin, J. Kilpatrick, J. Kelly, J.B. Harley
Oklahoma Medical Research Foundation, Oklahoma City, USA

For disorders with a poorly known biochemical basis like systemic lupus erythematosus (SLE), identification of the genes is a prerequisite to understanding its biological basis. We have previously identified (Namjou et al. 2002) a potential genomic region at 5p15.3 ($P < 0.0000001$), which contains a novel susceptibility gene, SLER1, for SLE. This finding was based on a selected group of 14 SLE families (SLE-RA) with European-American ancestry, where two or more family members also had been diagnosed with rheumatoid arthritis (RA).Replication of initial linkage signals from independent samples is considered an important and crucial step toward distinguishing between true positives and false positives. To confirm linkage at 5p15.3, we performed linkage analyses on six markers from the linked region in an independent dataset of 88 such SLE-RA families. Since the disease model was uncertain, especially when families are from multi-ethnic backgrounds (present study), we initially analyzed our data to confirm the established linkage at 5p15.3, based on a non-parametric allele-sharing method. Our new results replicated the initial linkage signal at 5p15.3 ($Zlr=2.58$, $P < 0.005$), which is well beyond the recommendation for linkage confirmation. This evidence of linkage substantially increased when analysis was restricted to the subset of families who had 3 or more individuals with alleged RA. The results of previous findings, together with our new results, have provided strong evidence that confirm SLER1 at 5p15.3 is a susceptibility locus for SLE especially in the families multiplex for both SLE and RA. Our results clearly demonstrate that carefully constituted sub-group analysis can be extremely powerful for detecting susceptibility gene(s) for complex diseases like SLE.

93

Comparison of Allele Sharing Linkage Methods Utilizing Covariates Under Locus Heterogeneity

K.K. Nicodemus, D. Fallin
Johns Hopkins School of Public Health, Baltimore, Maryland, USA

Although linkage methods have been successful in locating disease genes for diseases that are monogenic they have not been a complete success in locating genes for complex diseases. Reasons for this include locus and phenotypic heterogeneity and low power to detect genes with small effects. To attempt to account for genetic heterogeneity several covariate-based methods of non-parametric linkage analysis have been proposed, including ASM and LODPAL. However, it is unclear which methods are optimal under which conditions. To assess the performance of these methods, we simulated data for

one chromosome (190 cM) in 1000 datasets of 200 families. Marker spacing was set at 10 cM apart. Two disease loci were simulated (20 cM, 170 cM). Each family had only one disease locus. In addition, a covariate was simulated with different means in each linkage group (difference between means =1.5 standard deviations). The disease allele was set to a frequency of 0.01 and disease penetrance was set to 0.75 for individuals having at least one copy of the disease allele. To assess statistical significance, data was simulated for each family under the hypothesis of no linkage and analyses were repeated for each method. We considered observed linkage statistics significant if they were greater than 95% of the statistics generated under the null hypothesis. ASM within 10 cM on either side of the disease loci identified significant linkage in 100% of the replicates, whereas LODPAL analyses identified significant linkage in 99% of the replicates. Further work will include the assessment of additional methods (Genefinder, Ordered Subsets Analysis) and will vary sample size, inclusion of ARPs in addition to ASPs, the number of disease loci, marker heterogeneity, family size, penetrance and magnitude of difference between covariate means in the differently linked subgroups.

94

Development and application of DNA repair capacity and DNA damage indices to prostate cancer

N.L. Nock(1), B.A. Rybicki(2), 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

Polymorphisms in DNA repair genes modify an individual's ability to repair DNA damage caused by exogenous compounds, possibly leading to carcinogenesis. One or more distinct DNA damage recognition and repair pathways may be invoked in response to a xenobiotic agent, depending on the type of damage induced, the level of existing damage from both endogenous and exogenous sources, and the cellular state. Although quantitative kinetic models have been developed, human DNA repair pathways and their interrelations remain incompletely characterized, prohibiting construction of an accurate gene regulatory network. Direct measurement of DNA repair capacity and DNA damage is also problematic because of transient effects and insufficient reliability of the assays. In an attempt to overcome these limitations, we present an approach that simultaneously models several key candidate genes from all relevant DNA repair pathways and all known sources of the exogenous compound of interest. Our approach utilizes ''biologically supervised'' hierarchical factor analysis (BSHFA) to construct DNA Repair Capacity and DNA Damage indices followed by structural equation modeling (SEM). We illustrate this method (BSHFA-SEM) with an application of the relation between polyaromatic hydrocarbons (PAHs) and prostate cancer. We also compare BSHFA-SEM to conventional ''one-gene/exposure-at-a-time'' regression and discuss its viability in revealing new insights into the causal mechanisms of complex diseases.

95

Gene influences cholesterol and triglycerides on chromosome 21q

KE North (1), MB Miller (2), H Coon (3), LJ Martin (4), B Zhang (1), M Province (5), A Oberman (6), L Almasy (7), J Blangero (7), D Arnett (2), J Peacock (2), RC Ellison (8), G Heiss (1)
(1) Epi Dep, UNC-CH, (2) Epi Div, U Minn, (3) Psych. Dep, U Utah, (4) Epi & Bios, Children's Hosp, (5) Bios Div, Wash U Med, (6) Prev Med, UAB, (7) Genetics Dep, SFBR, (8) Prev Med & Epi, BU Med

Strong correlations between cholesterol, triglycerides (TG), and body mass index (BMI) were reported in the Hypertension Genetic Epidemiology Network (HyperGEN), a multi-center study of genetic and environmental factors related to hypertension. Motivated by these findings and previous evidence for pleiotropy between cholesterol and TG levels, we conducted bivariate linkage analysis of total cholesterol and TG. Sibships in HyperGEN were recruited from five field centers located in Massachusetts, North Carolina, Minnesota, Utah, and Alabama. All available hypertensive siblings and their non-hypertensive-medicated offspring and/or parents were recruited. Among 1, 408 white and 1, 524 African-Americans, we performed a genome scan for quantitative trait loci influencing total cholesterol and TG. The Marshfield Genotyping Service typed a total of 391 microsatellite markers, spaced roughly equally throughout the genome. Plasma TG and cholesterol levels were measured using glycerol-blanked TG reagent on a centrifugal analyzer and a commercial cholesterol oxidase method, respectively. Both phenotypes were similarly adjusted for race, study center, sex, age, age-by-sex interactions, smoking, alcohol consumption, menopause, hormone use, diabetes medication use, and waist circumference. Variance component linkage analysis was performed as implemented in SOLAR, using race-specific marker allele frequencies derived from founders and multipoint IBDs calculated in MERLIN. A maximum genome-wide LOD score of 4.5 was detected on chromosome 21 at 51 cM, between markers D21S2055 and D21S1446. Analyzing each race separately, a maximum peak of 2.9 was detected on chromosome 21 at 48 cM in African Americans and of 1.7 on chromosome 21 at 55 cM in white participants. This signal overlaps with positive findings for total cholesterol, LDL cholesterol, and apolipoprotein B in three other genome scans and is suggestive of one or more genes on chromosome 21q jointly regulating total cholesterol and triglyceride concentration.

96

Association of polymorphisms in SOD2 in Alzheimer's disease patients

R.T. Perry(1), H. Wiener(1), L.E. Harrell(1), D. Blacker(2), R.E. Tanzi(2), M. McInnis(3), S.S. Bassett(3), R.C.P. Go(1)
(1)Univ. Alabama at Birmingham, USA, (2)Massachusetts General Hospital, USA, (3)Johns Hopkins Univ. USA

Oxidative damage due to free radicals and ROS appear to contribute to the pathogenesis of 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. A T → C substitution (V → A) at position –59 of the mitochondrial targeting sequence of SOD2 could affect its transport or cleavage. We found a significant increase in the Ala allele frequency under the dominant model in patients from a set of families in the NIMH AD data set with at least two affected and one unaffected sibling (p=0.024) and, in subsets of families that were not homozygous for APOE4 (p=0.033), with onset >50 (p=0.028), and >70 (0.029). However, the frequency of the Val allele was marginally increased (p=0.084) under the same model in patients from a cohort of African American (AA) cases and controls. We obtained similar p values for the G allele at a second polymorphism in SOD2 (position –59 of intron 3) in the same families of the NIMH data set, but like the Val-9Ala change, the frequency of the other (T) allele was significantly increased in the patients of the AA cohort under additive (p=0.019) and recessive models (p=0.007). The haplotypes in linkage disequilibrium may explain the observed differences. Interestingly, SOD2 is located close (<9 Mb) to a significant chromosomal region, 6q27detected in the AD genomic scan we reported this year.

97

Premature death of adult adoptees: Analysis of a case-cohort sample

L Petersen (1), P.K. Andersen (2), T.I.A. Sørensen (1)
(1) Danish Epidemiology Science Centre at the Institute of Preventive Medicine, Cph. Univ. Hosp. Denmark. (2) Dep. of Biostat.Univ. of Cph. Denmark.

Familial influence on human lifespan is well established. Twin- or adoption studies are used to estimate genetic and environmental effects. The unique Danish register of 14,427 non-familial adoptions, carried out during 1924-1947, contains information on biological and adoptive relatives. In two earlier studies, a cohort study of 1003 adoptees born in the period 1924–1926 (NEJM 1988; 318:727), and a case-control nested in the cohort study, encompassing all adoptions until 1947 (Genet Epidemiol 2002; 23:123), showed a moderate genetic influence on risk of premature death. This most likely stems from genetic influence on risk of dying from vascular causes. A genetic effect on death due to infections was found only in the cohort study. In an extension of the adoption study, we chose case-cohort sampling. As the case-control design, the case-cohort design has the advantage of economic data collection and little loss in efficiency, additionally rate estimates may be obtained, and re-use of the cohort sample in future studies of other outcomes are possible. Assuming a Cox proportional hazards model in the full cohort, several estimators of the covariate effect have been proposed. Based on a simulation study designed to cover the situation our data, we chose to use Prentice's estimator for hazard ratio, and robust estimation for variances. The preliminary results suggest that genetic effects on infectious diseases are caused by distinct effects on tuberculosis. Further results will be presented.

