Genetic Epidemiology 19:235–279 (2000)

ABSTRACTS FROM THE NINTH ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

SAN ANTONIO, TEXAS
OCTOBER 27-28, 2000

1

Analytic Studies of Penetrance under Inaccurate Reports of Family History

C.I. Amos
UT MD Anderson Cancer Center, Houston, TX

Recent estimates of the penetrance of BRCA1 or BRCA2 for breast or ovarian cancer from either highly selected, high-risk registries or from population-based studies have shown discrepancies. Explanations of these differences include i) effects from modifier alleles or other factors that are more prevalent in the high-risk group than in the population-based studies, ii) differences in penetrance conferred by different disease alleles. Here, I studied an alternate possibility, by exploring the effect that inaccurate reporting of family history can have upon estimates of penetrance. For the high-risk families, penetrance has been estimated by the MOD score approach because it allows for the unknown method of ascertainment. Denote the sensitivity as the probability of correctly reporting an individual as affected, given that s/he is affected, and the specificity as the probability of reporting an individual as unaffected given s/he is unaffected. For population-based studies, the penetrance estimate depends directly upon the sensitivity. For many common cancers, the sensitivity in first-degree relatives is typically reported to be between 70-90% and the sensitivity decreases for more distant relatives. For the MOD score method of analysis, the estimation of penetrance is more complex. However, the estimator depends upon the specificity of disease reporting, which has been reported to be over 99%. Thus, the usual biases from inaccurate disease reporting are shown to have a more significant impact on population-based studies than on studies from high-risk registries, if the high-risk registries use MOD score or other methods that condition on the affection statuses of the pedigree members.

2

A Case-Combined Design Using Both Population Based- and Related-Controls: A Potential Alternative for Increasing Power in Gene-Environment Interaction Detection

N. Andrieu[1] and A.M. Goldstein[2]
[1]U521 INSERM, Villejuif, France;
[2]NCI, Bethesda, MD, USA

The interest in studying gene-environment (G×E) interaction is increasing for complex diseases. Among the few methods fully evaluated for their power to detect G×E interactions, most appear inefficient for detecting interaction involving rare event(s) (< 0.20).

In order to gather advantages from related and population-based controls, a design combining both of them is proposed. Cases and two types of controls, population-based controls and sibling controls are recruited. The two types of controls are combined assuming homogeneity for variables involved in G×E interaction. Rela-

tive efficiency of the case-combined control study compared to a classical case- unrelated control study is defined as the ratio of needed case sample sizes. Efficiency is evaluated for different scenarios.

The number of available sibling controls per case and the frequencies of the risk factors of interest are the most important parameters for determining efficiency. The case-combined control design assuming that 50% of cases have an available sib) to detect G×E interaction appears more efficient than a classical case-control study for interaction involving rare events. For common genes, the case-combined control design is often less efficient than a classical case-control study. The range of scenarios where the case-combined control design is both efficient and feasible may appear narrow (i.e., prevalence of G between 0.001 and 0.3). However, it appears to be a useful alternative to study G×E interaction where classical approaches remain unrealistic.

The efficiency of the case-combined design is conditional on the assumption of homogeneity for variables of interest between control groups and between cases with and without sibs. Analytical approaches for this design consist of numerous steps. These issues will be further presented and discussed.

3

A Population-Based Family Study on Colorectal Cancer: Familial Risk Estimation

N. Andrieu[1], G. Launoy[2], R. Guillois[1], C. Ory-Paoletti[1], and M. Gignoux[2]
[1]INSERM 521, Villejuif, France; [2]RTD Calvados, Caen, France

In France, colorectal cancer (CRC) is the second most frequent cancer affecting both males and females (about 33400 new cases per year). A population-based family study was carried out in Calvados, France, from September, 1993 to December, 1998. The main aim of this study was to define the role of genetic factors in the disease transmission taking into account environmental factors from a sample of systematically recorded family data. Firstly, we undertook to determine the magnitude of the risk of CRC among family members.

1351 new cases of CRC were diagnosed in Calvados during the study period. Seven hundred and sixty one families were included. The mean age for the proband is 68.5 years (age range: 25 to 95 yrs). These families were composed of 10,512 relatives (children, siblings, parents, aunts/uncles and grandparents). Familial CRC risks were estimated from a cohort analysis of the relatives. At risk relatives entered the cohort on January 1st, 1970 or at birth if it occurred after (i.e., 117,407 person-years). Expected numbers of CRC were computed from Calvados age-sex and four period-specific incidence rates (from 1978 to 1995).

The risk of CRC among relatives was 1.54 (95%CI: 1.26-1.86). The familial risk appeared greater (not significantly) among the proband 1st degree relatives than among the proband 2d degree relatives (1.71 and 1.22 respectively). If the 1st degree relative was a parent, the risk estimate was greater than if the 1st degree relative was a child or a sibling (3.70, 1.90 and 1.67 respectively). There was no clear pattern in the familial risk according to age. The age of relatives and probands seemed to have a different effect on CRC risk according to gender. Indeed, among men relatives with younger probands (≤ 60yrs), the cumulative risk of CRC increased at early ages, among women, this risk increased at a later age. Caution should be exercised when interpreting these results. They will be discussed more extensively.

4

Models for Familial Breast Cancer Incorporating BRCA1, BRCA2 and Other Genes

A.C. Antoniou[1], P.D. Pharoah[2], N.E. Day[3], B.J. Ponder[2], and D.F. Easton[1]
[1]CRC Genetic Epidemiology Unit, IPH; [2]CRC Human Cancer Genetics Research Group; [3]Dept of Public Health and Primary Care, University of Cambridge, UK

We used data from both a population based series of breast cancer cases diagnosed below age 55, and high risk families in the UK, with information on BRCA1 and BRCA2 mutation status, to investigate the genetic models that can best explain familial breast cancer outside BRCA1 and BRCA2 families. We also evaluated the evidence for genes (or other familial factors) that modify the risk in BRCA1 and BRCA2 carriers. We estimated the simultaneous effects of BRCA1, BRCA2, a third hypothetical gene "BRCA3", and a polygenic effect, using segregation analysis. The hypergeometric polygenic model was used to approximate polygenic inheritance and the effect of modifying genes. The best fitting model for the residual familial breast cancer was the polygenic with no evidence for a third major gene. There was significant evidence for modifying genes on the risks of breast cancer in BRCA1 and BRCA2 mutation carriers. Under this model, the frequency of BRCA1 was estimated to be 0.051% (95%CI: 0.021-0.125%) and of BRCA2 0.068% (95%CI: 0.033-0.141%). The breast cancer risk in mutation carriers by age 70 (based on the average incidence over all modifying genotypes) was estimated to be 35.3% for BRCA1 and 50.3% for BRCA2. The corresponding ovarian cancer risks were 25.9% for BRCA1 and 9.1% for BRCA2. The findings suggest that several common, low penetrance genes with additive effects may account for the residual non-BRCA1/2 familial aggregation of breast cancer. There was no significant difference between

the variance of the modifying effect and that of the polygenic effect, suggesting that the same common low penetrance genes may act on both carriers and non-carriers and confer similar relative risks. The modifying effect may explain the previously reported differences between population based estimates for BRCA1/2 penetrance and estimates based on high-risk families.

5

BRCA1 Genotyping with Temporal Temperature Gradient, Constant Denaturant Capillary Electrophoresis (CDCE)

C. Aragaki*[1,2,3], L. Jing[1], and H. Zarbl[1]
[1]Program in Cancer Biology, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; [2]Department of Epidemiology, University of Washington, Seattle, WA; [3]Dallas MPH Program, University of Texas-Houston Health Science Center School of Public Health, Dallas, TX

One difficulty in conducting molecular epidemiology studies is the genotyping costs for large populations. Previously, we showed in simulation studies that a pooling protocol could be cost effective for rare allele frequencies less than 10% and bipartite pooling is less efficient than splitting into 3-5 subpools. However, the pooling protocol requires rare mutation detection methods. Temporal Temperature Gradient, Constant Denaturant Capillary Electrophoresis (CDCE) has been used with great success for very rare mutation detection. In previous studies, we have detected one k-ras mutation in 1 million wild-type copies. Because of the extremely sensitive detection methods, we proposed to use this technique to screen large populations in pooled samples for infrequent alleles, subpooling if the pool contains the rare allele. As the rare allele frequency becomes low, as in the case for *BRCA1*, the efficiency of individual genotyping becomes very cost ineffective. Using CDCE, the authors detected the 186-AG deletion on exon 2 in *BRCA1* and recapitulated the wild-type allele homoduplex, the AG deletion homoduplex, and both heteroduplexes, based on the differing melting patterns of these sequences. We intend to expand this method to cover the entire *BRCA1* coding and neighboring intronic regions.

6

Homogeneity Tests for Genome Scans of Multifactorial Diseases

M.C. Babron, H. Selinger-Leneman, and F. Clerget-Darpoux
INSERM U535, Le Kremlin-Bicêtre, France

Detecting linkage with a genetic marker is a way to provide evidence for the presence of a genetic risk factor. For multifactorial diseases, this is performed by analyzing a sample of affected sibs through model-free statistics, such as the Maximum Lod Score (MLS, Risch, 1990). Most often, a single analysis does not reach the recommended level of significance for unambiguously concluding to linkage. The strategy is then to assemble a new independent sample of affecteds to attempt to replicate the first finding. However, even when both samples are similar in ascertainment criteria, results often seem contradictory both in MLS value and in position.

We propose here a test to determine whether those two samples are compatible both in MLS value and position, assuming the existence of a risk factor in this region with characteristics (IBD vector) corresponding to that estimated from the joint analysis of the two samples.

This test is applied to data on Celiac disease. Analysis of two samples of 100 pairs leads to an MLS value of 2 and 1.1 respectively, with the maximum reached 35cM apart on the long arm of chromosome 5. We show that these results are compatible with the existence of risk factor with a moderate effect on chromosome 5qter.

It is also applied to the simulated data from the Genetic Analysis Workshop 11 (1998) where the underlying genetic model is known.

7

Effect of Stratifying by Linkage at Known Loci in Assessing a Genome-Wide Scan of Familial Prostate Cancer

M.D. Badzioch[1], J.L. Stanford[2], M. Gibbs[2], M. Janer[3], E.L. Goode[1], M. Peters[2], S. Kolb[2], L. Hood[3], E.A. Ostrander[2], and G.P. Jarvik[1]
[1]University of Washington; [2]Fred Hutchinson Cancer Research Center; [3]Institute for Systems Biology, Seattle, WA

Five putative familial prostate cancer (FPC) susceptibility loci, *HPC1*, *PCAP*, *HPCX*, and *CAPB*, and recently, *HPC20* (chr. 20q13), have been reported. To increase genetic homogeneity, most genome scan studies stratify by average age at diagnosis, number of affected men, and clinical data, but evidence of linkage to known loci has not yet been used. We present now effects of stratifying by multipoint NPL scores at the above 4 former loci in assessing a 10cM scan of 94 FPC families. Using age alone, the highest scores were at *CAPB* (NPL=2.28, avg. age at diagnosis <61 years, n=16 families) and on chrs. 4 (2.06, <66, 50), 8 (2.02, ≥66, 44), and 15 (2.04, ≥66, 44). Since combined linkage to the 4 loci is estimated to be <50% we split our families into 2 equal groups based on an NPL >0.7 at any of the 4 regions. We then resurveyed the genome using sets of families defined by age and prior linkage status and found peaks (other than *CAPB*) only on chrs. 4 (2.68, <66, 27;

p=0.006), 10 (2.49, <66, 23; p=0.014), and 20 (2.76, ≥66, 23; p=0.005). Interestingly, the peaks on chrs. 4 and 10 occurred in families that also linked to one of the candidate loci. Since the linkage to chr. 20 in older onset families was consistent with HPC20 we maximized the NPL to 4.20, in 11 families by increasing the NPL score to be considered linked to the other loci to 1.10. These results further support locus heterogeneity in FPC and the existence of a locus at chr. 20q13.