98

Sample size calculations for population and family based case-control association studies on marker genotypes

RM Pfeiffer, MH Gail
NCI, Bethesda, USA

Most previous sample size calculations for case-control studies to detect genetic associations with disease have assumed that the disease gene locus is known, whereas, in fact, markers are used. We calculated sample sizes for unmatched case-control and sibling case-control studies to detect associations between a biallelic marker and a disease governed by a biallelic disease locus. We evaluated the sample size requirements for two-sided trend tests with additive scores applied to marker genotypes, for both designs. The main factors influencing sample size apart from α-levels and relative risk parameters are: 1) the degree of agreement between marker allele and disease allele frequencies, which determine the maximal linkage disequilibrium; 2) the percent of maximal linkage disequilibrium present; 3) the attributable risk from the disease allele. Type of inheritance also plays a role. For a fixed attributable risk, disease prevalence, and, in the sibling case-control design, residual familial aggregation and recombination have much smaller impact. Qualitatively similar results have been found for the transmission disequilibrium test. We found that additive scores, which do not require knowing which marker allele is in positive linkage disequilibrium with the putative disease allele, are not very inefficient, and can even be advantageous in some settings. The large sample size requirements represent a formidable challenge to studies of this type and may partly explain why many genetic associations based on SNPs have not been confirmed in subsequent studies.

99

How many markers are necessary to infer correct familial relationships in follow-up studies?

S. Presciuttini(1), C. Toni(2), I. Spinetti(2), R. Domenici(2), J.E. Bailey-Wilson(1)
1) Inherited Disease Research Branch, NHGRI, NIH, Baltimore MD, USA2) Unit of Legal Medicine, University of Pisa, School of Medicine, Italy

In follow-up studies, genome-wide scan data are usually not available, and verifying relationships among relatives could only be obtained by genotyping of additional markers outside the candidate region. We investigated the number of markers (M) that are necessary to assign a given proportion of individual pairs (1-β=90%, 95%, and 99%) to their correct relationship at three predefined significance levels (α=1%, 0.1%, and 0.01%) against several alternative hypotheses. Five relationships were considered: parent-child, full sibs, half sibs, first cousins, and non-relatives. For each relationship, 10,000 true pairs were simulated, and the likelihood ratio that each pair was falsely attributed to each of the other relationships was calculated using both exact Bayesian probability equations and an approximate method based on marker

heterozygosity. Simulations were carried out separately for 25 unlinked markers commonly used in the forensic practice, and they were repeated a second time to reach a total number of 50 markers. We also investigated the reduction in the number of total genotypes achieved by using a sequential test, based on a first set of ten markers and the addition of other sets of five markers each, while keeping α and 1-β constant. Our analysis may be useful to determine the costs of performing pre-screening genotyping in second-phase studies with the only purpose of verifying reported relationships.

100

Score tests for detecting linkage to quantitative traits in selected samples

H. Putter, J. Lebrec, J. van Houwelingen
Dept. of Medical Statistics, Leiden University Medical Center, The Netherlands

It is well known that linkage studies for complex diseases achieve higher power when selecting pedigrees and individuals on the basis of their phenotypic values. How to analyze these selected samples is the subject of ongoing research. We propose an approach using score tests in a variance components model for quantitative traits, related to [1]. Inference is now based on an inverted model, where the distribution of identity-by-descent sharing is considered, conditional on the trait values. The approach can also be used for the optimal selection of individuals, as shown in another abstract at this meeting [2]. It is similar to the inverted regression of [3], but gives explicit expressions for simple designs like sib pairs. Extensions to dichotomous traits are also considered, using the concept of a liability threshold model. [1] Putter et al. Genet. Epidemiol. 22, 2002, pp. 345–355 [2] Lebrec et al. Abstract, 2003 IGES conference [3] Sham et al. Am. J. Hum. Genet. 71, 2002, pp. 238–253

101

Linkage Analyses Using Autoantibody Traits as Inter-mediate Phenotypes in Lupus Pedigrees

P.S. Ramos(1), J.A. Kelly(2), C.M. Meyer(1), A.N. Leiran(1), W.A. Ortmann(1), K.J. Espe(1), G.R. Bruner(2), J.M. Olson(3), J.B. Harley(2), K.L. Moser(1)
(1) Dept. Medicine, Univ. Minnesota, USA, (2) Oklahoma Medical Research Foundation, USA, (3) Case Western Reserve Univ.USA

Numerous studies have suggested that various autoimmune phenotypes may share underlying genetic components. Given the higher prevalence of autoimmune diseases and serologic abnormalities in family members of systemic lupus erythematosus (SLE) probands, we hypothesized that redefining phenotypes of individuals in SLE pedigrees with a propensity for autoimmunity may provide a useful intermediate phenotype for unraveling the complexity of SLE and possibly other autoimmune diseases. We have characterized autoantibody profiles of 1668 total subjects in 229 families multiplex for SLE. We used measures of association between siblings for each autoantibody to determine which specificities were most likely to aggregate in families and provide potentially useful information for subsequent linkage analyses. We identified evidence of familial aggregation for La/SSB ($p=0.001$), nRNP ($p=0.009$), and Sm ($p=0.04$). Genome wide linkage analyses were performed using both the revised Haseman-Elston algorithm (SIBPAL) and an affected relative pairs approach (LODPAL) to identify chromosomal regions of increased allele sharing. Significant evidence for linkage ($p<0.00005$) was found using La/SSB autoantibodies as the trait for 8 loci: 6p23, 10q23, 11p15, 12q24, 15q22, 16q23, 17q25, and 19p13. Several additional loci with suggestive evidence for linkage ($p<0.002$) were identified using various other autoantibody traits. These results demonstrate that using autoantibody traits as intermediate phenotypes may increase the power to detect linkage, and provide evidence of the presence and locations of genes that are involved in the development of lupus-related humoral autoimmunity.

102

Improved inference of missing parental data in a log-linear test of association by inclusion of unaffected siblings

E. Rampersaud, M.C. Speer, E.R. Martin
Center for Human Genetics, Duke University Medical Center, Durham, NC, USA.

Candidate gene studies remain a powerful approach for studying complex diseases, but rely heavily on tests of genetic association. Weinberg et al. (1998) have developed a log-linear model for family-based association studies as a flexible likelihood-based alternative to the transmission/disequilibrium test (TDT). Like the TDT, the method uses genotype data from family triads (affected offspring and both parents), but has the advantage that the model can be generalized easily to different causal scenarios since the likelihood framework allows for inclusion of parameters to test for imprinting, maternal effects, and environmental interactions. In the situation where parental genotypes are missing, the expectation maximization (EM) algorithm is used to allow incomplete triads to contribute to the likelihood ratio test. We present an extension to this model, which incorporates additional information from the genotypes of unaffected siblings to improve assignment of incompletely typed families to theoretical mating type categories, thereby improving inference of missing parental data. We use computer simulations to evaluate type I error and power of the extended model. These simulations demonstrate the validity of the extended log-linear model under the null hypothesis of no association. We compare the power of the extended model to that of the original model, under varying levels of missing data. We examine the impact of violations of the low penetrance assumption, which is necessary for inclusion of the unaffected sibling into the model. We conclude that the proposed log-linear model will be an important tool for future candidate gene studies, particularly in the area of birth defects where unaffected siblings often can be ascertained and where epigenetic

factors such as imprinting may play a role in disease etiology.

103

Mining disease-relevant genes from DNA microarray data by an ensemble decision approach

S.-Q. Rao(1), X. Li(2, 3), T.-W. Zhang (2), Z. Guo(2, 3), K. L. Moser(4), E. J. Topol(1), Q. Wang(1)
(1)Depts. of Cardiovascular Med and Molecular Cardiology, CCF, USA; (2)Dept. of Computer Sci, Harbin Inst of Tech, China; (3)Dept. of Math, Harbin Medical Univ.China; (4)Dept. of Med, Univ. of Minnesota, USA

The advent of the microarray technology has promised to provide new insights into the complex biological systems by monitoring activities of thousands of genes simultaneously. Current analyses of microarray data focus on classification of biological types, for example, tumor versus normal tissues. A future scientific challenge is to extract disease-relevant genes from the bewildering amount of raw data. This is one of the most critical themes in the post-genomic era, but it is generally ignored due to the lack of an efficient approach. In this paper, we developed a novel recursive-partition-tree-based ensemble method for gene extraction that can be tailored to fulfill multiple biological tasks including (1) precise classification of biological types; (2) disease gene mining; and (3) target-driven gene networking. We also gave a numerical application for (1) and (2) using a public microarray data set and discussed its application for (3).

104

A Comparison of Data Mining and Logistic Regression Approaches for Detecting Gene-Gene Interactions

Marylyn D. Ritchie(1), Chris S. Coffey(2), and Jason H. Moore(1)
(1)Program in Human Genetics, Vanderbilt University Medical School, USA, (2)UAB School of Public Health, USA

The detection of gene-gene ($G \times G$) interactions is both a statistical and a computational challenge. In response to this challenge, we have developed two novel data mining approaches. The first, multifactor dimensionality reduction (MDR), is a data reduction strategy while the second, genetic programming neural networks (GPNN), is based on pattern recognition. The goal of the present study is to compare the power of MDR and GPNN for detecting $G \times G$ interactions with that of stepwise logistic regression (SLR). We simulated data using 20 different $G \times G$ interaction models varying in rare allele frequency (0.2 or 0.4), heritability (0.5% to 3%), and the number of interacting loci (2 or 3). All models represent the extreme case of interactions in the absence of main effects. We estimated power as the number of times that each method identified the correct two or three functional SNPs for each model. We found that MDR and GPNN were able to identify the correct two or three functional SNPs in most models with at least 75% power while SLR had 0% power. When the interacting SNPs were explicitly modeled using LR the power was less than 75% for most two-locus models and less than 25% for all three-locus models. These results suggest that data mining strategies such as MDR and GPNN may provide a powerful alternative to LR.