8

African-American Hereditary Prostate Cancer Consortium: A Genetic Linkage Study

A. Baffoe-Bonnie[1,3], C. Royal[1,2], R. Kittles[2], I. Powell[4], J. Bennett[4], G. Hoke[4], C. Pettaway[4], S. Weinrich[4], S. Vijayakumar[4], C. Ahaghotu[2,4], T. Mason[4], W. Boykin[4], J. Carpten[1], J. Bailey-Wilson[1], J. Trent[1], K. Berg[1], G. Dunston[2], and F. Collins[1]
[1]NHGRI/NIH; [2]Howard University; [3]Fox Chase Cancer Center; [4]Collaborative Recruitment Centers USA

African American males have the highest incidence and mortality rates of prostate cancer in the United States. The African American Hereditary Prostate Cancer (AAHPC) Study is an ongoing multicenter genetic linkage study organized by Howard University and the NHGRI, with support from the Office for Research on Minority Health and the NCI. The goals of the study are: 1) To look for evidence of involvement of chromosome 1q24-25 (HPC1) in African American men with hereditary prostate cancer (HPC) and 2) To conduct a genome-wide search for other published loci associated with HPC. In order to accomplish these goals, a network has been established including Howard University, the NHGRI, and six Collaborative Recruitment Centers [Michigan, Georgia, New York, Texas, Illinois and South Carolina] to identify and enroll 100 African American families.

To date, 43 families with at least four first-degree relatives diagnosed with prostate cancer have been enrolled. Recruitment strategies have included mass media campaigns, physician referrals, community health-fairs/prostate cancer screenings, support groups, tumor registries, as well as visits to churches, barbershops, and universities. Physician referrals and tumor registries alone yielded a total of 35 (81%) families. This is the first large-scale genetic linkage study being conducted almost entirely by African American clinical investigators and scientists.

The plan is to recruit more families to increase the power to test linkage to known chromosomal regions, and to search for other regions of the genome which may harbor genes involved in prostate carcinogenesis. Preliminary results will be presented.

9

Sequential Designs for Genetic Epidemiological Studies: A Review on the Literature

Inke Böddeker and Andreas Ziegler
Medical Center for Methodology and Health Research, Institute of Medical Biometry and Epidemiology, Philipps-University of Marburg, Germany

The cost of a genetic epidemiological study is largely determined by the number of persons to be recruited, phenotyped, and genotyped. The efficiency can be increased by using a sequential procedure that reduces time and cost by average.

Two strategies for sequential designs in genetic epidemiological studies can be distinguished: One approach is to increase the sample size sequentially and to conduct multiple significance tests on accumulating data. If, according to a given criterion, significance or futility can be assumed with a certain probability, the study is stopped. Otherwise it is carried on to the next stage. With an adaptation of the group sequential design by Pocock (1977), this strategy has been followed by Chotai (1984).

The second approach is to conduct early linkage analyses on a coarse marker grid, and to increase marker density in later stages. Interim analyses are performed to select interesting genomic areas for the follow up. In this line, Elston (1996) and Guo and Elston (2000) suggested two stage procedures for the analysis of affected sib pairs.

The aim of this presentation is to give a review on sequential procedures applied or suggested for application in genetic epidemiological studies. The majority of proposed study designs is developed to meet the demands of specific studies and lacks a theoretical foundation. A second group of procedures is based on simulation results and principally restricted to the specific simulated situation.

Finally, some theoretically founded procedures have been proposed that will be discussed in detail. Although interesting and promising procedures have been suggested, they still have some theoretical flaws. It will be concluded that further developments are required to recommend sequential strategies for practical use in genetic epidemiological studies.

10 [Invited Speaker]

Critical Issues in Quantitative Trait Linkage Analysis

J. Blangero
Soutwest Foundation for Biomedical Research, San Antonio, TX, USA

There is a growing interest in the genetic dissection of the numerous quantitative traits that serve as potential risk factors for common complex diseases. The likelihood-based variance component method is now widely used for the

linkage analysis of such quantitative traits. While this approach has many benefits, including modeling flexiblity, the importance of the underlying assumption of multivariate normality for a trait within pedigrees has not been studied extensively. Simulation studies have shown that traits with leptokurtic distributions yield linkage test statistics that exhibit excessive Type I error rates when analyzed naively. A number of robust alternatives, such as the utilization of the multivariate t distribution, have been proposed to eliminate this problem but the properties of these tests are only now beginning to be studied. In this presentation, we describe a new analytical approach that can be used for traits with arbitrary distributions. Using misspecification theory, we derive analytical formulae relating the kurtosis and total heritability of a quantitative trait to this deviation from the expected asymptotic distribution of the LOD score. A simple correction constant that leads to a robust LOD score is provided for any deviation from normality and for any pedigree structure. Use of this robust version of the LOD score will eliminate the problem of underlying probability model misspecification in variance component-based linkage analysis.

11

Evaluation of Multivariate Reduction Techniques for Deriving Informative Phenotypes for Genetic Linkage Analysis

I.B. Borecki, M.A. Province, T. Rice, M. Feitosa, and D.C. Rao
Washington University School of Medicine, St. Louis, MO

Suites of correlated phenotypes may contain additional information for gene discovery over their univariate components if they are influenced by pleiotropic loci or if they represent alternative, but imperfect, measures of the same underlying true phenotype. Principal components (PC) / factor analysis represent potential approaches for deriving phenotypes with improved signal-to-noise ratio for linkage studies. We have investigated their utility using SEGPOWER, a flexible simulation program that provides likelihood ratio tests of linkage in a variance components framework. So as to not confound our observations with simple power issues, we have simulated 1000 nuclear families, each with 3 offspring for every condition. The basic model is trivariate with locus g_1 affecting all three traits, g_2 affecting only two, and g_3 affecting only one. These trait loci are unlinked to one another, but each is completely linked to a fully-informative marker; an unlinked control marker to assess false positive rates is also included. We evaluate the relative utility of fitting PCs to the phenotypic vs. the genotypic covariance matrices, over a variety of combinations of locus-specific heritabilities and identify the situations in which PCs are more informative in these relatively ideal circumstances. Further, we assess the effects of adding residual correlations among the traits on the relative efficiency for detecting trait loci. We conclude that, under certain circumstances, PCs may improve the ability to detect trait loci, however in real situations, interpretation of the results may be challenging.

12

Complex Diseases in Founder Populations: Use of Closely Related Affected Individuals

C. Bourgain[1], E. Génin[1], J. Partanen[2], and F. Clerget-Darpoux[1]
[1]INSERM U535, Kremlin-Bicêtre, France; [2]FRC Blood Transfusion Service, Helsinki, Finland

We have proposed a method to locate genetic risk factors for complex diseases in founder populations, the Maximum Identity Length Contrast statistic (MILC). MILC compares the identity length among parental haplotypes transmitted to affected offspring with identity length among non-transmitted parental haplotypes.

Initially, closely related affected individuals were discarded from the analysis and only one affected by sibship was considered.

Since nuclear families with multiple affected sibs as well as large pedigrees are often available in founder populations, we performed simulations to investigate the properties of MILC in the presence of closely related affected individuals.

We show that the use of closely related affected individuals greatly enhances the power of the statistic. For a given sample size and type one error, we found that the power may be increased by a factor of 2.5 when using affected sibpairs instead of random affected individuals from the population.

This increase is related to an increase of kinship coefficient contrast between haplotype groups when considering closely related individuals. An interest of the MILC approach is that it allows the simultaneous use of affected individuals from a founder population with any kind of relationship.

We illustrated our purpose by an analysis of the role of HLA in the determinism of celiac disease. We showed that the effect of HLA in celiac disease may be detected with MILC by typing only 11 affected individuals members of a single large pedigree.

13

Westernization and Family History as Determinants of Breast Cancer Risk in Asian American Women

C.M. Bromley[1], M.H. Greene[1], T. Fears[1], R. Falk[1], A.M.Y. Nomura[3], M.C. Pike[2] D.W. West[4], A.H. Wu[2], R.N. Hoover[1], and R.G. Ziegler[1]
[1]DCEG, NCI, MD; [2]USC School of Medicine, LA; [3]University of Hawaii, HI; [4]Northern California Cancer Center, CA

Breast cancer (BC) incidence rates have historically been 4-7 times higher in the United States than in China and Japan, although the reasons remain elusive. When Asian-American women migrate to the U.S., BC risk rises over several generations and approximates that among U.S. Whites. We explored the interaction between family history of BC in first-degree relatives (FH), and level of Westernization as determinants of BC risk. In a population-based study of BC among women of Chinese, Japanese, and Filipino ancestry living in San Francisco, Los Angeles and Oahu. Interviews were conducted during 1985-1989 with 597 BC cases (62 with a positive FH) and 966 control subjects (43 with a positive FH). Unconditional logistic regression was performed to obtain relative risks (RR) and 95% confidence intervals (CI), adjusting for ethnicity, center and age. Among first-degree relatives of probands, the risk of BC was significantly elevated [RR=2.6; 95% CI=1.7-3.9]. Women *born in the East*, with a positive FH, had a similar RR to women *born in the West* with a positive FH [RR=2.3; 95%CI=1.2-4.6 vs. RR=2.2; 95% CI=1.3-3.9]. Thus, given the increase in BC risk with Westernization, the absolute increase in BC risk associated with a positive FH may be substantially greater in Westernized women. These preliminary results suggest a possible synergism between environmental risk factors and genetic effects of BC among women of Asian ancestry who migrate to the United States.

14

Adjusting for Ascertainment in Gibbs Sampling Based Generalized Linear Mixed Models (GLMMs)

P.R. Burton[1,2], K.J. Scurrah[1,2], and L.J. Palmer[2,3,4]
[1]University of Leicester, UK; [2]TVW Telethon Institute for Child Health Research, Perth, Western Australia; [3]Case Western Reserve University, Cleveland, OH; [4]Harvard Medical School, Boston, MA

Using Gibbs sampling in BUGS we have shown how to construct generalized linear mixed models (GLMMs) for a range of applications in genetic epidemiology. The GLMMs generate consistent parameter estimates and can be used for Normal and binary traits and for traits reflected in a right censored survival time, including an age-at-onset. We have recommended that the random effects associated with σ^2_A (the variance attributable to additive polygenic effects) may be used as an "adjusted" phenotype and input to a standard linkage package. The "SSARs" (Sigma Squared A Random effects) have a nicer distribution than standard "residuals" and are corrected both for observed covariates and for unobserved but shared environmental determinants. This approach provides a particularly useful way to undertake a linkage analysis based on an age-at-onset. Unfortunately, when interest focuses on σ^2_A and the SSARs, the GLMMs are relatively sensitive to ascertainment and the introduction of an adjustment for ascertainment has historically proven to be difficult in Markov chain Monte Carlo (MCMC) based variance component models. However, using a Metropolis-Hastings based feature that has recently been added to WinBUGS, we will show how the GLMMs may readily be corrected for ascertainment. We will consider some of the theoretical and practical issues underlying this adjustment and will illustrate our methods using both simulated and real data.

15

Segregation Analysis of the Respiratory Disturbance Index (RDI) in African-Americans and European-Americans

S. Buxbaum, R.C. Elston, J. Aylor, B. O'Malia, S. Morton, K. Clark, G. Graham, P.V. Tishler, and S. Redline
Case Western Reserve University, Cleveland, OH

Differences in age of onset and anatomic risk factors for obstructive sleep apnea (OSA) in European- and African-Americans suggest possible racial differences in the genetic etiology of the disorder. In this study, we assess sex-specific transmission patterns in an African-American sample of 123 families comprising 710 family members and a European-American sample of 177 families comprising 1202 family members. In each sample, these patterns were assessed using two variables: 1) the respiratory disturbance index (RDI), log transformed and adjusted for age and 2) the RDI, log transformed and adjusted for age and body mass index (BMI). Subjects underwent home sleep monitoring with recording of thermistry, oximetry, chest wall impedance and heart rate, which were used to determine the RDI. The analysis suggests Mendelian inheritance with separate distributions for each sex, accounting for more of the variation in males than in females. These results provide support for an underlying genetic basis for OSA independent of the contribution of BMI to the disease.