105

Heritability of Urine Calcium in French-Canadian Families Ascertained for Kidney Stones

N.M. Roslin(1), J.C. Loredo-Osti(1, 2), J. Tessier(3), A. Bonnardeaux(3), K.Morgan(1, 2, 4)
(1)Research Institute of the McGill University Health Centre, Canada, (2)Department of Human Genetics, McGill University, Canada (3)Centre de recherche Guy-Bernier, Hôpital Maisonneuve-Rosemont, Université de Montréal, Canada, (4)Department of Medicine, McGill University, Canada

The prevalence of kidney stones in North America has been estimated to be about 5%, and the recurrence rate of episodes can be as high as 80%. One of the main risk factors for idiopathic kidney stone formation is hypercalciuria which tends to aggregate in families. 136 French-Canadian families were ascertained through probands with kidney stone episodes who attended nephrology clinics or lithotripsy units in Québec. The heritability of urine calcium was estimated by variance components analysis using a derivative-free restricted maximum likelihood procedure. A square root transformation of the amount of calcium excreted by an individual over a 24-hour period was used to normalize the trait. Eight statistical models were chosen based on combinations of the following co-variables: gender, age at exam, weight, body mass index, 24-hour urine sodium level, household effects, and the proband's untransformed urine calcium level. To minimize the effects of ascertainment bias on the heritability estimates, the probands were removed from the analysis. The heritability estimates ranged from 0.51 to 0.56, with standard errors of 0.14 to 0.16. Mixture analysis suggested the presence of a major gene. Formal segregation analysis is in progress.

106

SDF1 3'A is associated with breast feeding HIV-1 transmission and breast milk cell-free viral load

C Rousseau(1, 4), G John-Stewart(3, 4), R Nduati(6), B Richardson(2, 5), J Kreiss(3, 4), J Overbaugh(1, 2)
(1) Div of Human Bio, (2)Pub Hlth Sci, F Hutchinson Cancer Res Cntr, Depts. of (3)Med, (4)Epi, and (5)Biostat, U of WA, Seattle, and (6)Dept of Ped, U of Nairobi

A polymorphic allele (SDF1 3'A) in the 3' untranslated region of SDF1β has been previously shown to be associated with HIV-1 disease progression in adult males and postulated to increase the level of SDF1 expression. In vitro, SDF1 has been previously shown to block infection of virus that uses CXCR4 as its coreceptor and enhance replication of virus that uses CCR5 as its co-receptor. To examine the effect of SDF1 3'A on mother-to-child transmission, women participating in a randomized clinical trial of breastfeeding transmission were genotyped at the SDF1 3'A locus. Heterozygous women

were more likely to transmit via breastfeeding (RR 3.1, 95% CI 1.1–8.6) than women homozygous for the most common allele. To understand the mechanism of the SDF1 3′A association with breastfeeding transmission, cell-free and cell-associated breast milk viral load were measured. Heterozygous women had higher levels of cell-free virus (p=0.005) but not cell-associated virus (p=0.422) compared to homozygotes with the most common allele, suggesting a greater rate of viral replication among heterozygotes. This data support the hypothesis that the allele causes an increased expression of SDF1β, which enhances viral replication of strains that use CCR5, the strains that are most often transmitted. It is also possible that the allele is in linkage disequilibrium with another causative allele.

107

Longitudinal Analysis of Breast Milk HIV-1 Level and Association with Clinical Outcomes

C.M. Rousseau (1, 5), R. Nduati (2), B.A. Richardson (3, 4), J.K. Kreiss(5), J. Overbaugh(1, 4)
(1)Div of Human Bio, (4)Pub Health Sci, Fred Hutchinson Cancer Res Cntr, Seattle, WA, USA (2)Dept of Ped, Univ of Nairobi (3)Depts of Biostats, (5)Medicine, and Epi, Univ of WA, Seattle, USA

It is not known whether breast-feeding transmission of HIV-1 is caused by cell-free virus, cell-associated virus, or both. To examine whether these viral levels in breast milk are associated with mother-to-child transmission, cell-free (RNA) and cell-associated (DNA) HIV-1 was quantified from breast milk supernatant and cells collected throughout the first two years of lactation from 291 Kenyan women. The level of HIV-1 DNA was associated with the level of HIV-1 RNA in breast milk (R=0.251, p=0.002). The level of RNA and DNA were both significantly associated with maternal plasma viral load, CD4 T cell count, and the detection of cervical and vaginal secretions. For each 10 fold increase in either RNA or DNA level, the risk of transmission increased (OR 1.4, 95% CI 1.05–1.85; OR 1.5, 95% CI 1.83–2.20, respectively). The concentration of HIV-1 RNA was significantly higher in colostrum/early milk than in mature milk. The concentration of HIV-1 DNA per milliliter was estimated from the measured values, which were per million cells. Estimates of HIV-1 DNA concentrations were higher in colostrum/early milk that in mature milk. In conclusion, infants are exposed to the greatest concentration of infected cells and cell-free virus early after delivery, during the time when most transmission occurs. Thus, both cell-associated and cell-free virus are likely to be important for transmission.

108

Combined Regression of Offspring on Mid-Parent (ROMP) and Regression of Offspring on One Parent (ROOP) approach

M-H Roy-Gagnon (1, 2), AF Wilson (1)
(1) Genometrics Section, IDRB, NHGRI, NIH, Baltimore, MD; (2) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

The Regression of Offspring on Mid-Parent (ROMP) approach is an extension of the traditional linear regression of offspring on mid-parent used to estimate trait heritability. ROMP also provides a test of association and an estimate of the heritability attributable to a candidate locus. ROMP requires parent-offspring trios with phenotype data on both parents and their offspring and genotype data on the offspring. Combining a Regression of Offspring on One Parent (ROOP) approach to ROMP allows the use of trios where phenotype data is available on only one parent. Combined heritability estimates are obtained by weighting the ROMP and ROOP estimates by the inverse of their respective variance. The power of the combined ROMP-ROOP approach was investigated using 2000 replicates of simulated samples of 400 parent-offspring trios. The phenotype was determined in part by 3 loci with locus-specific heritabilities ranging from 0 to 50%. Missing phenotype on one parent was introduced in 20 and 40% of the trios. At the 1% level, the reduction in power due to missing data was on average 5 +/− 9% and 12 +/− 17% for missing rates of 20% and 40%, respectively. The reduction was worse for smaller effects and for total compared to locus-specific heritability. With 20% missing values ROMP-ROOP did not improve the power on average (5 +/− 4% reduction). ROMP-ROOP improved the power when the missing value rate was 40% (4 +/− 8% reduction). Imputation procedures were also investigated.

109

Capturing LD structure with a subset of SNPs

S.F. Saccone(1), P. Taillon-Miller(1), R.D. Miller(1), N.L. Saccone(1), P.-Y. Kwok(2), J.P. Rice(1)
(1) Washington University School of Medicine, St. Louis, MO, USA, (2) University of California, San Francisco, CA, USA

For association studies of complex human disease it is important to find subsets of single nucleotide polymorphisms (SNPs) that capture the key linkage disequilibrium (LD) information in a region. Proposed approaches include definitions of "LD blocks" or "haplotype blocks" from which "tag SNPs" can be selected. We describe an alternative approach using conditional LD. Given three SNPs A, B, and C, we define conditional measures of disequilibrium by computing the usual measures, such as D and D′, conditional on the genotype at the intervening marker B. Assuming that linkage disequilibrium is due solely to recombination, we can determine changes between generations in the $2 \times 2 \times 2$ table of frequencies and have constructed theoretical examples where the conditional measures can be computed algebraically. Given a set of SNPs, we study the process of selecting a minimal subset with moderate LD between each consecutive pair and minimal conditional LD, and evaluate how this subset captures the LD information in the region. We have completed extensive SNP genotyping for 94 CEPH males on chromosomes 5, 7, 17 and X. As an example, using our X chromosome data we examined a set of 17 SNPs spanning

a 400kb region and selected 5 SNPs where each SNP from the original set is in LD with some member of the subset, while the conditional D-prime in the subset is small. The same process can be applied to autosomes, where the conditioning is performed on three genotypes as opposed to two in the male X data.

110

New functional PON1 variant predicts prostate cancer

J.T. Salonen (1), M. Marchesani (2), A. Hakkarainen (1), J. Kaikkonen (2), E. Pukkala (3), P. Uimari (2), T.-P. Tuomainen (1)
(1) Research Institute of Public Health, University of Kuopio, Finland, (2) Oy Jurilab Ltd, Finland, (3) Finnish Cancer Registry, Finland

Prostate cancer is a common highly heritable neoplasm in men. The human serum paraoxonase (PON) is an antioxidative enzyme, which eliminates carcinogenic radicals. To find new mutations in the serum paraoxonase-coding gene, PON1, we applied phenotype-targeted hierarchical sequencing in a prospective random population sample of men. We studied the association of one new (I102V) and two previously known (M55L and Q192R) functional mutations with incident prostate cancer by Cox' modeling in cancer-free male participants of the KIHD cohort from Eastern Finland, who were followed by the National Cancer Registry of Finland for 9-14 years. We found a previously unknown single nucleotide polymorphism 304 A/G in codon 102 in exon 4 of the PON1 gene that produces an amino acid change of isoleucine (I) to valine (V). The V coding allele was associated with a decreased serum paraoxonase activity. Of the 1, 543 men, 54 (3.5 percent) were carriers of the PON1 I102V mutation. In a model adjusting for the strongest other predictors, the carriers of the 102V allele were at a 6.3-fold (95 percent confidence interval, 2.1 to 19.1) risk to develop prostate cancer during the follow up, as compared with non-carriers. These prospective data as well as confirmatory data suggest that carriers of the PON1 102V allele are at increased risk to develop prostate cancer. If our findings hold in other populations the mutation can be used in gene tests predicting prostate cancer.

111

Bias and power in unmatched analyses of association studies for gene-environment interaction

CL Saunders, DT Bishop, JH Barrett
Cancer Research UK Genetic Epidemiology Division, University of Leeds, Leeds, UK

Several variations on the population-based case-control design have been proposed for the study of genetic association and gene-environment interaction. These include the use of unaffected siblings as controls and the selection of cases who have an affected relative or two primary cancers. We discuss the assumptions that are made in the unmatched analysis of such designs and the bias in the estimates of the population interaction odds ratio, φ, that are obtained. For several designs the expected value of these estimates can be expressed in terms of φ and a term that describes the association between genotype and exposure among the siblings of cases, allowing the magnitude of bias to be quantified. The efficiency of a study to detect interaction depends on the size of the interaction estimate and its variance. Designs that result in an upwardly biased estimate of φ are potentially more efficient than a population-based case-control study, but this also depends on the variance of the estimate. If a frequency matching approach is taken in a population-based study, the frequency at which to sample exposure among controls to minimize this variance can be evaluated. If risk factors are independent, this frequency is 50%, but if genotype and exposure are associated then the most efficient frequency depends on the magnitude of the association and on the population genotype and exposure frequencies. We discuss how this result relates to the performance of the family and second-primary designs. In these designs the frequency of rare risk factors is implicitly raised among study subjects towards 50%, decreasing the variance of the estimate; however the increased risk factor associations among study subjects lead to downwardly biased estimates or increased variance in some cases. Power to detect interaction in unmatched analyses of these studies will be illustrated.

112

The –514 C>T polymorphism in Hepatic Lipase Gene and Plasma Lipids:A Meta-analysis

F.A. Sayed-Tabatabaei, A. Isaacs, O.T. Njajou, C.M. van Duijn
Genetic Epidemiology Unit, Department of Epidemiology & Biostatistics, Erasmus MC, Rotterdam

Hepatic lipase (HL) plays an important role in the metabolism of both pro- and anti-atherogenic lipoproteins. A polymorphism in the HL gene, designated alternately as the –480 C>T or the −514 C>T, has received considerable attention in the literature. The findings of association between this polymorphism and plasma lipids were contradicting. In this study we examine the association in a meta-analysis using published articles until June 2003. Among 33 identified articles, 20 met the inclusion criteria containing 11748 subjects, although this total fluctuates according to the data available for a given outcome. Significant differences between both the CT and the TT genotypes compared to the CC genotype occurred in HL activity (weighted mean difference (WMD) [95% CI] =−5.65 [−8.05, −3.25] and −10.76 [−14.13, −7.40] mmol/L/hr, respectively) as well as plasma HDL levels (WMD=0.03 [0.02, 0.05] and 0.11 [0.04, 0.17] mmol/L respectively). Analysis of other plasma lipids, including total cholesterol, LDL and triglycerides failed to reveal any significant differences. Stratification by gender, ethnicity, and risk category did not affect these findings. Funnel plots were quite symmetrical. In this meta-analysis, we found a significant association between presence of the T allele of –514 C>T polymorphism in hepatic lipase gene and HDL level and HL activity. Although the difference between the CT and TT groups was not significant the step-wise change in the point estimate suggests an allele dosage effect with this polymorphism.

113

Evidence of linkage on chromosome 15 for age-related cortical cataracts

J. Schick (1), S. Iyengar (1), K. Reading (1), R. Liptak (1), C. Millard(1), K. Lee (2), E. Moore (2), G. Jun(1), R. Klein(2), R. Elston (1), B. Klein (2)
(1) Dept. of Epi. & Biostat.Case Western Res. Univ.; (2) Dept. of Ophth. & Vis. Sci.Univ. of Wis. Med. School

Age-related cortical cataract is a common complex disorder with a multifactorial etiology. We performed the first genome-wide linkage scan for cortical cataracts in a selected sample from the Beaver Dam Eye Study, a longitudinal study evaluating risk factors for age related macular degeneration, lens opacities and visual impairment. The trait of interest was a quasi-continuous 12-category severity score determined using continuous gradings of duplicate Neitz photographs that included nine segments of the lens of the eye. We genotyped 353 autosomal markers in 325 participants (N=257 sib pairs). Second stage mapping was conducted in regions that demonstrated marginal evidence of linkage ($P \leq 0.01$). Multipoint linkage analysis was performed using Haseman and Elston regression as implemented in SIBPAL (S.A.G.E. 4.3). Prior to the linkage analysis, we adjusted our trait for age, age2, sex, age $\times$ sex interaction and vitamin usage. There was empirical evidence of linkage along two regions of chromosome 15 at a threshold significance level of 0.01 or less; one along an 18 cM region on chromosome 15q14 that includes markers ACTC (P=0.0090), D15S118 (P=0.0015), GATA50C03 (P=0.0031) and D15S1012 (P=0.0038) and another along a 4 cM region of 15q23 that contains marker D15S1000 (P=0.0081). Our analysis indicates that one or more genes for development of cortical cataracts is located on chromosome 15q.

114

Empirical Comparison of A General Approach for Analyzing Sibling-Pair Data

F.R. Schumacher(1), D.V. Conti(2), J.S. Witte(3)
(1)Dept. of Epi & Biostat, CWRU, USA, (2)Dept. of Prev. Med. USC, USA, (3)Dept. of Epi & Biostat, UCSF, USA

Sibling-based designs are commonly used for both linkage and association studies. In particular, concordantly affected sib-pairs are used for evaluating linkage, whereas discordant (i.e. case-control) pairs are used to investigate association. When a study recruits a combination of concordant and discordant pairs, however, conventional linkage and association analyses do not optimally use all information on the subjects. For example, evaluating association with conditional logistic regression (CLR) will ignore concordantly affected sib-pairs. To address this problem one can use a general approach that separates genotype information into within- and between-family (WB) components (Fulkner, 1999;Abecasis, 2000). We present the general theory underlying this approach, and illustrate its potential value with an application to association data from a study of candidate genes and prostate cancer. Specifically, we compare results from using the WB and CLR approach. Across the two methods of analysis the results for several candidate genes differed substantially. For example, the WB approach yielded an odds ratio=2.14 (p-value=0.07) for a SNP in the candidate gene CYP3A4, whereas using CLR gave an OR=1.70 (p=0.26). However, the reverse situation was seen for other SNPs. Since the potential improved efficiency of this approach may change drastically depending on the analysis, additional work will more fully compare (e.g. by simulation) this approach to conventional methods. Fulker D.W, et al. AJHG 64(1999):259–67; Abecasis G.R.et al. AJHG 66(2000):279–92

115

Ascertainment bias and its effect on estimating gene-environment interaction

K.D. Siegmund, W.J. Gauderman
Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA

Siegmund and Langholz (Am J Epidemiol 2002; 155:875–80) showed that in family-based case-control designs, the use of unaffected siblings that live outside the ascertainment region of the case as controls, can lead to biased estimates of environmental main effects. Gene-environment interaction effects remain unbiased if the environmental factor (E) is uncorrelated with genotype (G). Since the case-only design requires this same gene-environment independence assumption, one might ask, are family-based controls needed? We compare the bias and mean squared error (MSE) of the case-sib control and case-only design in the presence of gene-exposure correlation. Data are simulated using the model from Albert et al. (Am J Epidemiol 2001; 154:687–93), logit $P(D=1)=-3.7+0.054G+1.13E \quad -0.815 \quad GE$. When environmental exposure is independent of genotype, both designs give unbiased estimates and the case-only design has more than twice the efficiency of the case-sib design. When there is correlation between gene and exposure, the case-only design is biased whereas the case-sib control design is unbiased provided there is no ascertainment bias. When ascertainment bias in the selection of controls is added (exposure frequency is 20% higher in the case-ascertainment region than in the control-ascertainment region), both designs are biased however the bias and MSE are lower in the case-sib design. This suggests that ascertainment bias in the case-sib design results in less bias in $G \times E$ interaction estimates than genotype-exposure correlation in the case-only design.

116

Prevalence of Hypertension and Family Aggregation of Blood Pressure in Middle DalmatiaCroatia

T. Skaric-Juric, N. Smolej Narancic
Institute for Anthropological Research, Zagreb, Croatia

Isolated, particularly island populations gain an increased interest in genetic studies because of expectedly reduced number of potential genes having a pronounced role in the

determination of complex traits. The populations inhabiting Middle Dalmatian islands are relatively isolated and inbred populations sharing a rather homogenous environment (climate, nutrition, life style, etc.), which makes them a very promising target for genetic investigations. The analysis of blood pressure family data encompassing 1,040 individuals, inhabitants of the islands of Braè, Hvar, Korèula and the Pelješac peninsula (Middle Dalmatia, Croatia) have been performed. The results suggested higher heritability of diastolic blood pressure compared to the systolic: h^2_{DIA}=44%, h^2_{SYS}=24% (age & sex adjusted data), and the same relation remained after adjustment for morphology (h^2_{DIA}=29%, h^2_{SYS}=13%) and additionally for smoking habits (h^2_{DIA}=25%, h^2_{SYS}=10%). The obtained results combined with the information on high prevalence of developed hypertension (34.7%) observed in the four investigated Middle Dalmatian island/peninsular populations, especially with the phenomenon of rather high prevalence of isolated diastolic hypertension (15.3%) with a trend to be increasingly higher as the island gets more distant from the mainland (and consequently, more isolated and more inbred) point to the adequacy of further genetic investigations of blood pressure in those populations.

117

Application of a longitudinal mixed effects model to transmission disequilibrium testing.

M.K. Sontag(1,2), J.E. Hokanson(2), J.A. Marshall(2), M. Corey(3), G.O. Zerbe(2), F.J. Accurso(1)
(1)Pediatrics, Univ. of Colorado, Denver, CO. (2) Prev. Med. & Biometrics, Univ. of Colorado, Denver, CO, USA. (3)Hospital for Sick Children, Toronto, ON, CAN.

Longitudinal, quantitative data analysis within the framework of a traditional transmission disequilibrium test (TDT) is cumbersome, requiring initial estimation of parameters to be used as a quantitative trait. We propose a model that incorporates longitudinal observations, the mating type of the family, and the resulting genotype of the child, to avoid the multiple models and to allow for simultaneous estimation of all of the parameters. This model was tested in trios comprised of children with cystic fibrosis (CF) and their parents, using markers on the CFM1 region on the long arm of chromosome 19 (q13.2), investigating modifiers of pancreatic disease. Longitudinal immunoreactive trypsinogen (IRT) levels from 288 children were modeled and patients were divided into three groups: rapid, moderate, and slow decliners. Genotyping in the CFM1 region is complete on 37 children with CF and both parents. In rapid decliners (n=7), TDT indicated linkage to D19S217 (chi-square=3.57, p=0.06). Using longitudinal mixed effects model analysis that incorporates family structure with simultaneous parameter estimates and covariate adjustment we obtained similar results (p=0.05). We conclude that mixed effects models provide an efficient tool for incorporating longitudinal data into a TDT model in a parallel manner.