16

Power to Detect Gene-Environment Interactions in Complex Diseases by Combined Segregation-Linkage Analysis With and Without Linkage Disequilibrium

V. Chaudru, M. Rosenberg-Bourgin, and F. Demenais
INSERM EPI 00-06, Paris, France

The interest in conducting studies to examine gene-environment interactions (G×E) is increasing for most chronic diseases. Our goal is to investigate, through simulations, the statistical properties of combined segregation-linkage analysis to detect G×E. Disease status and marker data were generated on five-member nuclear families. The liability to the disease was simulated under the mixed model including a major gene effect interacting with an environmental factor plus a polygenic component. The proportion of total variance due to each component was varied as follows: 10 to 30% for the major gene, 10% to 30% for G×E, and 0 to 30% for the polygenic component. The disease prevalence was fixed at 5% or 10%. The linked marker was assumed to be diallelic (SNP) tightly linked and in linkage disequilibrium (LD) with the disease locus. One hundred replicates of family samples were generated for each set of parameters. Samples sizes were chosen to yield at least 80% power of detecting the disease gene under the true model. Data were analyzed by the regressive threshold model taking into account or ignoring LD. Our results show that the power of detecting G×E is reduced by a factor of 0.8 to 0.4 when ignoring LD as compared to including LD in the analysis. This reduction in power is not influenced by the mode of inheritance of the gene but increases as the gene frequency and the overlap of the underlying genotypic distributions increase. These simulations will be extended to more complex situations including various sampling schemes and missing data.

17

TP53 Gene Polymorphisms and Risk for Breast Cancer by Age 50 in Germany

J. Chang-Claude[1], I.B. Runnebaum[2], and
S. Wang-Gohrke[3]
[1]Deutsches Krebsforschungszentrum, Heidelberg, Germany; [2]University of Freiburg, Freiburg; [3]University of Ulm, Ulm, Germany

In a population-based case-control study of breast cancer among women diagnosed by the age of 50, conducted in two geographic areas in Germany, 560 cases and 550 age-matched controls were genotyped for three biallelic polymorphisms in introns 3 and 6 and exon 4 of the *TP53* gene to evaluate the effect on risk for breast cancer. Breast cancer risk was significantly increased by 30% in women carrying the 16bp duplication polymorphism in intron 3. Compared to wild-type, the adjusted odds ratios for A1/A2 genotype and A2/A2 genotype were 1.27 and 1.66 respectively, showing a significant multiplicative risk effect per A2 allele. There was evidence for a differential effect by family history of breast cancer (p=0.03), however, the higher risk associated with the A2 allele in women with first-degree family history was primarily observed among parous women. Similar albeit weaker associations were found with the A allele of the intron 6 *Msp1* RFLP and the Pro allele of the Arg72Pro polymorphism. Since the *TP53* intron 3 and 6 polymorphisms show strong linkage and to a lesser extent to the Arg72Pro polymorphism, analysis of haplotype associated risk will be pursued.

18

Why Are Genome Scans of Multifactorial Diseases So Discordant?

F. Clerget-Darpoux, H. Selinger-Leneman, and
M.C. Babron
INSERM U535, Le Kremlin-Bicêtre, France

A popular strategy to detect genetic risk factors in a multifactorial disease is to search for linkage by performing a systematic screening of the genome using the Maximum Lod Score (MLS) on affected sibpairs (Risch, 1990).

We show that, for a genetic factor with moderate effect, the MLS variance is very large, leading to the possibility of very different MLS values in two samples of same size. For a factor such as APOE in Alzheimer's disease considering a sample of 100 sibpairs and a 10 cM map, the probability to obtain an MLS greater than 3 is the same as the probability of having it smaller than 1 (20%). In addition, when the linkage threshold is reached, the position of the MaxMLS varies over 35 cM. The variation is even more dramatic for a risk factor with a smaller effect such as INS in Insulin-Dependent Diabetes or HLA in Multiple Sclerosis. Our results demonstrate not only the low power to replicate a significant result in an independent sample, but also the poor resolution of linkage analysis to highlight a region

This may explain the apparent discrepancy between the results of genome scans performed on most multifactorial diseases.

19

Oligogenic Segregation Analysis of Prostate Cancer Pedigrees Using Markov Chain Monte Carlo Methods

E.M. Conlon[1], E.L. Goode[1], M. Gibbs[2], J.L. Stanford[2], M. Badzioch[1], M. Janer[3], L. Hood[3], E.A. Ostrander[2], G.P. Jarvik[1], and E.M. Wijsman[1]
[1]University of Washington; [2]Fred Hutchinson Cancer Research Center; [3]Institute for Systems Biology, Seattle, WA

Previous studies have suggested strong evidence for a familial component to prostate cancer (PC) susceptibility. Here, we analyze 2,418 individuals in one hundred

forty-nine prostate cancer families recruited as part of the ongoing Prostate Cancer Genetic Research Study (PROGRESS). We use Markov chain Monte Carlo (MCMC) methods for oligogenic segregation analysis to estimate the number of quantitative trait loci (QTL) and their contribution to the variance in age at onset of PC. We find evidence that a mean of 3.78 QTLs make contributions to the variance in age at onset of PC. We estimate an 89.8% posterior probability of at least 3 QTLs contributing to the variance in age at onset of PC. Environmental effects contribute ~18% to the variance in age at onset, with genetic variance contributing the remaining ~82%. We estimate that the largest QTL contributes ~74% to the genetic variance, with the second largest contributing ~17%, the third largest contributing ~6%, and the remaining QTLs contributing less than 3% to the genetic variance. We estimate that the largest QTL contributes square-root variance of 10.6 years to shift in age at onset of PC. Our estimates are based on simulations with 500,000 iterations using a DEC Alpha XP1000 workstation, with run time ~2 hours, 45 minutes.

20

An Agent-Based Computer Simulation Program to Analyze Genetic Systems

D.V. Conti and J.S. Witte
Case Western Reserve University, Cleveland, OH, USA

Agent-based models are ideal for the study of complex nonlinear systems and nested subsystems, such as the interaction of genes, culture and environment in the study of disease. They allow for the inve)stigation of temporal and spatial processes with adaptive, heterogeneous individuals whose behaviors change as their characteristics change. In this talk, we present a recently developed and freely available simulation program, the Genetic Epidemiologic Research Analyst (GenERA). GenERA is an agent-based model that simulates individuals, with unique characteristics and behaviors, interacting in an artificial environment. GenERA can assign individuals a variety of genetic and cultural traits, such as Mendelian markers, additive polygenic diseases, and mating and movement rules. In GenERA, as individuals evolve through time they undergo mutation, recombination, selection, and migration, resulting in emergent population dynamics. The program's graphical user interface allows observation of population patterns and study of the population with family-based, case-control or case-only sampling. The simulated data can then be used to investigate methods of linkage, segregation, and association for validity, power, and robustness. This type of simulation is particularly suited for research where population history influences one's ability to detect traits with standard genetic analysis. Additionally, the use of GenERA as an interactive educational tool enables students to visualize and explore the interaction of human history, genes, culture and the environment on methods of genetic analysis.

21

Trend-TDT: A New Transmission Disequilibrium Based Test to Detect Trends in Response to Increasing Allele Length of Functional Mini/Microsatellites

Marilys Corbex[1], Laurent Oget[2], and Bruno Fallissard[3]
[1]LGN CNRS UMR C9923, Hôpital Pitié-Salpêtrière, Paris, France; [2]LIAFA, Université Paris 7, Paris, France; [3]INSERM U169, Villejuif, France

Mini- and micro-satellites are associated with human disease, not only as markers of risk but also directly in disease aetiopathogenesis. They may play significant roles in replication, repair and mutation of DNA, regulation of gene transcription and protein structure alteration. Notably, numerous *in vitro* studies have shown that mini/microsatellites affect gene transcription, the transcription being proportionally increased (or decreased) to the number of repeated sequences (i.e. length of alleles). These results point to the fact that phenotypes can be affected by mini/microsatellites proportionally to the length of allele. We address this point in the context of complex traits and review briefly how and which mini/microsatellites could be involved in such a mechanism. Actual family-based association tests do not allow to test such phenomenons. We have then developed a new test of association between putative functional mini/microsatellites and qualitative phenotypes that assess the significance of a linear trend across allele lengths. The test is a paired t-test based on the quantitative value assigned to the length of alleles transmitted from the heterozygous parents to their affected children and is named trend-TDT.

A program to perform the test is available at http://liafa.jussieu.fr/~oget/trend-TDT.html

22

Statistical Modeling of Inter-Locus Interactions in a Complex Disease: Rejection of the Multiplicative Model of Epistasis in Type 1 Diabetes

Heather J. Cordell[1], Natasha J. Hill[1], Christopher J. Lord[1], Paul A. Lyons[1], Lawrence B. Peterson[2], Linda S. Wicker[2], and John A. Todd[1]
[1]University of Cambridge, UK; [2]Merck Research Laboratories, NJ

In general, common diseases do not follow a Mendelian inheritance pattern. Their genetic dissection may be assisted by evaluation of linkage in mouse models of human disease. Statistical modeling of multiple-locus

linkage data from the nonobese diabetic (NOD) mouse model of type 1 diabetes has previously revealed strong epistasis between alleles of several *Idd* (insulin-dependent diabetes) loci. The construction of NOD congenic strains containing selected segments of the diabetes-resistant strain genome allows analysis of the joint effects of alleles of different loci, in isolation without the complication of other segregating *Idd* loci. In this report, we analyze data from congenic strains carrying two chromosome intervals (a double congenic strain) for two pairs of loci: *Idd3* and *Idd10*, and *Idd3* and *Idd5*. The joint action of both pairs is consistent with models of additivity on either the log odds of the penetrance, or the liability, scale, rather than with a multiplicative model of epistasis. For *Idd3* and *Idd5* we would also not reject a model of additivity on the penetrance scale, which would indicate a disease model mediated by more than one pathway leading to β-cell destruction and development of diabetes.

23

GoM Maps IDDM11 to Chromosome 14q24.3-q31 and Demonstrates the Relative Contribution of HLA-DR and -DQ to IDDM in Multiplex Families

E.H. Corder[1], M.A. Woodbury[1], D.K. Madsen[1], and L.L. Field[2]
[1]The Center for Demographic Studies, 2117 Campus Drive, Duke University, Durham, NC 27701;
[2]Department of Medical Genetics, University of Calgary, Calgary, Alberta, Canada T2N 4N1

GoM (Grade-of-Membership) analysis was employed to map the IDDM11 diabetes locus on chromosome 14 (Field et al.., 1996), simultaneously demonstrating the importance of HLA-DR and –DQ loci for IDDM risk in multiplex families. There were 603 sibling pairs. A GoM model was constructed from information on the extent of IBS sharing within pairs for eight markers spanning IDDM11 and also shared HLA-DR and –DQ alleles. Five profiles were identified labeled 'I' to 'V'. They varied in terms of IDDM11 region allele sharing (0, 1, 1, 1, 2 for markers close to IDDM11) and HLA sharing (1, 0, 1, 2, 1). The genetic information predicted the frequency of doubly affected pairs: (30%, 28%, 42%, 94%, 55%). Only profile IV with complete HLA sharing and 94% 'recurrence risk' was composed of pairs sharing HLA-DR 3,4 and –DQ 97,109 (associated with IDDM). This profile was infrequent (7%). Complete IDDM11 region sharing (V) increased the frequency of double affected pairs to 55% from 42% (III). The distribution of shared HLA alleles for these profiles was similar, i.e., little gene-gene interaction. Profiles I and II had substantial recurrence risk and composed 43% of the sample. Further genetic dissection of IDDM might be accomplished by mapping other loci simultaneously with IDDM11 and HLA. In conclusion, GoM may represent a useful approach to genetic linkage analysis for IDDM and other disorders.

24

Modeling Susceptibility to Colorectal Cancer Using Segregation Analysis

G.P. Crockford[1], D.T. Bishop[1], D.J.B. St John[2], E.A. Debney[2], F.T. McDermott[2], and E.S.R. Hughes[2]
[1]Imperial Cancer Research Fund, Leeds, England;
[2]Department of Gastroenterology, The Royal Hospital, Melbourne, Australia

Segregation analyses of susceptibility to colorectal cancer (CRC) assuming the mixed model have uniformly identified a rare, dominantly inherited susceptibility with lifetime penetrance 0.5 or higher. We now know that germline mutations in mismatch repair genes (MMR) cause the syndrome, Hereditary Non-Polyposis Colorectal Cancer. Approximately 50% of Amsterdam criteria families (3 CRC, one less than 50 yrs) and 10% of multicase bowel families can be attributed to MMR mutations while families with a weaker family history are unlikely to be due to MMR genes. We are interested in (i) estimating the contribution of mismatch repair genes to colorectal cancer incidence in the general population and (ii) evaluating the evidence for other genetic factors being involved in bowel cancer susceptibility in families without MMR. We obtained detailed and verified family histories from 525 bowel cancer patients of Sir Edward Hughes in Melbourne. Four of these families are known to have MMR repair mutations. Segregation analysis of the remaining families identified a rare dominant gene with frequency 0.0072 and lifetime penetrance 0.30. Recognizing that a proportion of the Amsterdam criteria and multicase families are expected to have MMR mutations we randomly removed these from the segregation analysis according to the above proportions. The overall best fitting model is a dominant with gene frequency 0.0068 and lifetime penetrance of 0.29. Even if we exclude all Amsterdam criteria families the dominant model again provides the best estimate with gene frequency 0.0099 and penetrance 0.23. The results suggest other genes are involved in susceptibility to colorectal cancer.