118

Recombination rate variation and the description of genetic diversity

M.P.H. Stumpf(1), G.A.T McVean(2), E. DeSilva(1), S. Myers (2)
(1) Dept. of Biology, University College London, UK, (2) Dept. of Statistics University of Oxford, UK

We estimate recombination rates, their local variation and the density of recombination hotspots in the human genome. Two estimators, a composite likelihood approach that uses genotypic data without having to infer haplotypes, and a non-parametric estimate for the minimum number of local recombination events are applied to the data of Gabriel et al. (Science, 296(2002) 2225–2229). We find considerable variation in the local recombination rate between and within the different regions but high levels of concordance for estimates obtained from different populations (correlation of recombination rates between populations: 0.85–0.95). There is also excellent agreement between the different estimators. Estimated recombination rates appear to give a much more consistent description of the linkage disequilibrium (LD) structure than do the simple summary statistics is use today. We also provide evidence that demographic uncertainty introduces no or only little bias into our estimators. A detailed study of local recombination rate variation detects several features that can be described as recombination hotspots. Comparison with the data of Jeffreys et al. (Nat.Gen.29 (2001), 217–222) allows us to estimate the number of recombination hotspots that can be detected at the average marker density (approximately 1/5kb) in the data. Our conservative estimate yields approximately 15 000–20 000 hotspots/hotspot clusters in the human genome. We show that this number is likely to go up as denser marker sets are considered. Using estimated recombination rates it is possible to assess the origin of haplotype blocks. We find that often blocks appear to be due to stochasticity in the recombination process and do not show a clear correspondence with regions of low recombination rate. Extensive simulations and empirical analysis show how the inferred recombination structure of the human genome is likely to affect the performance of tagging approaches.

119

Sequential testing methods for pedigree error detection based on genome-screen data

L. Sun (1, 2), R.V. Craiu (3)
(1) Dept. of Public Health Sciences, University of Toronto, Canada, (2) Programs in Genetics and Genomics, Hospital for Sick Children, Canada, (3) Dept. of Statistics, University of Toronto, Canada

Detection of pedigree errors via inference of pairwise relationship among individuals is now a standard practice prior to genetic mapping studies. However, there are often thousands (pairs within families, e.g. Sun et al. 2001) or

millions (pairs across families, e.g. Epstein et al. 2000) of pairs to be tested in a typical data set, and the problem of multiple comparisons is crucial. While, one can apply the conservative Bonferonni correction, the prior expectation is that only a small number of pairs may be misspecified, thus by correcting for all the pairs, one unnecessarily trades the power for controlling the overall false positive rate. The sequential multiple decision procedure (SMDP) discussed in Province (2000) in the context of gene-wide linkage scans is developed for detection of pedigree errors. The method exploits genome-screen data sequentially until a set of likely mis-specified pairs separates themselves from the remaining pairs. The method essentially uses part of the available data to exclude a majority of the pairs, which are expected to be background noises, from the final hypothesis tests performed, while it simultaneously controls both type I and type II errors. We perform simulation studies to demonstrate the efficiency of the proposed method.

120

Automated Detection of Informative Combined Effects in Genetic Association Studies of Complex Traits

N. Tahri-Daizadeh(1, 2), D.A. Tregouet(1), V. Nicaud(1), N. Manuel(1), F. Cambien(1), L. Tiret(1)
(1) INSERM U525, Faculté de Médecine, Hôpital Pitié-Salpêtrière, 91 Bld de l'Hôpital, 75634 Paris, France, (2) Genset-Serono Group, RN7, 91030 Evry, France

There is a growing body of evidence suggesting that the relationships between gene variability and common disease are more complex than initially thought and require the exploration of the whole polymorphism of candidate genes as well as several genes belonging to biological pathways. When the number of polymorphisms is relatively large and the structure of the relationships among them complex, the use of data-mining tools to extract the relevant information is a necessity. Here, we propose an automated method for the Detection of Informative Combined Effects (DICE) among several polymorphisms (and non-genetic covariates) within the framework of association studies. The algorithm combines the advantages of regressive approaches with those of data exploration tools. Importantly, DICE considers the problem of interactions between polymorphisms as an effect of interest in itself and not as a nuisance effect. We illustrate the method with three applications on the relationship between (1) the P-selectin gene and myocardial infarction, (2) the Cholesteryl Ester Transfer Protein gene and plasma high-density-lipoprotein cholesterol concentration, and (3) genes of the Renin-Angiotensin-Aldosterone system and myocardial infarction. The applications demonstrated that the method was able to recover results already found using other approaches, but in addition detected biologically sensible effects not previously described.

121

The Muscarinic Cholinergic Receptor Gene: Single Nucleotide Polymorphisms in the Chinese Population, Linkage Disequilibrium Analysis and Association with Severe Myopia

Tan EC, Ng SH, Yap SH, Karupathivan U, Teo YY, Wu HM, Yap EPH
Defence Medical Research Institute, Defence Science & Technology Agency, Singapore

The muscarinic system is a major component of the cholinergic neurotransmission system. Although an excellent candidate gene for cognitive processes, identification of a possible role of muscarinic receptor genes in psychiatric disorders has been hampered by a lack of reported polymorphisms and genotyping assays in the published literature. In this study, we have developed simple PCR-RFLP assays for the M1 receptor gene (CHRM1) and established the frequencies of three polymorphisms in control individuals and those with severe myopia in a Chinese population. Atropine and pirenzepine (both antagonists of muscarinic receptors) have been used successfully in clinical trials and animal studies for the treatment of myopia. It is thus postulated that dysregulation in the muscarinic cholinergic pathway may play a role in the pathogenesis of severe myopia the most common ocular disorder worldwide. This study sought to establish presence of single nucleotide polymorphisms (SNPs) in the muscarinic cholinergic receptor 1 (CHRM1) gene in the local Chinese population and investigate possible association of these SNPs with myopia. Nine SNPs in both the coding and untranslated regions were investigated. Only 5 (3 synonymous changes and 2 in UTR) were found to be present in the local Chinese population. The remaining 4 were not found in the samples assayed (number of chromosomes=40). Controls (n=152) and 220 cases with severe myopia (refractive errors < -8.00 diopters) were genotyped for all 5 markers. Marginally statistically significant difference was obtained for the –9697G > A polymorphism in the 5′ UTR although haplotype analysis did not increase the level of significance. There was also no significant linkage disequilibrium between marker pairs. Further investigations will be carried out to evaluate other polymorphisms in the M1 receptor gene and the association with severe myopia.

122

Graphical modeling of the joint distribution of alleles at associated loci. HapGraphs not HapMaps.

A Thomas (1), N J Camp (1)
(1) Genetic Epidemiology, 391 Chipeta Way Suite D, Salt Lake City, UT 84108, USA

Pairwise linkage disequilibrium haplotype blocks and hot spots provide only a partial description of the patterns of dependences and independences between the allelic states at proximal loci. On the gross scale, where recombination and spatial relationships dominate the associations can be reasonably described in these terms. However, on the fine scale of current

polymorphism maps the mutation process is important and creates associations which are independent of the physical ordering and which can not be summarized with pairwise measures of association. Graphical modeling provides a standard statistical framework for characterizing precisely this sort of complex stochastic data. While graphical models are often used in situations where assumptions lead naturally to specific models, it is less well known that estimation of graphical models is also a developed field. In this presentation decomposable graphical models are fitted to data from 25 SNPs, one triallelic, in the ELAC2 gene. The objective function is the maximized log likelihood for the model penalized by a multiple of the model's degrees of freedom. Simulated annealing is used to find good solutions. The results show clear non spatial and non pairwise dependences. The great potential of this approach is that categorical phenotypes can be included in the same analysis and association with polymorphisms assessed jointly with the inter locus associations. This is illustrated in the above example with phenotypic data on sex and incidence of prostate cancer.

123

Using family history information in genetic association studies

D. Thompson(1, 2), D.E. Goldgar(1)
(1)Unit Genetic Epi, IARC, France (2)CR-UK Genetic Epi Unit, Cambridge, UK

Association studies assessing the relationship between a common polymorphism and disease generally compare allele frequencies in cases and controls. In such studies, a limited amount of information is often available about disease incidence in relatives. We hypothesized that more power could be obtained by incorporating the constraints imposed by the properties of a genetic polymorphism, and that power could be further increased by using family history (FH) information. We have developed a simple method for incorporating basic FH information from cases and controls into a genetic association study. We model the likelihood of the data in terms of the allele frequency and its relative risk (RR) of disease. Likelihood ratio tests are performed and maximum likelihood estimates of the parameters are obtained. Using simulations, we compared the power to detect an association using this approach with that of a 2×2 Chi-squared test, for a range of allele frequencies, RRs and familial RRs. We also considered the effect of similarly modeling the likelihood without stratifying by FH. The sample size required to detect an association at a 1% significance level is consistently lower when stratifying by FH. The reduction is clearest for a higher familial RR, a lower allele frequency, and a higher carrier RR. Stratifying by FH improves the precision of the RR estimates, although for an allele of frequency 0.5% both estimates are less stable. In situations where basic FH data are already available this study shows that efficiency can be improved by the inclusion of even this small amount of extra information.

124

Modeling Heterogeneity in the Integrated Meta-Analysis of Gene-Disease and Gene-Phenotype Association Studies

J.R. Thompson, C. Minelli, M.D. Tobin, K.R. Abrams, P. Burton
Genetic Epidemiology Unit, Centre for Biostatistics, University of Leicester, UK

An integrated random effects meta-analysis of the pathway from gene to intermediate phenotype to disease can combine information from studies of gene-disease and studies of gene-phenotype to create an indirect estimate of the phenotype-disease association. Because of mendelian randomization, this indirect estimate should be unconfounded and free from the effects of reverse causation, and so in these respects is preferable to direct epidemiological measurement. Individual studies that report data for both gene-disease and gene-phenotype may show correlation in the between-study heterogeneities of these two measures. Using simulation studies we show that even large meta-analyses often have insufficient information to estimate this correlation, but that the size of the correlation is critical to the indirect estimate of the phenotype-disease association. We propose an alternative model in which the heterogeneities of the gene-phenotype and phenotype-disease associations are modeled as being independent. This model still induces a correlation between the heterogeneities of gene-phenotype and gene-disease but now it can be estimated by maximum likelihood or Bayesian methods. This approach is applied to the integrated meta-analysis of the MTHFR-homocysteine-coronary heart disease pathway.

125

Power Study of a Novel Approach to Identifying a Minimum Candidate Gene Region in Complex Diseases

Thornton TA (1, 3), Kenealy SJ (2, 3), Haines JL (2, 3)
(1) Center for Integrative and Cognitive Neuroscience, Department of Biomedical Informatics, Vanderbilt University, (2)Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, (3)Program in Human Genetics, Vanderbilt University Medical Center, USA

Genetic heterogeneity is a major confounding factor in the statistical analysis of complex human diseases. In linkage analysis, locus heterogeneity reduces the power to detect a true signal originating from a subset of families. A novel approach to this challenge, consensus haplotyping, was proposed by Hutcheson et al. (Am J Med Gen 117B: 90–96, 2003) in which locational mapping of candidate genes was combined with recombination breakpoint analysis. We have performed simulation studies to evaluate the power of this approach. We simulated nuclear families with three children under a recessive disease model with 50% locus heterogeneity. A sample of 100 families was selected in which both parents were unaffected and at least two children were affected.

Ten bi-alleleic markers were simulated 10cM apart along with one disease locus, which was midway between two markers. A second disease locus was simulated to have no linkage to the first disease locus or any other marker. We performed linkage analysis on the ten markers, identified the marker with the highest lod score, and selected those families that showed linkage to any marker within 20cM of this marker. We then performed haplotype analysis and identified which loci were shared on both haplotypes among the affected children. Our analyses found that this approach had greater than 80% power to correctly localize the gene within a 40cM region and 60% power within a 30cM region. The simulation is being expanded to use more finely mapped markers, multiple genetic models, and various sample sizes. These results suggest that consensus haplotyping is a powerful approach toward localizing a complex disease gene given locus heterogeneity.

126

Comparisons of two-stage TDT type design with genomic control and structural association approaches using Simulations

Tiwari HK(1), George V(1), Laud P(2), Beasley TM(1)
(1)University of Alabama at Birmingham, Birmingham, AL, USA, (2)Medical College of Wisconsin, Milwaukee, WI, USA

Concerns for spurious significance due to population stratification led to the development of the family-based transmission/disequilibrium test (TDT) design. The TDT tests for linkage disequilibrium by comparing the proportion of alleles transmitted vs. not transmitted from informative parental matings to affected offspring. TDT type designs are effective ways of eliminating false positives due to stratification and admixture, but at a high cost in terms of power and efficiency. There may be substantial loss of power due to the fact that all individuals who are uninformative for allelic transmission status are excluded from the analysis, effectively utilizing only a substantially smaller part of the sample compared to a case-control design. The TDT design for linkage testing has power only in the presence of allelic association. It should be advantageous to design a two-stage procedure. In the first stage, we test for population association using all individuals in the sample. If a statistically significant association is found, then using one of the appropriate TDT type designs, we can perform a family based association study to effectively evaluate linkage disequilibrium. There are mainly two types of methods to control for stratification has been proposed in case-control design, termed as genomic control (GC) and structural association (SA). There has never been thorough investigation of comparisons between TDT type methods with genomic control or structural association methods. In this study, we will compare the two-stage TDT design based on George et al. (1999) regression method for quantitative traits with various GC/SA approaches in case-control designs, with respect to power and type I error using extensive simulations.

127

Analysis of continuous traits affected by treatment in genetic studies

M.D. Tobin, N.A. Sheehan, J.R. Thompson, P.R. Burton
Genetic Epidemiology Unit, Centre for Biostatistics, University of Leicester, UK

Studies of the genetic determinants of continuous intermediate traits offer a powerful insight into the aetiological architecture of complex diseases. However, for many such traits, measurements are subject to the effect of treatment that effectively right censors the relevant observations and estimates of genetic and environmental effects can be seriously distorted. Some adjustment for these effects is needed, but many common approaches to such adjustment do little or can make matters worse. We compare methods to adjust for treatment in studies of genetic and environmental determinants of blood pressure in population-based association studies, using simulated datasets and cross-sectional data from the Speedwell study. We demonstrate the extent of bias and loss of power when failing to adjust appropriately and show that censored normal regression provides a straightforward alternative to adding a constant when prior knowledge of average treatment effects is missing. Such models can allow for treatment in the estimation of genetic and environmental components of variance as well as fixed effects in cross-sectional and longitudinal family studies using extensions of our standard Gibbs sampling-based variance component models (eg Scurrah et al, 2000). We illustrate how censored normal models can be adapted to assume informative or non-informative censoring depending on the trait and population studied. These methods have wide applicability to any study of a continuous trait mitigated by treatment. Scurrah KJ, Palmer LJ, Burton PR. Genet Epidemiol 2000; 19:127-141.

128

Haplotype analysis of IL18 gene polymorphisms in relation to coronary artery disease and serum levels of IL18. The AtheroGene Study

D.A. Tregouet(1), S. Blankenberg(2), T. Godefroy(1), C. Bickel(2), H.J. Ruprecht(2), L. Tiret(1)
(1) INSERM U525, Paris, France. (2) Johannes Gutenberg University, Department of Medicine II, Mainz, Germany

Interleukin (IL)-18 plays a central role in orchestrating the cytokine cascade and is therefore suspected to play a key role in atherosclerosis. The objective of this work was to investigate the relative contribution of IL18 gene polymorphisms to the variability of serum levels of IL18 as well as their relation to coronary artery disease(CAD). For this purpose, the coding and regulatory sequences of the IL18 gene were screened and 11 polymorphisms were identified, 6 in the promoter (T-1529C, G-887T, C-838A, G-368C, C-105T and T–119C), 1 in the coding, S35S (A/C), and 4 in the 3'(C+167G, A+183G, G+230A and T+533C) regions. 4 polymorphisms (G-887T, C-105T, S35S (A/C) and A+183G) were then genotyped in the AtheroGene study composed of 1085 CAD patients and 469 controls.

Haplotype analyses revealed that the A+183G polymorphism was the only polymorphism associated with serum IL18 levels. The G183 allele was associated with decreased levels (68.5 vs 61.2 vs 55.7; R2=2%; p <10–4, in AA, AG and GG carriers respectively), consistently in cases and controls. Haplotype frequencies distribution was different between cases and controls (p=0.006) independently of IL18 levels. In particular, the –105C/35C haplotype was less frequent in cases than in controls (0.026 vs 0.053; p<10–4). In conclusion, these analyses suggested a role of IL18 gene polymorphisms in CAD risk and in the variability of IL18 serum levels.

129

Population mikro-isolates and the genome project in South Tyrol

F.D. Vogl, I. Pichler, G.K. Pinggera, P.P. Pramstaller
GenNova, Dept. of Genetic Medicine, European Academy Bolzano, Italy

Small founder populations with long-lasting isolation have esteemed qualities for gene-mapping studies. Due to their geographic situation in the Alps and their socio-cultural heritage, the German-speaking population of the autonomous province of South Tyrol (Northern Italy) can be considered a population isolate. This is strengthened by the presence of recurrent founder mutations. We present an interdisciplinary and multi-step strategy to study the genetics of frequent diseases. First, criteria were prepared to define suitable sub-populations ("mikro-isolates") within the province. Subsequently, appropriate villages were selected based on historical information and local demography. Church records served to reconstruct complete genealogies of current inhabitants, spanning at least 16 generations. Voluntary participants underwent a detailed interview and physical screening examination. Blood was sampled for DNA extraction. The total population of South Tyrol is approximately 460,000. We identified several villages in remote side valleys that qualified as mikro-isolates according to our criteria. In Stilfs (1,300 inhabitants) we enrolled half of the population above 18 years of age. Endogamy, calculated as percentage of marriages between individuals from the same village, was above 80% until 1950. In several large families, clusters of diseases (RLShypercholesterolemia) have been observed which are currently subject to genetic analysis. Focusing on mikro-isolates with known genealogies is a promising strategy to ascertain extended pedigrees ensuring a homogeneous environment.

130

Self-reported race: Having more than one choice

D.K. Wagener (1), J.D. Parker(2)
(1) Epidemiology and Public Health, RTI International, USA, (2) National Center for Health Statistics, CDC, USA

In 1997 the government redefined standards for the collection of Federal data on race and ethnicity, allowing for more racial categories and the designation of more than one race. Most epidemiologic research continues to use only single race classification for subjects. Concern has been previously expressed about the possible heterogeneity of people with single race categories. The availability of data from the National Health Interview Survey, which has collected self-reported information since 1982 on multiple race choices and ''best representative race,'' enables the estimation of the effect of forced single race choice on heterogeneity in the population groups. The percent of the population claiming multiple-race status increases for the smaller racial groups, e.g. American Indian or Alaska Native and Asian or Pacific Islander. This varies substantially by state. It is difficult to predict which race will be chosen as the primary race when multiple races are designated. For instance, American Indian/White people tend to designate White as the primary race, whereas Black/White tend to designate Black. Further, this varies by State and county. The consequence is that people indicating multiple race backgrounds, when forced to designate a single race, make different choices, in part based on their community. Single race and ethnicity classifications, as social concepts, encompass biologically heterogeneous populations.