25 [Invited Speaker]

Approaches to LD-Mapping Using Multiple Markers

David Curtis
St. Bartholomew's and Royal London Hospital School of Medicine and Dentistry

Because of their ubiquity, SNPs have considerable potential for fine-mapping susceptibility loci using

methods to detect linkage disequilibrium. However the fact that they are biallelic means that each SNP provides less information than would be available from a more polymorphic marker. This problem might be overcome if linked SNPs could be combined into haplotypes. However it is difficult to detect the haplotypes present in case-control samples. Also, recombination means that disease-related haplotypes may exist only in an incomplete form. The ability of artificial neural networks (ANNs) to recognize patterns, even when data are missing, means that they may have a useful role in detecting partial haplotypes concealed within multilocus genotypes. We have used a neural network to analyze simulated data and show that it complements single-locus analyses to produce a valuable increase in power to detect association with a disease locus. This occurs mainly when more than one mutation at the locus can increase susceptibility, meaning that there is more than one disease-associated haplotype. We continue to explore additional approaches in an attempt to maximize the ability to detect association using multiple markers simultaneously.

26

Distribution of the Cystic Fibrosis Allele DF508 in North and South America

D. Daley[1], R.C. Elston[1], and E. Palsson[2]
[1]Case Western Reserve University, Cleveland, OH; [2]College of Staten Island, Staten Island, New York, NY

Information on over 7,000 cystic fibrosis (CF) chromosomes, with identification of different CF alleles, was obtained for eight countries in North and South America (Canada, United States, Mexico, Argentina, Columbia, Venezuela, Chile, and Brazil) from the CF Mutation Database and previously published data. Weighted multiple regression analyses of the relative frequency of the DF508 DF allele were performed, using longitudes and latitudes of the reporting cities as predictor variables. When possible, ethnic populations for each reporting city were identified and analyzed separately. Three geographical areas were analyzed: the United States, North America, and North and South America combined. Two sets of weights were used, the number of chromosomes reported for each city, and the inverse of the estimated variance of the relative frequency for each city. In each case, latitude was found to be a significant predictor of the frequency of the DF508 allele ($p<0.0001$), with the trend showing increasing frequency of the DF508 with northward progression. We discuss the possible reasons for this cline, such as a greater fitness of the DF508 allele in the North, a greater proportion of Caucasians in the North, and the hypothesis that the ancestral DF508 mutation arose in a Caucasian population.

27

Oligogenic Segregation Analysis to Estimate Modifier Loci in PS2 Alzheimer's Disease Pedigrees

E.W. Daw[1], T.D. Bird[1,2], E.J. Nemens[2], D. Nochlin[1], G.D. Schellenberg[1,2], and E.M. Wijsman[1]
[1]University of Washington, Seattle, WA; [2]VA Medical Center, Seattle, WA

After a major locus is identified, an important additional task in dissecting a genetic trait is determining if other loci modify the trait. Here, we apply Monte Carlo Markov chain (MCMC) methods for oligogenic segregation and linkage analysis (Heath, AJHG, 61:748) to Volga German (VG) kindreds in which a PS2 mutation leading to early-onset Alzheimer's disease (AD) was identified. The mutation is the same in the 9 VG families (245 individuals) with ≥1 confirmed affected carrier, yet range in age-at-onset is wide (40-75 years with one inferred carrier who died unaffected at 89). This range suggests effects in addition to PS2 are present. Our analysis simultaneously estimated variance in age-at-onset of AD due to environment/residual effects, family-specific effects, PS2, apoE, and unknown loci, while also estimating the number of unknown loci. Environment/residual effects accounted for ~6% of variance, family-specific effects ~14%, PS2 ~60%, apoE ~2%, and unknown loci ~19%. We estimate a 96% posterior probability of ≥1 loci in addition to PS2 and apoE, and 73% of ≥2 loci. The variance of the largest two unknown loci is ~13% and ~4%, more than apoE. Parameter estimates for these loci suggest they might be a subset of those estimated in analysis of late-onset AD (Daw et al., AJHG, 66:196). These results demonstrate the utility of these MCMC methods and provide evidence that both apoE and other unknown loci modify age-at-onset of AD caused by PS2 mutation.

28

Candidate Genes for Alopecia Areata

M. de Andrade[1], R. McMillen[2], Q. Zhang[3], and M. Duvic[2]
[1]Biostatistics, Mayo Clinic/Foundation, Rochester, MN; [2]Dermatology and, [3]Epidemiology, UT MD Anderson Cancer Center, Houston, TX

Alopecia Areata (AA), patchy hair loss which may progress to total scalp or body hair loss, affects 1-2% of the population. AA is a T cell mediated immune response directed at hair follicles, and may be another organ specific autoimmune disease, like diabetes, in which there is failure of self-tolerance. The study cohort consists of families identified through several Dermatology Clinics at Texas Medical Center, Houston, TX. Presently, a total of 118 multi-generation families with at least one affected

member have been collected. We selected 4 candidate genes (CG) for this analysis: HLA Class II DR and DQB, T-cell differentiation antigen CD8 and CGRP (calcitonin gene-related peptide). These CG are located on 6p21.3, 2p12 and 11p15.1-p15.2, respectively. We used the transmission disequilibrium test (TDT) to test for association between alleles at specific candidate loci and risk of AA in 204 nuclear families. The results showed an association between the alleles of HLA DQB (p= 0.00033) and DR (p=0.00025) but not for CD8 (p=0.2) and CGRP (p=0.80).

29

Localizing Small Gene Effects in Complex Diseases

M. de Andrade
Biostatistics, Mayo Clinic/Foundation, Rochester, MN

After performing a genome wide screen to identify genes responsible for complex diseases, researchers typically focus on the largest LOD scores, excluding genes with small effects due to their lower LOD scores. But, these genes could jointly provide an interesting finding, if they belong to the same physiological pathway that causes the disease under study. Quantitative measures or traits can represent this disease. We develop a statistic test that uses all information provided by these small genes effects. Presently, this test is applicable for one quantitative trait and multiple genes. With this test we have more power to detect multiple genes involved in the same pathway. We use the variance component approach for quantitative trait as the method to detect linkage. Results using published data will be presented.

30

Fetal Growth and the Heritability of Fasting Insulin Concentrations in Healthy Adults

E. Demerath, S. Czerwinski, R.M. Siervogel, and B. Towne
Wright State University School of Medicine, Dayton, OH

The "fetal programming" approach to the study of chronic diseases has focused attention on aspects of early growth and development in the etiology of diabetes. Numerous studies have shown a relationship between intrauterine growth retardation and the onset of insulin resistance in adulthood. However, pre-natal rate of growth, as well as traits related to glucose tolerance in adulthood have significant genetic determinants that have been disregarded in most epidemiological investigations. Using a maximum likelihood variance components approach, we examined the heritability of fasting insulin concentrations in healthy adults and tested whether or not weight at birth had any significant effect after adjusting for age, sex, and BMI. The sample included 460 individuals aged 18 to 60 years (mean age 37 years) in 115 kindreds from the Fels Longitudinal Study. Fasting insulin concentration was significantly heritable ($h^2 = 0.21 \pm 0.10$), and age, sex, and BMI explained a further 25% of the variation in the trait. In addition, as predicted by the "fetal programming" hypothesis, birth weight was negatively associated with fasting insulin ($p = 0.03$) such that adults with higher fasting insulin were smaller at birth than those with lower fasting insulin. However, birth weight explained only 1% of the total variation in fasting insulin. In conclusion, while we found evidence for an effect of birth weight on insulin levels in adulthood, the magnitude of this effect is small compared to the large effect of known risk factors such as age and BMI.

31

Transformation of Left Ventricular (LV) Phenotypes in a Genome Scan in the NHLBI's Framingham Heart Study (FHS)

A.L. DeStefano[1], E.J. Benjamin[1,2], M.L. Larson[1,2], C.J. O'Donnell[2,3], R.S. Vasan[1,2], L.A. Cupples[1], R.H. Myers[1], and D. Levy[1,2,3]
[1]Boston University School of Medicine; [2]The Framingham Heart Study; [3]NHLBI

LV hypertrophy is an independent risk factor for congestive heart failure and other cardiovascular diseases. Linkage analysis was conducted on two moderately heritable LV phenotypes: end-diastolic LV internal diameter (LVID) (h^2=.34); LV wall thickness (LVWT) (h^2=.23). A 10cM scan was performed in 329 pedigrees from the FHS (n=1268 with both genotype and phenotype data). LVID and LVWT were determined by M-mode echocardiography. Exclusion criteria were prior myocardial infarction, significant valvular heart disease or an inadequate echocardiogram. Sex-specific residuals for LV phenotypes were calculated, adjusted for age, height, weight and systolic blood pressure. Because the residuals for the traits exhibited varying degrees of skewness (LVID=.5; LVWT=1.2) and kurtosis (LVID=2.5; LVWT=4.4), indicating a potential violation of the normality assumption in variance component linkage analysis, linkage was also assessed for normalized residuals and residuals from 1n transformed traits. Maximum multipoint LOD scores obtained using SOLAR on the untransformed traits were: 1.6 on chr 10 at 4cM for LVID; 2.1 on chr 3 at 10cM for LVWT. For LVID there was minimal change in magnitude of the maximum LOD on chr 10 with either transformation. Transformation of LVWT decreased the LOD on chr 3 (LOD=1.8 for 1n (LVWT); LOD=1.5 for normalized LVWT). A location on chr 22 at 10cM emerged with LOD=1.6 for 1n (LVWT)

and LOD=2.3 for normalized LVWT. For the untransformed LVWT LOD=1.5 at the chr 22 location. Although moderate, these LODs are in the range often used as a criterion for follow up (e.g. LOD=2.0). These results highlight the effect of phenotype distribution on linkage results and indicate that transformation of non-normal traits should be considered.

32

No Association Between Alpha-2-Macroglobulin and Alzheimer Disease in Texans

E. Drigalenko, S.E. Poduslo, X. Yin, and R.B. Schiffer
Texas Tech University Health Sciences Center, Division of Neurology, Lubbock, TX

Two mutations in the gene for alpha-2-macroglobulin (A2M) on chromosome 12 have been reported to be associated with increased risk Alzheimer disease (AD) in population studied under NIMH Genetics Initiative (Blacker et al., 1998) and in another sample (Liao et al., 1998).

We have studied 89 Caucasian families from our DNA Bank, each ascertained via a proband resident in Texas (Drigalenko et al., *Genetic Epidemiology* 17 208, 1999). The sample includes 276 sibs, among them 130 affected and 146 unaffected. Parents were unavailable. The A2M polymorphism was genotyped for each subject in accordance with previously published laboratory method. The sib transmission/disequilibrium test was used to study genetic association between the A2M polymorphism and the affected state (Spielman and Ewens 1998).

We confirmed the association of AD and A2M in the original NIMH data (17 families having 71 sibs, among them 51 affected and 20 unaffected), but we found no significant association in our Texas sample. Stratification of families by presence of ApoE4 allele of apolipoprotein E (chromosome 19) did not change the result.

Supported by funds from the State of Texas.

33

A Randomization Test for Meta-Analysis

C. Etzel[1] and R. Guerra[2]
[1]The University of Texas MD Anderson Cancer Center, Houston, TX; [2]Southern Methodist University, Dallas, TX

Meta-analysis is becoming an important tool in linkage analysis. Pooling results across linkage studies allows for an increase in the precision of estimation of genetic effects, and hence stronger conclusions, relative to small, individual studies with low statistical power. We develop a meta-analysis procedure to combine and evaluate evidence for linkage over several simulated studies, which report Haseman-Elston statistics for linkage to a quantitative trait locus (QTL) at multiple, possibly distinct markers on a chromosome. We also observe an increase in precision to estimate both the location of the QTL and the magnitude of the genetic effect. We further propose a randomization test to construct point-wise and chromosome-wise threshold values with which to conduct the meta-analysis procedure. When compared to the meta-analysis results obtained using a standard normal critical value, the randomization test reduces the rate of false-positives (false detection of linkage to the QTL) although minor loss of power to detect linkage is observed. This method also provides useful information concerning the location of a QTL and its genetic effect.

34

Generalized Statistics for Testing Haplotype Frequency Differences Between Groups

Daniele Fallin[1], Laurent Essioux[2] and Nicholas J. Schork[1,2,3]
[1]Case Western Reserve University, Cleveland OH, USA; [2]Genset Corp, La Jolla, CA, USA & Paris, France; [3]Harvard University, Boston, MA

Statistical tests for assessing differences in the frequencies of sequences, haplotypes or multilocus genotypes between groups of individuals defined by an outcome measure is fundamental to genetic association studies. Differences between groups of individuals at the DNA level can present themselves in subtle ways, as they are likely the result of the influence of many factors such as the number of variants in that genomic region that impact the phenotype of interest (allelic heterogeneity), recombination and mutation rates, and the variant-specific penetrance. Statistical tests for detecting such differences should thus be as flexible and sensitive as possible. In this paper we consider statistical tests of the equality of DNA sequence configurations across defined groups using a class of tests for multinomially distributed observations known as the 'power-divergence' statistics. The utility of these statistics is rooted in their ability to detect different types of departure from the hypothesis of equality of sequence configuration frequencies across groups. They are thus more powerful as a class of tests than any single test statistic. We examine the utility of the proposed tests in a number of applications and emphasize their flexibility and ability to detect a wide variety of genetic effects. The proposed tests can accommodate multiple group comparisons, epistasis, highly polymorphic loci, and special study designs such as the transmission disequilibrium test setting. We also comment on limitations of the proposed tests as well as areas of further research.

35 [Invited Speaker]

Maximum Likelihood Inference for Linkage Disequilibrium Mapping

Joe Felsenstein, Mary K. Kuhner, Jon Yamato, and Peter Beerli
Department of Genetics, University of Washington, Box 357360, Seattle, WA 98195-7360

While much attention has been paid to measures of linkage disequilibrium, there has been little progress in computing the joint distribution of all the disequilibria among a set of loci in a sample from a population. That is what is needed to do maximum likelihood inference in linkage disequilibrium mapping. The problem is easily shown to be the same as computing the distribution of the haplotype frequencies in a sample from a population. This can be done using coalescent likelihood methods, which use random sampling to approximate the likelihoods, averaging over the unknown coalescent tree connecting the sampled haplotypes by sampling from a space of possible trees. Griffiths and Marjoram (J. Comp. Biol., 1997) and ourselves (Kuhner et. al., submitted) have implemented different sampling methods aimed at calculating this likelihood. The implementation using a Markov Chain Monte Carlo (MCMC) method employing a Metropolis-Hastings sampler will be described, including methods of computing likelihoods for diverse positions of a locus from one run of the sampler, as well as correcting for ascertainment of the locus and uncertainties of haplotyping.

36

A Mixture Analysis Approach for Addressing Multiple Comparisons, with an Application to Gene-Expression Data

B.A. Fijal and J.S. Witte
Case Western Reserve University, Cleveland, OH

The rapid increase in the amount of genomic data available to genetic epidemiologists is stimulating novel research into the pathogenesis of disease. This increase, however, will undoubtedly compound the existing statistical and philosophical issues surrounding multiple comparisons. For example, although the advent of microarray technology allows the expression levels of thousands of genes to be compared, the statistical interpretation of ensuing results remains somewhat problematic. Traditional methods of adjusting for multiple comparisons control the false positive rate at the expense of power. For instance, one of the most common approaches for addressing multiple comparisons is the Bonferroni correction in which the comparison-specific α-level (e.g., 0.05) is divided by the number of comparisons to obtain an "adjusted" α-level. While the simplicity of this approach makes it attractive, the Bonferroni correction can ultimately lead to high false negative rates. Therefore, we present here a mixture analysis approach for selecting critical values when making multiple comparisons. This is a decision theoretic approach that allows one to estimate and control false positive and false negative rates. We evaluate the performance of this method via simulation of gene-expression array data in which the expression levels of several thousand genes are compared in two cell lines.

37

Statistical Models of Prediction of Tolerance to Alcohol

L. Flury, T. Foroud, V.A. Ramchandani, S. Morzorati, D. Kareken, T. Blekher, and S. O'Connor
Indiana University School of Medicine, RLR VAMC, Indianapolis, IN

Human brain function adapts to the presence of moderate levels of alcohol. This study examines various statistical models predicting acute tolerance to alcohol from data obtained during brain exposure to alcohol at a constant breath alcohol concentration (BrAC).

Alcohol was infused intravenously to 44 subjects to evaluate the effects on 58 variables from 8 subjective, ERP, ocular motor, and neuropsychological tasks. All variables were measured at baseline (b_0), immediately after a target BrAC of 60%mg was achieved (b_1), and 105 minutes later (b_2) with virtually no interim change in BrAC. An acute tolerance index was computed for each variable as $sign((b_2\text{-}b_1)/(b_1\text{-}b_0))*|b_2\text{-}b_1|$. The sum of the standardized indices defined a tolerance phenotype, which could then be used to predict an individual's overall response to alcohol. Of interest was whether a subset of the 58 variables could predict an individual's overall tolerance to alcohol. Subsets were selected using two methods: (1) robust effects of alcohol on the variable, (2) principal component analysis (PCA), selecting only the first PC from each of the four sets of variables, and using the PCs from the reduced subset in (1). Least squares and logistic (using the top and bottom 35% of the tolerance distribution) regression were employed to identify significant predictors of tolerance.

All models yielded consistent results, with the PC based on the subjective variables or the one subjective variable representing the individual's perception of feeling "high", accounting for 28-43% of the variability. These results suggest a consistent and significant predictor of tolerance ($0.006<p<0.10$) is an individual's perception of the intoxicating effects of alcohol, and that "tolerance" as defined above, is a stable phenotype for further analyses in which family history and genetic components of alcoholism are studied.

38

Joint Oligogenic Segregation and Linkage Analysis of Plasma HDL Levels in Familial Combined Hyperlipidemia Pedigrees

F. Gagnon, G.P. Jarvik, A.G. Motulsky, S.S. Deeb, J.D. Brunzell, and E.M. Wijsman
University of Washington, Departments of Medicine and Genetics, Seattle, WA

Familial combined hyperlipidemia (FCHL) is defined by elevated levels of plasma cholesterol and/or triglyceride. High-density Lipoprotein cholesterol plasma levels (HDL) are also lower in these individuals. Etiologic heterogeneity of FCHL has been suggested. Under these conditions, it is difficult to localize causative genes, and linkage studies of FCHL families have produced inconsistent results for the chromosomal region of APOCIII (11q23). A method of joint oligogenic segregation and linkage analysis (based on Monte Carlo Markov chain methods (MCMC)) provides a novel way to evaluate this chromosomal region. Using this approach, we estimated the number of genes contributing to the variance of HDL levels, as well as the contribution of age, sex, and APOCIII genotype. We used familial data from 183 individuals in 3 FCHL pedigrees. We estimate that 2 to 3 genes contribute to HDL levels, but that none were linked to APOCIII. When sex is added as a covariate, this estimate slightly drops toward 2 loci. Adding age as a covariate had no effect. These results support the involvement of 2 to 3 genes in the control of HDL levels. A locus near APOCIII does not have a role in the regulation of HDL levels in these FCHL families. A complete genomic search, also exploring the genetic influence on HDL subfractions, HDL2 and HDL3, is under way.

39

Gene-Environment Interaction and Affected-Sib Pair Linkage Analysis

W. James Gauderman and Kimberly D. Siegmund
Department of Preventive Medicine, University of Southern California

Objectives: Gene-environment (G×E) interaction influences risk for many complex disease traits. However, genome screens using affected sib pair linkage techniques are typically conducted without regard for G×E interaction. We propose a simple extension of the commonly used mean test and evaluate its power for several forms of G×E interaction. Methods: We compute expected IBD sharing by sibling exposure profile, that is by whether two sibs are exposed (EE), unexposed (UU), or are discordant for exposure (EU). We describe a simple extension of the mean test, the "mean-interaction" test, that utilizes heterogeneity in IBD-sharing across EE, EU, and UU sib pairs in a test for linkage. Results: The mean-interaction test provides greater power than the mean test for detecting linkage in the presence of moderate or strong G×E interaction, typically when the interaction relative risk (R_{ge}) exceeds 3 or is less than 1/3. In the presence of strong interaction (R_{ge}=10), the required number of affected sib pairs to achieve 80% power for detecting linkage is approximately 30% higher when the environmental factor is ignored in the mean test, than when it is utilized in the mean-interaction test. Conclusion: Linkage methods that incorporate environmental data and allow for interaction can lead to increased power for localizing a disease gene involved in a G×E interaction.

40

Genotype Wise Calculation of Detection Rates for Pedigree Errors in SNPs

Frank Geller and Andreas Ziegler
Institute of Medical Biometry and Epidemiology, Philipps University, Marburg, Germany

One well-known approach for the analysis of transmission-disequilibrium is the investigation of single nucleotide polymorphisms (SNPs) with trios consisting of a child and its both parents. If genotypes are given erroneously we speak of pedigree errors.

These pedigree errors can have various reasons that include sample swap or wrong pedigree structure. However, under good study conditions they should be mainly caused by genotyping errors. Here, heterozygous individuals are often ascertained as homozygous because PCR amplification fails to amplify one allele or homozygous genotypes are read as heterozygous due to stutter bands introducing a second nearby allele into the genotype. Some of these errors can be detected by Mendel checks whilst others are compatible with the pedigree structure. The proportion of detected errors is termed detection rate.

Gordon et al. (Hum Hered 1999; 49:65-70) considered the case of pedigree errors that occur randomly and independently with some fixed probability α for the wrong ascertainment of an allele. In practice, however, SNP genotypes are read instead of single alleles. Thus, pedigree errors should be analyzed based on genotypes. In our presentation, we investigate this genotype wise approach.

In contrast to Gordon et al., who reported detection rates between 25 and 30%, we obtain detection rates between 20 and 50%. We conclude that in practice the detection rate may be substantially larger compared with the numbers presented by Gordon.

41

Statistical Properties of the TDT in Presence of Inbreeding

E. Génin[1], A.A. Todorov[2], and F. Clerget-Darpoux[1]
[1]INSERM U535, Le Kremlin-Bicêtre, France;
[2]Washington University School of Medicine, St Louis, MO

Family-based association tests such as the Transmission Disequilibrium Test (TDT) that compare the rates of transmission of alleles from parents to affected offspring are widely used to detect the role of genetic risk factors in diseases. In prior studies of the statistical properties of the TDT, it has always been assumed that the parents of affected offspring were neither inbred nor related. This is not the case in many human populations and it is thus of interest to determine if the TDT is valid in inbred populations.

In this study, we show that, in the presence of inbreeding, alleles non-transmitted to affected offspring provide biased estimates of population allelic frequencies. However, the TDT remains a valid test for linkage and association and in some instances, inbred families will actually provide more linkage information than outbred families. Under a recessive mode of inheritance, selecting families with parents who are both first cousins and offspring of first cousins will reduce the number of trios needed to detect the role of a disease susceptibility locus with a moderate effect (GRR = 2) by a factor of 1.1 to 2.5. Under a dominant mode of inheritance, there is also a noticeable gain in power when the susceptibility allele is frequent.

We provide software to estimate the power of a TDT study for various modes of transmission and degrees of relationship between parents of affected offspring.