131

A Genome-Wide Scan for Carotid Intima-Media Thickness in Mexican-American Families with a Coronary Artery Disease Proband

D. Wang(1), H. Yang(1), M.J. Quiñones(2), I. Enriquez(2), X. Jimenez(2), G. Hernandez(2), R. De La Rosa(2), Y. Li(3), W.A. Hsueh(2), H.N. Hodis(3), J.I. Rotter(1)
(1) Cedars-Sinai Medical Center, Los Angeles, CA, USA, (2) UCLA Medical Center, Los Angeles, CA, USA, (3) USC, Los Angeles, CA, USA

Carotid intima-media thickness (IMT) is a sub-clinical measure of atherosclerosis. Increased carotid intima-media thickness is associated with increased risk of coronary artery disease (CAD) and is an important predictor of CAD. To identify the genetic determinants of IMT, we performed a genome-wide scan using data from 91 two-generation Mexican-American families ascertained via a parent diagnosed with CAD. IMT was measured in 268 adult offspring using ultrasound. 407 adult offspring and their parents were genotyped using Marshfield screen set 12 (408 microsatellite markers at ~10cM interval). Heritability estimates and multipoint linkage analysis was performed using the multi-point variance component method implemented in SOLAR. Heritability estimate for this phenotype was 0.60 (p<1E-5). The strongest evidence for linkage was found on chromosome 13 at 99cM with a LOD score of 2.9. Other suggestive linkages were found on chromosome 6 (LOD=1.8 at 56cM) and 18 (LOD=1.5 at 116cM). After adjusting for the age, gender, and BMI, we still observed the strongest linkage (LOD=1.9) on chromosome 13 at 94cM. These results indicate a strong genetic effect for IMT in Mexican-American CAD families. The consistency of the linkage location on chromosome 13 before and after covariate adjustment suggests there is a locus at this location influencing IMT.

132
On asymptotic properties of affected-sib-pair linkage tests
K. Wang
Department of Biostatistics Division of Statistical Genetics, University of Iowa, USA

One popular ''model free'' statistical method for analyzing affected-sib-pair data is the ''possible triangle'' method, which maximizes the likelihood function within a triangle for the probabilities of identical by descent excluding points that are not biologically meaningful. However, part of the ''possible triangle'' corresponds to traits with low, including 0, prevalence and is hence not biologically interesting. We illustrate this point by considering a single-locus trait model. Our results are the following. Considering the disease allele frequency, penetrances of three genotypes with respect to the disease allele and the recombination fraction as model parameters, we show that 1) Subject to the constraint that the total trait variance is not smaller than any given positive value, the likelihood ratio statistic is asymptotically equivalent to some ''model-free'' methods, including mean test and two-allele test. One consequence of this result is that ''model-based'' methods, regardless of the specified trait model, are all asymptotically equivalent to each other and all are asymptotically equivalent to the mean test and the two-allele test; 2) The likelihood ratio test with fixed prevalence is asymptotically equivalent to the likelihood ratio test with maximized prevalence, where the prevalence is maximized over a set in which the prevalence is not smaller than any given positive value. This seemingly counter-intuitive result is due to the rather limited degrees of freedom for affected-sib-pair data.

133
Using trait data and marker data simultaneously: QTL mapping adaptive to the extent of selection
K. Wang
Department of Biostatistics Division of Statistical Genetics, University of Iowa, USA

For selected samples, there are two sources of linkage information, one is the trait data and the other one is the marker data. Traditional statistical methods for mapping quantitative trait loci typically use only one source of linkage information but not the other, or use both sources through the use of ad hoc weights. In addition, existing methods are mostly devised for selected sib-pairs and do not take the extent of selection into consideration. A theoretical framework is proposed for general pedigrees selected randomly through a phenotype-dependent-only criterion. This approach models the marker data and the trait data through common parameters and thus eliminates the need for weights to combine the linkage information from two sources. It explicitly takes the extent of selection into consideration and provides a unified analysis of population samples and selected samples. Score tests and their asymptotic distributions are introduced. Simulation studies were performed using nuclear families with discordant sib-pairs. The type I error rate and the power of this method and some existing methods are reported. This work generalizes Wang (2002, Hum Hered, 54:57-68) from population samples to selected samples. The current report considers a single-locus trait model with a single phenotype. However, it is straightforward to consider two-locus epistasis trait models or bivariate phenotypes that may or may not be involved in the selection criterion, given the work of Wang (2003, Hum Hered in press) or the work of Wang (2003, Genet Epidemiol in press).

134
Locus heterogeneity models for quantitative traits and related test statistics
K. Wang(1), Y. Peng(2)
(1)Department of Biostatistics Division of Statistical Genetics, University of Iowa, USA, (2)Department of Mathematics and Statistics, Memorial University of Newfoundland, Canada

Locus heterogeneity is believed to be common for complex traits but is seldom studied for quantitative traits. We propose a locus heterogeneity model (with one variant) for quantitative traits. This model generalizes the locus heterogeneity model of Smith (1961, Proc Sec Int Congr Hum Genet 1:212–213) from dichotomous traits to continuous traits. Several test statistics parallel to those in Lemdani and Pons (1995, Biometrics 51:1033–1041), Chiu et al. (2002, Biostat 2:195–211) and Shoukri and Lathrop (1993, Biometrics 49:151–161) are considered for this model for testing the null hypothesis that there is no linkage. The type I error rate and the power of these statistics are assessed through simulation studies.

135
Sample Size to Detect Gene-Gene Interactions Using Association Designs
S. Wang(1), H. Zhao (1,2)
(1)Dept. of Epidemiology and Public Health, Yale Univ. USA, (2)Dept. of Genetics, Yale Univ. USA

It is likely that many complex diseases are the results of the interactions among many genes as well as environmental factors. The presence of such interactions poses challenges to identify susceptibility genes, to understand biological pathways, and to predict and control disease risks. Recently, Gauderman (1) reported the first systematic study on the statistical power to detect gene-gene interactions in association studies. However, different statistical models were used to model disease risks for different study designs, and very low disease prevalence was assumed to make different models more comparable. In this article, assuming a logistic model for disease risk for different study designs we investigate the power of population-based association designs and family-based association designs for the detection of gene-gene interactions for common diseases. Our results

indicate that population-based designs are more powerful than family-based designs to detect gene-gene interactions when the disease prevalence is moderate in the study population.

136

Genome-wide linkage analysis identifies novel susceptibility loci for premature myocardial infarction

Q. Wang(1)*, S.-Q. Rao(1)*, G.-Q. Shen(1), L. Li(1), K. Newby(3), W. J. Roger(4), R. Cannata(1), E. Zirzow(1), R. C. Elston(2), E. J. Topol(1)
(1)Depts. of Cardiovascular Med and Molecular Cardiology, CCF, USA; (2)Dept. of Epi & Biostat, CWRU, USA; (3)Duke Univ. Med Ctr, USA; (4)The Univ. of Alabama Med Ctr, USA. *Contributed equally.

Atherosclerotic coronary artery disease (CAD) and acute myocardial infarction (MI) are thought to have a polygenic basis with a complex interaction with environmental factors. We recruited 428 multiplex families with premature CAD and MI consisting of 2030 individuals – 712 with MI, 974 with CAD and the age of onset was 44.4 ± 9.7 yrs. Genotyping was performed at the NHLBI mammalian facility using 408 markers that span the entire human genome every 10 cM. Linkage analysis was performed with the modified Haseman-Elston regression models and SIBPAL program. Three genome wide scans were conducted for single point, multipoint, and multipoint Caucasian only analysis (92% of cohort). Of eight novel susceptibility loci detected for MI, the most significant linkage was found to chromosome l (multipoint allele-sharing P values of $-\log(P) > 12$, or log of the odds ratio (LOD) score of 11.98); linkages to chromosomes 2, 4, 5, 7, 12, 13, 14 meet the criteria for genome-wide significance ($-\log(P)=4.66$ or LOD score of 3.6). For the less restrictive phenotype of CAD, no genetic loci were detected. Our data suggest that CAD and MI are genetically distinct disorders. This study thus identifies novel genetic susceptibility loci for MI, and provides a framework for the ultimate cloning of genes for premature, familial MI.

137

A locus on chromosome 13 influences levels of TAFI antigen in Mexican Americans

D. M. Warren(1), S. Cole(1), J.M. Soria(2), J.C. Souto(2), J. Hixson(3), J. Fontcuberta(2), J. Blangero(1), L. Almasy(1)
(1)Southwest Foundation for Biomedical Research, San Antonio, TX, USA (2)Hospital de la Santa Creu i Sant Pau, Barcelona, Spain, (3)University of Texas Health Science Center, Houston, TX, USA

Thrombin activable fibrinolysis inhibitor (TAFI) is a zymogen whose activated form inhibits fibrinolysis by modifying fibrin, depressing its plasminogen binding potential. TAFI levels have been associated with diseases including coronary artery disease, ischemic stroke, thrombosis, and type 2 diabetes. We examined TAFI antigen levels in 635 subjects from 27 randomly ascertained Mexican American families participating in the San Antonio Family Heart Study. Mean age of subjects was 42.9 years (range 18 to 96). A genome scan was performed using 419 highly informative autosomal microsatellite markers spaced at circa 10 cM intervals. Additive genetic heritability of TAFI levels was 0.53 ($p < 0.0001$). TAFI antigen levels were higher in males and decreased with age. Environmental covariates (smoking, alcohol use, menopause, exogenous hormones, diabetes status, use of aspirin, anti-lipids or anti-diabetics) and shared household effects did not contribute significantly to TAFI antigen level variation. The genome screen yielded a maximum multipoint LOD score of 3.09 near marker D13S788 on chromosome 13. This is very near the TAFI structural gene (CPB2) located at 13q14.11, and suggests that polymorphisms in the TAFI structural gene or its nearby regulatory elements may contribute strongly to variation in TAFI antigen levels in this population. No other LOD scores > 2 were observed. Our results support those of the Genetic Analysis of Idiopathic Thrombophilia (GAIT) project, which also identified a potential TAFI QTL in the chromosome 13q region in a genome-wide linkage, scan in Spanish families. We are currently typing SNPs in the TAFI gene to further investigate its role in TAFI antigen level variation.

138

Modeling Complex Phenotypes in Association Studies

J.S. Witte
Department of Epidemiology and Biostatistics, University of California, San Francisco, USA

Conventional association studies compare genotypes or haplotypes among cases versus controls. However, cases, commonly include individuals with heterogeneous phenotypes. For example, cases of prostate cancer range from those in which the disease is aggressive and life-threatening to men in which it will have no ill-health effects. Such phenotypes may arise from different genes that have varying effects at particular points in the disease process. Hence, treating all cases as a single group may substantially reduce one's ability to detect associations. This issue is commonly addressed by undertaking subgroup analyses that compare cases with distinct phenotypes (e.g. aggressive, non-aggressive) to controls. But stratifying the cases in this manner can reduce efficiency, and assumes that the heterogeneous phenotypes result solely from modification of the genetic effect. Instead, I show here how one can more appropriately evaluate the effects of genetic and environmental factors on complex phenotypes using ordinal logistic regression models (i.e. polytomous logistic regression, continuation ratio, and stereotype models). I illustrate the increased efficiency and intuitiveness of these approaches with an application to single nucleotide polymorphisms and prostate cancer. Moreover, I extend this approach to matched (e.g. family-based) data. These ordinal logistic regression models can improve conventional analyses by providing a flexible and potentially valuable

framework for evaluating complex phenotypes in association studies.