42

A Bayesian Approach to the Transmission Test for Linkage Disequilibrium

V. George and P.W. Laud
Medical College of Wisconsin, Milwaukee, WI, USA

The transmission/disequilibrium test (TDT) proposed by Spielman et al (Am J Hum Genet 52:506-516, 1993) for binary traits is a powerful method for detecting linkage between a marker locus and a trait locus in the presence of allelic association. The TDT uses information on the parent-to-offspring transmission status of the associated allele at the marker locus to assess linkage or association in the presence of the other, using one affected offspring from each set of parents. For testing for linkage in the presence of association more than one offspring per family can be used. However, without incorporating the correlation structure among offspring, it is not possible to correctly assess the association in the presence of linkage. In this presentation, we propose a Bayesian TDT method as a complementary alternative to the classical approach. Relationships between the recombination fraction (θ), disequilibrium coefficient (δ), allele frequencies and the observations allow evaluation of the full likelihood of the data. Using priors constructed from historical information for allele frequencies and vague information on θ and δ (conditional on $\delta>0$), joint and marginal posterior distributions for θ and δ are obtained via standard MCMC methods. These lead naturally to Bayesian intervals for both parameters. This method can easily be extended to quantitative traits, and more complex pedigrees.

43

Evaluating the Power and Validity of a Two-Stage TDT for Quantitative Traits in Pedigree Data

V. George[1], H.K. Tiwari[2], and R.C. Elston[2]
[1]Medical College of Wisconsin, Milwaukee, WI;
[2]Case Western Reserve University, Cleveland, OH

Recently, George et al. (Am J Hum Genet 65:236-245, 1999) proposed a regression-based transmission/disequilibrium test (TDT) for linkage between a marker locus and a quantitative trait locus in pedigree data. This method allows for the correlation structure among the members of the pedigree and incorporates other predictors and confounders in the model. All TDT methods are valid tests for linkage, but they only have power in the presence of population association. We have extended this method to a two-stage procedure in which association and linkage are tested sequentially. At the first stage, a test of population association is performed using a regression model with the presence or absence of the marker allele in all individuals in the sample as the independent variable. If a significant association is found, a regression-based TDT is performed using only the informative offspring in the sample. In this presentation, we evaluate the validity and the power of the two-stage procedure through simulation, for various combinations of disequilibrium coefficient and recombination fraction. The simulation is repeated for the case of the two, four, eight and sixteen alleles at the marker locus.

44

Deciphering the Genetic Architecture of a Multivariate Phenotype

Saurabh Ghosh and Partha P. Majumder
Anthropology and Human Genetics Unit, Indian Statistical Institute, Calcutta

A heritable multivariate quantitative phenotype comprises several correlated component phenotypes that are usually pleiotropically controlled by a set of major loci

and environmental factors. One approach to decipher the genetic architecture of a multivariate phenotype, in particular, to map the underlying loci, is to reduce the dimensionality of the data by a data-reduction technique, such as principal component analysis. The extracted principal components are then analyzed in conjunction with marker data to map the underlying loci. We have examined the efficiency of this approach with and without taking the correlation structure of the multivariate phenotype when extracting principal components. We have assumed that genome-wide scan data on sib-pairs are available for low-density (widely-spaced) and high-density markers. Using extensive simulations, based on three models of the multivariate phenotype, we have shown that although the ignoring of the correlation structure of the multivariate phenotype does not have any serious impact on the efficiency of mapping the underlying trait loci in wide marker intervals, there is a significant adverse effect of this for fine-mapping. We, therefore, recommend that the correlation structure of the multivariate phenotype be carefully examined to decide on the strategy of extracting principal components for deciphering the genetic architecture of the multivariate phenotype.

45 [Presidential Address]

Genetic Epidemiology: 2001 and Beyond

Lynn R. Goldin
National Cancer Institute, Bethesda, MD

Now that we are about to begin a new millennium, it is worth considering what genetic epidemiology will look like over the next few years. New high throughput technologies in genomics are providing us nearly unlimited genetic information about individuals that will enhance our ability to identify genes for complex diseases and elucidate underlying mechanisms, including the role of environmental factors. A typical genetic epidemiological study of the future may include features such as: 1) large sample of cases, population controls, and family controls who have been genotyped for 500,000 SNPs, evaluated for environmental risk factors, and measured for other relevant biological traits 2) sample of multiplex families with relevant phenotypes and genotyped at a density of 1 cM and 3) expression levels for a few thousand "candidate" genes in cases.

These emerging technologies provide many challenges to the genetic epidemiologist. What new analytic approaches are needed? What are the limitations of these studies? How interdisciplinary do we need to be? IGES is in a strong position to take a leadership role in moving the science forward by continuing to build bridges among our related disciplines, promoting training opportunities for genetic epidemiologists, and participating in the public debate about the strengths and limitations of genetic studies in improving public health.

46

Comparison of Case-Control Designs Using Unrelated or Related Controls for Detecting Gene-Environment (G×E) Interaction

A.M. Goldstein[1] and N. Andrieu[2]
[1]NCI, Bethesda, MD, USA; [2]U521 INSERM, Villejuif, France

Two types of control groups are generally used for examining G×E interactions: unrelated (e.g., population-based) or related (e.g., sibling) controls. Previous work has shown that relative controls have good power for detecting interaction when risk factors are rare (Witte et al., 1999). However, some of these studies may have assumed unrealistic numbers of relative controls. Limited reviews (e.g., of breast and stomach cancers) have suggested that about 50% of cases may have an available sibling control. We, thus, compared a 1:0.5 case-sibling control design to 1:0.5 or 1:1 unrelated control designs to determine the design that maximizes power for reasonable effect estimates (RR_E=1,3,5,10; RR_G=1,2,3,5,10; RR_{int}=3,5,10) according to different G and E frequencies.

Our results show that when the risk factors of interest are common (> 0.2), the unrelated controls are more efficient than the sibling controls for detecting interaction. For rare genetic factors, the sibling controls are more efficient. For example, when the genetic factor of interest is ≤ 0.05, the case-sibling control design is more efficient than the 1:1 case-unrelated control design. When the frequency of the genetic factor is 0.1 or 0.2, then the case-sibling control design is more efficient than the 1:0.5 case-unrelated control design but less efficient than the 1:1 case-unrelated control design. When one or both risk factors are rare, the case-sibling control design is always more efficient than the 1:0.5 case- unrelated control design and often more efficient than the 1:1 case-unrelated control design for detecting interaction.

Thus, the choice of related or unrelated controls will depend mainly on the prevalence of the risk factors and their effect estimates. An alternative design that collects both controls might improve power for G×E interaction detection and allow for the widest range of investigations.

47

Analysis of Prostate Cancer Loci Considering Clinical Characteristics

E.L. Goode[1], J.L. Stanford[2], M. Gibbs[2], M. Janer [3], M. Peters[2], S. Kolb[2], M.D. Badzioch[1], L. Hood[3], E.A. Ostrander[2], and G.P. Jarvik[1]
[1]University of Washington; [2]Fred Hutchinson Cancer Research Center; [3]Institute for Systems Biology, Seattle, WA

Hereditary prostate cancer is a heterogeneous complex disease with 4 susceptibility loci mapped to date. We

assessed linkage to multiple markers at *CAPB* (1p36), *HPC1* (1q24-25), *PCaP* (1q42.2-43), and *HPCX* (Xq27-28) in 149 families with 3 or more living affected men using clinical data from medical records of 505 affected men. LINKAGE, HOMOG, and GENEHUNTER were used to examine linkage and heterogeneity. Overall, the following maximum 2-point lods were seen: 0.86 (θ=0.18) at *CAPB* (D1S407), 0.43 (θ=0.24) at *HPC1* (D1S1660), 0.57 (θ=0.26) at *PCaP* (D1S2785), and 0.16 (θ=0.34) at *HPCX* (DXS984). To increase homogeneity, families were stratified by race, median age of diagnosis, distribution of stage and grade, and lod scores at other loci. A maximum NPL score of 1.83 (p=0.04) was seen at *CAPB* (D1S407) for 37 white families with at least 1 high-grade cancer; this rose to 2.34 (p=0.01) removing families with evidence for other linkage. The most suggestive *HPC1* results were for 42 white families with median age $\geq$ 70 years, with a peak NPL of 1.59 (p=0.06, D1S1660) and lods and NPLs > 0 at all markers. A peak NPL of 1.03 (p=0.15) was seen at *PCaP* (D1S1609) for 26 white families with moderately differentiated tumors and no evidence for other linkage. A maximum 2-point lod of 1.14 (θ=0) was seen at *HPCX* (DXS1193) for 33 white families with median age 60-64 years. Considering such clinical data may increase our understanding of the multiple loci responsible for hereditary prostate cancer.

48

Joint Analysis of Linkage and Linkage Disequilibrium on Multiallelic Marker Data: Reducing the Number of Degrees of Freedom

H.H.H. Göring[1] and J.D. Terwilliger[2]
[1]Southwest Foundation for Biomedical Research, San Antonio, TX; [2]Columbia University and NY State Psychiatric Institute, New York, NY

Linkage analysis and linkage disequilibrium (LD) analysis are two related techniques used to localize genes. We have recently proposed a "model-free" approach to joint analysis of both phenomena using a hypothetical "pseudomarker" locus. The underlying LD model was completely general, allowing for differences in the frequencies of all marker alleles between affected and unaffected individuals. The approach was shown to be more powerful than the transmission disequilibrium test (TDT) in a wide variety of situations when a diallelic marker is analyzed. If the marker has many alleles, however, this may no longer be the case due to the increase in the number of degrees of freedom (df). If one assumes the existence of only a single ancestral disease allele of which clonal copies are present on some of the chromosomes of affecteds in the dataset, only a single marker allele is expected to be over-represented in affecteds relative to unaffecteds. By making this assumption, it is possible to reduce the number of df to 1 in LD analysis, no matter how many marker alleles there are (e.g., Terwilliger, Am J Hum Genet 56:777, 1995). We have incorporated a similar model into "model-free" joint linkage and LD analysis on nuclear pedigrees, trios and singletons combined. We have found that this approach can be more powerful for analysis of multiallelic markers than the completely general approach as well as the TDT. The same technique is also applicable to "model-based" analysis as well as analysis on multigenerational pedigrees.

49

Regression Modeling of Sibpairs and Affected Relative Pairs Used to Identify Linkage and Epistasis in Families Multiplex for Systemic Lupus Erythematosus

C. Gray-McGuire[1], K.L. Moser[1,2], J. Kelly[1], J. Olson[2], C.M. Jedrey[2], K.B. Jacobs[2], B.R. Neas[1], J.B. Harley[1]
[1]Oklahoma Medical Research Foundation, Oklahoma City, OK; [2]Case Western Reserve University, Cleveland, OH

Systemic lupus erythematosus (SLE) is a complex autoimmune disorder involving at least hormonal, environmental, and genetic factors. High heritability, high monozygotic twin concordance rates, familial aggregation, association with candidate genes, and the results of five genome scans demonstrate support for a genetic component. We present here the results of a genome scan of 126 pedigrees multiplex for SLE. Using the revised multipoint Haseman-Elston regression technique for concordant and discordant sib pairs (SIBPAL2) and a conditional logistic regression technique for affected relative pairs (ARP), we identify four new candidate linkages significant at p<0.01, as well as identify a significant epistatic interaction between candidate linkage regions on chromosome 4p16-15 and 5p15 in European Americans (variance components = 0.047 and 0.000 compared to 0.356 for the interaction). We present three regions for which the results agree with a previously published subset of this data, including an effect on chromosome 4p16-15 (p=0.0003, and lod=3.84) in the European Americans and chromosome 1q22-24, the region containing FcγRIIA (p=0.001, and lod=2.47) in the African-Americans. We confirm the published results of four independent studies for nine regions, including an effect on chromosome 6p22-21, a region that contains several HLA components. We discuss the role of ethnicity, clinical heterogeneity, and method of analysis in determining the genetic architecture of SLE.

50

Model-Free Linkage Analysis Incorporating Disease Epidemiology

C.M.T. Greenwood[1] and K. Morgan[1,2]
[1]Departments of Human Genetics and [2]Medicine, McGill University, Montreal

For common diseases, reliable estimates of disease prevalence or incidence are often available for different demographic subgroups of individuals. For example, disease prevalence may be known for different age groups, or for individuals exposed to a particular risk agent. When searching for evidence for genetic linkage, this information should be incorporated into the analysis since it can increase the power to detect linkage. An algorithm is developed which fits a 'model-free' linkage analysis incorporating known disease epidemiology through the use of prevalence estimates in different liability classes. The approach is based on an extension of the 'model-free' approach of Curtis and Sham (1995) [Am J Hum Genet 57:703-716], where the likelihood is maximized over penetrances consistent with the known prevalence, under linkage and no linkage separately. In the extended algorithm, liability classes are defined based on the disease epidemiology, and the constrained maximization explores the parameter space between no genetic effect and a predefined maximum genetic effect for all classes. The maximum effect can be chosen, or calculated from estimates of sibling risk. Following Curtis and Sham, the search space is restricted to dominant or recessive models with ordered penetrances. Two data sets are used to demonstrate the performance of this algorithm. Also, a small set of simulations shows that this robust analysis method performs well, and is a good alternative to some commonly used robust or model-free approaches that do not use epidemiologic information. CMTG is supported by a postdoctoral fellowship from MRC.

51

A Generalized Regression Procedure for Mapping Complex Diseases by Linkage-Disequilibrium and Population-Based Samples

C. Gu[1], D.C. Rao[1], and M. Xiong[2]
[1]Division of Biostatistics, Washington University School of Medicine, St. Louis; [2]Human Genetics Center, University of Texas, Houston

Nonrandom association observed between alleles of genetic markers and putative disease loci is a result of many factors including linkage-disequilibrium (LD) due to mutation, migration and population admixture. It has been argued that designs using population-based unrelated individuals may have higher power and resolution than using data observed on pedigrees in some cases, because of the large number of recombinations accumulated in the general population. However, for the LD method to be practically useful in mapping complex diseases, the many non-genetic factors influencing the observed association must be differentiated from the genetic ones. It has been also increasingly recognized that the relationship between phenotype and genetic and environmental factors may be nonlinear. We introduce a general regression procedure in which the genetic effects contributing to a complex trait are modeled through a generalized link function f: $E(y|x) = f(\mu + g_A e)$, where g_A is a vector of genotypic values of the putative genes and e a vector for environmental effects. This modeling is general enough to allow for interactions between the LD effects and environmental factors, and nonlinear choices of f can be used to accommodate traditional logistic regression for categorical trait outcomes and survival models for age-dependant effects. A step-wise procedure is implemented to select variables of significant influences on the observed association and the underlying LD effects. Fitness to certain genetic models can be tested using likelihood ratio tests. It can be shown that this method is equivalent to the classical regression method when the link function is linear.

52

Why are We Weighting? The True Type I Error Rate of Likelihood-Ratio Affected Sib-Pair Analysis on Multiplex Sibships

Peter Holmans
Washington University, St Louis MO

"All-pairs" likelihood-ratio analyses, such as those performed by MAPMAKER/SIBS [Kruglyak & Lander 1995], require that a sibship containing N affected siblings be split into N(N-1)/2 sibships, each containing a different pair of affected sibs, before analysis. Each of these N(N-1)/2 sibships may also contain the other affected sibs from the original sibship, coded as unaffected, to infer missing parental genotypes, as is done automatically in MAPMAKER/SIBS. Then, the use of the same individuals both as affecteds to test for linkage and, elsewhere, as unaffecteds to infer missing parental genotypes leads to negative correlations in the estimated IBD sharing among affected pairs from the same original multiplex sibship. This gives a conservative test of linkage, even when no downweighting is applied. Conversely, if the other affected sibs from the original sibship are omitted, the correlations are positive and the linkage test is anti-conservative in the absence of weighting, as suggested by Greenwood and Bull [1999]. True Type I error probability also depends on numerous other factors, suggesting the use of simulation, rather than asymptotic theory, to assess significance levels.

The 2/N weighting proposed by Suarez & Hodge [1979] was found to increase power over an unweighted analysis in many situations, provided significance levels were adjusted appropriately by simulation.

53

A Genome Scan for Susceptibility Loci to Lipids in the Amish

W.-C. Hsueh[1], B.D. Mitchell[1], J. Schneider[1], M.J. Wagner[2], P.L. St. Jean[2], C.J. Bell[3], and A.R. Shuldiner[4]
[1]Southwest Foundation for Biomedical Research, San Antonio, TX; [2]Glaxo Wellcome Inc., RTP, NC; [3]Axys Pharmaceuticals, La Jolla, CA; [4]University of Maryland School of Medicine, Baltimore, MD

Serum lipid levels are significantly associated with the risk of CVD. To delineate the genetic influence in lipid variation, we conducted heritability and linkage analyses of several serum lipid traits, including the fasting total serum cholesterol (TSC), high-density lipoprotein cholesterol (HDL-C) and triglyceride (TG), in a founder population, the Old Order Amish. These lipid traits were significantly heritable ($p<0.0001$), the heritability estimates for TSC, HDL-C, and TG are 0.63±0.10, 0.56±0.08, and 0.36±0.09, respectively.

We performed a genome scan using a multipoint variance components procedure to detect susceptibility loci for these lipid traits. A set of 357 polymorphic markers were genotyped in 694 subjects ascertained from 28 families enriched with type 2 diabetes. The significances (p values) of all nominal LOD scores were verified by simulation studies, then converted back to LOD scores. The highest LOD score we observed was 1.90, reflecting a linkage of TG to a region on chromosome 11p (137cM). LOD scores of 1.5 ($p<0.005$) or greater were also observed in 4 other chromosomal regions as shown in the table. Signals for these 3 traits appeared to be in different regions. Despite of moderate heritability, no significant evidence for linkage to lipid traits was observed, suggesting the complexity of genetic influence on the lipid metabolism.

Chr	cM	Trait (LOD score)	Candidate Gene
2	32	TSC (1.58)	POMC, POB
3	25	TSC (1.75)	PPARG
6	130	HDL (1.47)	CYP51
10	48	Ln (TG) (1.58)	
11	137	Ln (TG) (1.90)	APOC3,APOA1
19	40	TSC (1.50)	

54

Admixture Association in Genetic Epidemiology

J.J. Hottenga[1,2], J.J. Houwing[1], L.A. Sandkuijl[1,2], and C.M. van Duijn[1]
[1]Department of Epidemiology & Biostatistics, Erasmus University, Rotterdam, The Netherlands; [2]Department of Human Genetics, Leiden University, The Netherlands

The genetic association study is a powerful tool to study candidate genes via case-control design. Like all case control studies however, it is susceptible to confounding leading to false positives. One form of confounding occurs when the cases and controls are sampled from two or more sub-populations that differ with respect to allele frequencies. Association can then be found even if there is no causal relationship between the DNA variant and the disease. The aim of the present study is to quantify the problem of admixture by means of simulation. In this study two populations were simulated - one of which was a large population with stable allele frequencies, while the other population was relatively small and therefore susceptible to genetic drift. With the allele frequencies from the simulated populations odds ratio's were calculated under various conditions of admixture. Given an odds ratio threshold, the probability of finding a positive association was calculated. Our results show that only major differences in allele frequencies will lead to spurious admixture association. Such pronounced differences will only be observed if a sub-population has been exposed to a high level of drift due to prolonged isolation.

55

The Null Distribution of the Heterogeneity Lod Score (HLOD) Does Depend on the Assumed Genetic Model for the Trait

J. Huang, V.J. Vieland, and Kai Wang
University of Iowa, Iowa City, IA

It is well-known that the asymptotic null distribution of the homogeneity lod score (LOD) does not depend on the genetic model specified in the analysis [Williamson & Amos, 1990]. When appropriately rescaled, the LOD is asymptotically distributed as a $.5\chi_0^2+.5\chi_1^2$, regardless of the assumed trait model. However,because locus heterogeneity is a common phenomenon, the HLOD [Smith, 1961] is often used in gene mapping studies rather than the LOD itself. We show here that, in contrast with the LOD, the asymptotic null distribution of the HLOD does depend upon the genetic model assumed in analysis. In affected sib pair (ASP) data, this distribution can be worked out explicitly as $(.5 - c)\ \chi_0^2+.5\chi_1^2+c\ \chi_2^2$, where c depends on the assumed model. E.g., for a simple dominant model (HLOD/D), c is a function of the disease allele frequency p: for p=0.001, c=0.000056; while for p=0.1, c=0.0059. For a simple recessive model (HLOD/R), c=0.098 independently of p. We show also that this latter (recessive) distribution turns out to be the same as the asymptotic distribution of the MLS statistic [Risch, 1990] under the possible triangle constraint [Holmans, 1993], which is thus no less "parametric" than the HLOD/R. Note that the null distribution of the HLOD/D is close to that of the LOD, because the weight c on the component

is small. Note also that the cutoff value for a test of size α will tend to be smaller for the HLOD/D than the HLOD/R. For example, the α=0.001 cutoff (on the lod scale) for the HLOD/D with p=0.01 is 2.076, while for the LOD it is 2.074, and for the HLOD/R it is 2.34. For general pedigrees, explicit analytical expression of the null HLOD distribution does not appear possible, but it will still depend on the assumed genetic model.

56

Adjusting for the Non-Independence of Pairs in the Conditional Logistic Model for Affected-Relative-Pair Linkage

K.B. Jacobs and R.C. Elston
Case Western Reserve University, Cleveland, OH

The Olson (*Am. J. Hum. Genet.* 65:1760-1769, 1999) conditional logistic model for affected-relaive-pair (ARP) linkage analysis provides a general and flexible method to test for genetic linkage under a variety of multi-locus models using information from all affected relative pairs. However, the model as proposed assumes independence of all relative pairs, which will not be the case for data from extended pedigrees, or, indeed, from sibships of size greater than two. Cordell et al. (*Am. J. Hum. Genet.* 66:1273-1286, 2000) note this will inflate type 1 error and recommend that significance levels be calculated by use of simulation. We confirm this by large-scale simulation over various pedigree structures and compute the null distributions. We then discuss several possible simulation methods and their computational and statistical effectiveness to compute empirical estimates of significance levels. While effective, simulation is computationally expensive and can be infeasible in some situations. Therefore, a weighting scheme is developed to approximately correct for the non-independence of pairs due to allele sharing at a single locus so that simulation is unnecessary.

57

Measurement of Paraoxonase Phenotype is Required to Detect *PON1-192* or *PON1-55* Genotype Effects in Carotid Artery Disease

G.P. Jarvik, L.S. Rozek, V.H. Brophy, T.S. Hatsukami, R.J. Richter, N.T. Tsai, G.D. Schellenberg, and C.E. Furlong
The University of Washington Department of Medicine (Medical Genetics), Seattle

Two paraoxonase (PON1) polymorphisms, *PON1-Q192R* and *PON1-L55M* have been inconsistently associated with vascular disease. However, plasma PON1 activity phenotypes vary markedly within genotypes. Thus, activity was expected to add to the informativeness of genotype for predicting vascular disease. The case-control study included 212 age and race matched men with mean age 66.4 yr. (range 49-82 yr.); 95% were Caucasian. The 106 carotid artery disease (CAAD) cases had >80% carotid stenosis and the 106 controls had <15% stenosis. Two PON1 substrate hydrolysis rates (paraoxon, POase; diazoxon, DZOase) were significantly lower in cases than in controls and were significant predictors of CAAD using logistic regression (POase, p=0.005, 25% reduced; DZOase, p=0.019, 16% reduced). POase and DZOase were both significant when included in the same model. DZOase predicted vascular disease independently of lipoprotein profile, HDL subfractions, apoAI, and smoking. The marginal effects of *PON1-192* (p=0.75) and *PON1-55* (p=0.83) genotypes or haplotype (0.70) did not predict case-control status. However, when phenotype was included as a predictor both *PON1-192* and *PON1-55* genotypes *or* haplotypes were significant predictors at the 0.05 level. The common methodology of examining *PON1-192* and/or *PON1-55* genotypes alone may lead to the erroneous conclusion that there is no PON1 role in CAAD. This may have broad implications for the utility of the "genotype only" approach. These results support the benefit of a "level crossing" approach that includes intervening phenotypes in the study of complexly inherited disease.