139

Multivariate Variance Components Linkage Models for Longitudinal PhenotypeData using Gibbs Sampling Approaches

Qiong Yang(1, 2), Larry D. Atwood(1, 2), L. Adrienne Cupples(1)
(1) Dept. of Biostatistics, Boston Univ Schl of Public Health, USA(2) Dept. of Neurology, Boston Univ Schl of Medicine, USA

We have previously extended univariate variance components models to examine linkage for traits measured repeatedly on the same subjects at distinct ages or other temporal measures. A mixed regression model with orthogonal polynomials of age was proposed to model random QTL effects, polygenic effects as well as permanent environmental effects in the repeated measurements. The total and QTL heritabilities were determined as functions of age with parameters estimated via maximum likelihood (ML) methods. In this work, we used Gibbs sampling approaches to estimate the parameters in total and QTL heritability functions. Though computational more intensive, Gibbs sampling approaches may enable us to obtain convergence easier than ML approaches, especially with a large number of regression coefficients in the model. Using a dataset of BMI measured at every two or four years in Framingham Offspring and Cohort subjects, we fitted mixed regression models to examine the variation of a QTL effect with age. Two sampling strategies were compared: the first one sampled parameters by effect: regression coefficients that were used to model the same effect were sampled at once; the second one sampled all the random parameters simultaneously. Bayesian model selection methods such as Bayes Factor were employed to determine the best combination of ranks for all the polynomials and to determine if the QTL effect changed significantly with age.

140

Evaluation of Different Haplotype Block Definition Methods and Characterisation of the Genomic Architecture of Human Chromosome 17q

E. Zeggini (1), A. Barton (2), S. Eyre (2), D. Ward (2), J. Worthington (1, 2), S. John (1)
(1) CIGMR, University of Manchester, UK, (2) arc EU, University of Manchester, UK

The finding of mosaic patterns of linkage disequilibrium (LD) in the human genome has triggered attempts to characterize the distribution of haplotype blocks as a useful tool in association studies. Recent reports have focused on determining the genomic architecture of different chromosomal segments, each by applying distinct haplotype block definitions based on LD measures or haplotype diversity. The effect that different block definitions have on association studies, however, remains unclear. A proposed strategy to utilize blocks in association studies is to initially select haplotype tagging single nucleotide polymorphisms (htSNPs) in the region of interest. We have applied various sets of criteria in characterizing blocks in a 6.1Mb region of chromosome 17q. 189 unrelated healthy individuals were genotyped for 137 SNPs, at a median spacing of 15.5 kb. Haplotype block maps of the region were constructed using published methods and novel methods we have developed. HtSNPs were identified for each map. Blocks were generally found to be shorter and fewer with definitions based on LD measures. Although the distribution of blocks was variable, the number of htSNPs was consistent, indicating that, for the marker spacing used in this study, selection of block definition is not important when used as an initial screen of the region to identify htSNPs. The choice of definition however could be of consequence in designing and interpreting genetic association scans.

141

Statistical Models for Linkage Analysis of Ordinal Traits

H Zhang, R Feng
Yale University School of Medicine, CT, USA

Many health conditions, particularly psychiatric disorders, are recorded in ordinal scales. Although methods and software are available for linkage analyses of dichotomous and quantitative traits, few methods and no standard software are available for conducting linkage analyses of ordinal traits. To fill this gap we exploit the use of an ordinal logistic, mixed effects regression model and compare the power of mapping the trait loci using the ordinal scale with that based on a dichotomized outcome.

142

Linkage Analysis for Complex Diseases Using Admixed Populations

X. Zhu (1), R.S. Cooper (1), R.C. Elston (2)
(1) Department of Preventive Medicine, Loyola University Medical Center, Maywood, Il, USA (2) Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

Linkage disequilibrium arising from the recent admixture of genetically distinct populations can be potentially useful to map genes for complex diseases. McKeigue (1998) proposed a method that condition on parental admixture to detect linkage. We show that this method only tests for linkage under specific assumptions such as equal admixture in the parental generation and admixture that only occurs in a single generation. In practice, these assumptions are unlikely to hold for natural populations, resulting in an inflation of type I error rate when testing for linkage. In this report, we formulate the ancestral probability of an affected person under two different admixture models: 1) intermixture admixture, and 2) continuous gene flow. We then propose a case-only method for detecting linkage that can be more powerful than allele-sharing analysis. We calculate the sample size required for a genome wide search by this method under different disease models, e.g. multiplicative, additive,

recessive and dominant. Our results show that the sample size required to obtain 90% power to detect a putative mutant allele at a genome-wide significance level of 5% can usually be achieved in practice if informative markers are available at a density of 1 cM.

143

Association of haplotype estimates with quantitative phenotypes: variations in the interleukin-21-receptor gene and IgE-Level

A. Ziegler(1), G. Bein(2), M. Hecker(2), A. Bohnert(2), H. Hackstein(2), I. R. König(1), U. Mansmann(3)
(1)Institute of Medical Biometry and Statistics, University Lübeck, Germany, (2)Institute of Clinical Immunology and Transfusion Medicine, University Gießen, Germany, (3)Institute of Medical Biometry and Informatics, University Heidelberg, Germany

Analyzing the association between total serum IgE levels and variations in the interleukin-21 receptor gene (IL21R), recent results revealed an association of carrying at least one IL21R polymorphism (T-83C) with high IgE levels in females (Hecker et al. 2003, Genes and Immunity, 4, 228–233). Haplotypes consisting of four SNPs were estimated separately in probands with high and low IgE levels; the frequencies were found to be different in the two groups. To increase overall power of the study, we used the quantitative information of IgE itself. As we are not aware of any approach for this, we compared two different procedures: In the first, we estimated haplotype probabilities using an Expectation-Maximization algorithm across all individuals, regardless of their IgE level. Secondly, haplotypes were estimated depending on the specific IgE level of an individual. In Monte-Carlo simulations, each person was assigned a combination of two haplotypes based on the estimated probabilities and a random number. To receive stable estimates, the number of replications was set to 100,000. Mean IgE levels were then compared between different groups of haplotypes and haplotype combinations. The investigated methods will be compared with regard to availability of implementations and power.

144

The interplay of LD allele frequencies and multifactorial inheritance in the outcome of complex disease association studies

K. Zondervan (1), A. Morris (1), L. Cardon (1)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford, UK

Case-control studies of complex traits and genetic variants have been notable by their lack of success and replicability. The marker allelic odds ratio (OR) is statistically determined by 4 parameters: the OR of the disease variant; disease and marker allele frequencies (DAF & MAF), and linkage disequilibrium (LD) between marker and disease variant. We derived this relationship, showing that under complete LD and MAF > DAF, the relationship between marker and disease OR reduces to a simple expression. We used 5 associations to illustrate: Deep Vein Thrombosis & FVL (DAF: 0.03); Crohn's disease & NOD2 (0.06); Alzheimer's & APOE (0.15); Bladder cancer & GSTM1 (0.7); and NIDDM & PPAR (0.85). Common MAF variability within haplotype blocks was investigated in empirical (Chr 19) LD data. Allelic ORs of rare/medium DAFs were large: 3.3–4.6. Using a range of marker frequencies and $D' > 0.5$–0.6 in a sample of 1000 cases and 1000 controls, ORs remained detectable with 80% power. The small allelic ORs (1.2–1.3) of common DAFs were only detectable in samples of 5000 cases and 5000 controls when MAFs closely resembled DAFs and $D' > \sim 0.7$. Chr 19 data showed that ~60% of markers were within 0.1 of the most common MAF within blocks. Associations with rare alleles of large effect and common alleles of small effect should be detectable in large case-control studies using common markers (MAF > 0.1) and information on LD. Rare alleles with small ORs are not detectable in feasible samples unless rare markers in very high LD are used.

145

Familial aggregation of endometriosis in the rhesus macaque

K. Zondervan (1), D.E. Weeks (2), L. Cardon (1), A. Goudy Trainor (3), C.L. Coe (4) B. Tier (5), J. Kemnitz (3), S. Kennedy (6).
(1) Wellcome Trust Centre for Human Genetics, Univ. of Oxford, UK, (2) Dept. of Human Genetics & Biostatistics, Univ. of Pittsburgh, USA, (3) Wisconsin National Primate Research Center & (4) Harlow Primate Research Laboratory, Univ. of Madison-Wisconsin, USA. (5) Animal Breeding and Genetics Unit, Univ. of Armidale, Australia, (6) Nuffield Dept of Obstetrics & Gynaecology, Univ. of Oxford, UK

Endometriosis is a complex trait in women characterized by endometrial-like tissue found outside the uterus, causing pelvic pain and infertility. We investigated the familial aggregation of spontaneous endometriosis in a large colony of rhesus macaques. 142 Rhesus macaques with endometriosis were identified between 1981-2002 at the Univ. of Madison-Wisconsin. Females that had died aged $\geq$10 years without endometriosis and had both ovaries until at least 1 year prior to death were considered unaffected. Affected were used to build a large multigenerational pedigree and 9 nuclear families comprising 1,602 females. Heritability was investigated using generalized linear mixed models. Segregation analyses were attempted using a Monte Carlo Markov Chain method. The prevalence of endometriosis in the colony was 31.4% (95% CI: 26.9–35.9%). The average kinship coefficient was significantly higher among affected compared to unaffected ($p < 0.001$), and a higher recurrence risk for full sibs (0.75, 95% CI: 0.45–1.0) was found compared to paternal half-sibs (0.4795% CI: 0.42–0.52). Heritability was estimated between 0.2–0.6. Segregation analyses were limited by the scarcity of disease information.