58

A New Hardy-Weinberg Disequilibrium Measure for Fine-Scale Genetic Mapping

R.F. Jiang[1], J.P. Dong[1], D. Wang[2], and F.Z. Sun[2]
[1]Michigan Technological University, Houghton, MI;
[2]University of Southern California, Los Angeles, CA

Hardy-Weinberg disequilibrium (HWD) among affected individuals has recently been proposed for fine-scale mapping of disease-susceptibility genes. We investigate the statistical properties of several available HWD measures and propose a new HWD measure for fine-scale mapping. We show both theoretically and through simulations that the available HWD measures depend not only on the genetic distance between the marker locus of interest and the disease-susceptibility locus, but also on the allele frequencies at the marker locus. On the other hand, the new measure is not affected by the allele frequencies at the marker locus under the assumption of initial complete association between the marker and the disease loci and no new mutations at the marker and the disease loci. The new measure is robust if the mutation rates at the marker and the disease loci are low while the new measure may not be useful if there are incomplete associations between the marker locus and the disease locus. We also propose a method to estimate the location of the disease-susceptibility

gene. The power of this method is comparable to simple linkage disequilibrium mapping using P_{excess} for case-control studies. Thus, HWD provides a new method for fine-scale mapping.

59

The Score Statistic of the LD-LOD Analysis: Detecting Linkage Adaptive to Linkage Disequilibrium

Yanming Jiang and Jian Huang
Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA

Due to rapid advancements in molecular technology, it is becoming feasible in the near future to use the single-nucleotide polymorphism (SNP) markers in a genome screen. With a dense SNP map, some of the markers could be in linkage disequilibrium (LD) with the disease-redisposing alleles. Significant LD may also result from other factors such as admixture of two or more subpopulations differing in allele frequencies. However, in practice, the extent of LD between a marker and disease is usually unknown. Therefore, it is of interest to develop linkage analysis methods that are adaptive to LD. We study the properties of a modified LOD score method for testing linkage that incorporates LD (LD-LOD). By examination of its score statistic, we show that the LD-LOD score method adaptively combines two sources of information: (a) the IBD sharing score which is informative for linkage regardless of the existence of LD; and (b) the contrast between allele-specific IBD sharing scores which is informative for linkage only in the presence of LD. In particular, we show that, for triad data, a simple recessive LD-LOD test is symptotically equivalent to the TDT; and for ASP data, it is an adaptive combination of the TDT and the ASP mean test. We demonstrate that the LD-LOD score method has relatively good statistical efficiency in comparison with the ASP mean test and the TDT for a broad range of LD and the genetic models we considered. Therefore, the LD-LOD score method is an interesting approach for detecting linkage when the extent of LD is unknown.

60

A Follow-Up Linkage Study Supports Evidence for a Bipolar Affective Disorder Locus on Chromosome 21q22

Suh-Hang Juo[1], Jianjun Liu[1], Joseph D. Terwilliger[1], Adina Grunn[1], Xiaomei Tong[1], Miguel Brito[1], Jo Ellen Loth[1], Kyra Kanyas[2], Bernard Lerer[2], Jean Endicott, Graciela Penchaszadeh[1], T.Conrad Gilliam[1], and Miron Baron[1]
[1]Columbia University, New York; [2]Hadassah-Hebrew University Medical Center, Jerusalem

Evidence for linkage between bipolar affective disorder and 21q22 was first reported by our group in a single large pedigree with a LOD score of 3.41 at the PFKL locus. In a subsequent study, with dense marker coverage in 40 multiplex BP pedigrees, we reported supporting evidence with a two-point LOD score of 2.76 at the D21S1260 locus, 5 cM proximal to PFKL. For cost-efficiency, the individuals genotyped in that study comprised a subset of our large pedigree sample. To augment our previous analysis, we have conducted a follow-up study including 16 new pedigrees, 3 additional markers (a total of 11 markers in 10 cM), and married-in spouses' genotyping data. The present study analyzed the complete dataset of 56 pedigrees (a total of 860 genotyped individuals vs. the 372 genotyped previously), the largest multigenerational BP pedigree sample reported to date. Affected-only parametric analysis was carried out using a dominant model with the penetrance of 0.8 for gene carriers and 0.01 for phenocopies. Results show that the highest two-point LOD score is 3.12 (θ= 0.25) at D21S1260. Six out of the 11 markers have LOD scores greater than 1. The 16 new pedigrees gave a two-point LOD score of 1.82 (θ=0.1) at D21S266, 2 cM proximal to D21S1260. Our results are consistent with a putative BP locus on 21q22.

61

Ascertainment Corrections in Two-Phase Sampling Designs for Segregation and Linkage Analysis

Kevin J. Keen and Robert C. Elston
Case Western Reserve University, Cleveland, OH

For linkage studies, it is common practice to first collect probands and then extended family data on only the multiplex families thus identified. Although the different categories of relatives classified by disease, sex, and other covariates may have a particular multinomial distribution among families of a given size, the numbers as ascertained do not have the same distribution because of unequal probabilities of selection of families. The effects of different two-phase sampling designs on the estimation of parameters in the classical segregation model are examined. An approximation to the classical segregation likelihood model based on the approach of weighted distributions is found in Monte Carlo simulations to produce results close to those of the exact likelihood function for a balanced two-phase design. This has implications for more complex models in which the computation of the exact likelihood is prohibitive, such as for the enhancement of a typical sampling plan designed to collect multiplex families initially for linkage analysis but then extended retroactively to include a sample of

simplex families in order to perform a combined segregation and linkage analysis.

62

Genetic Linkages to Human SLE in 126 Multiplex Pedigrees

J. Kelly[1], K.L. Moser[1,2], C. Gray-McGuire[1], S. Rao[2], J. Salmon[3], R.P. Kimberly[4], J.C. Edberg[4], B.R. Neas[1,5], J.M. Olson[2], and J.B. Harley[1,5,6].
[1]Oklahoma Medical Research Foundation, OKC, OK; [2]Case Western Reserve University, Cleveland, OH; [3]Hospital for Special Surgery, NY, NY; [4]University of Alabama at Birmingham, Birmingham, AL; [5]University of Oklahoma, OKC, OK; [6]VA Medical Center, OKC, OK

We have evaluated 126 pedigrees multiplex for systemic lupus erythematosus containing 698 family members in an effort to better define the genetics of this systemic autoimmune disease. Multiple genome scans with >300 markers demonstrate ~20 suspected linkages. Of these, several are of sufficient magnitude or also found in other studies to conclude that a susceptibility gene has a high likelihood of being present. Using the criteria (lod>3.3), significant evidence for linkage is found at 1q22-23, lod=3.97 and 3.43 for FcγRIIA and D1s2762, respectively, by maximum-likelihood model-based methods. Affected relative pair evaluation (S.A.G.E. 4.0) has revealed linkage at 4p16-15 with D4s2366 (lod=3.62). Additional genome scan analyses of both univariate SLE related traits and derived Principal Components revealed multiple signals. The three largest multivariate effects obtained when using the Principal Component analysis were at 7p13 with D7s1818 (p=0.003), 4q36.1 with D4s1652 (p=0.020) and 15q15.1 with D15s659 (p=0.027). When univariate SLE related traits were analyzed, the largest effect, influenced by immunological characteristics, was found at 2q34 with D2s1384 (p=0.00048). For the univariate principal components, the most significant linkage result was at 4q36 with D4s1652 (p=0.00007). The genes responsible for these linkages are not known. These data establish that there are several linkages that each represents a gene or genes that have a role in immunopathogenesis and contribute toward developing dysregulation of the immune system causing lupus.

63

Single Nucleotide Polymorphisms (SNPs) in Alcohol Dehydrogenases and Risk of Oral Clefts in Humans: Different Effects of Maternal and Child Genotypes

M.H. Khoshnevisan[1], T. Wu[1], C. Sun[1], M. Mazeheri[2], R. Long[2], A. Miller-Chisholm[1], and S. Diehl[1]
[1]Craniofacial Epidemiology and Genetics Branch, NIDCR, NIH, Bethesda, MD; [2]Lancaster Cleft Palate Clinic, Lancaster, PA

Nonsyndromic cleft lip and palate (CL/P) has high heritability, but multiple susceptibility genes and environmental interactions underlie its etiology. Since this birth defect occurs during the first trimester of pregnancy, both the mother's and the child's genotypes may affect disease risk, further complicating analyses of candidate gene SNP associations. We conducted a family-based association study of SNPs in alcohol dehydrogenase (ADH) genes and maternal alcohol consumption during pregnancy. We used a recently developed statistical approach, which assesses effects of the maternal and affected child's genotypes on disease risk. Our sample included 163 families with ≥1 CL/P-affected member and 448 subjects with DNA available for genotyping. Two SNPs in ADH2 and one in ADH3 were genotyped by PCR-RFLP or OLA methods. We found associations in opposite directions for mothers' and affected children's genotypes for two SNPs. For example, risk was increased in children having the ADH3 "12" genotype, but decreased in mothers with this same genotype. We also found a significantly elevated frequency of the slow metabolizing"12" and "22" genotypes of the ADH3 SNP in CL/P-affected children whose mothers consumed alcohol during the first trimester compared to affected children whose mothers didn't drink. This suggests a gene by environment interaction: CL/P risk associated with the child's genotype at this SNP depends on maternal alcohol exposure. These examples illustrate some of the issues that must be incorporated into SNP linkage disequilibrium studies of complex diseases.

64

Stochastic Properties of p-values Associated With the Haseman-Elston Sib-Pair Linkage Test Under Alternative Hypotheses

A. Kingman[1], A.J.M. Sorant[2], and A.F. Wilson[2]
[1]NIDCR, NIH, Bethesda, MD; [2]NHGRI, NIH, Baltimore, MD

Recently we investigated potential power gains realized by using cluster-based statistics based on combinations of marker-specific Haseman-Elston sib-pair t-tests (IGES, 1998). In that study we simulated samples of 100 families with 4 sibs under different genetic models, and the H-E sib-pair t-test was used to test for linkage between a series of markers and a quantitative trait. The study focused on the increase in power realized for these tests, which involved deriving fixed critical values against which individual p-values were compared. The appropriate choice of critical

value for hypothesis testing has been debated in the past several years. As a result, much effort is being expended investigating alternative strategies. In the present investigation we use the same simulated data sets to investigate the stochastic nature of the distribution of the p-value from the H-E test for different underlying genetic models. The exact distribution of the p-value associated with the H-E test statistic can be derived (using asymptotic theory) for any specific genetic model. Under the null hypothesis the distribution is U(0,1). However, under non-null genetic models this distribution is not uniformly distributed. We examine distributions of 2000 p-values associated with the H-E statistic under each genetic model, derive the expected p-values (EPV) and relate them to the significance level and power for this test procedure. We illustrate these relationships as a function of trait heritability and distance of the marker from the trait locus.

65

Power of QTL Variance Component Linkage Analysis Methods

Daniel L. Koller, Brian P. McEvoy, and Tatiana Foroud
Department of Medical and Molecular Genetics, Indiana University, Indianapolis

Several linkage analysis packages implementing variance components models for quantitative traits are now available. To compare the performance of several programs, we used GASP to simulate an 80 cM chromosome with markers of 80% heterozygosity 10 cM apart. We allowed two parameters to vary which are generally beyond the investigator's control: QTL effect size (0, 10, and 25%) and overall trait heritability (50 and 80%). The remaining parameters that we varied are generally limited by the cost of the study: number of sibships studied (400 and 1000), QTL position (0, 2, or 5cM from nearest marker), and the number of parents genotyped (0,1, or 2). We analyzed 1000 replicates for each combination of parameter values with Mapmaker/SIBS (MM/S; weighted and unweighted) and Sibpal2 (beta 3) from the S.A.G.E. package. A subset of the parameter combinations were also analyzed with SOLAR.

The mean LOD score failed to reach suggestive linkage (2.2) for 400 sibships with any program; it did reach this threshold for 1000 sibships with all programs, but only for QTL effect sizes of 25%. The mean LOD reached significant linkage (3.6) only for the larger sample (1000 sibships), and then only with MM/S unweighted, a QTL of 25% effect, and otherwise ideal conditions. LOD scores were generally greatest for MM/S unweighted, with MM/S weighted and SOLAR slightly higher than Sibpal2. With a screening LOD score of 1.0, the false positive rates were 8% for Sibpal2, 14% for MM/S weighted, and 19% for MM/S unweighted. Interestingly, Sibpal2's power to detect a QTL of a given effect size showed less dependence on the overall trait heritability than did Mapmaker/SIBS. In general, the degree to which parents were genotyped (0, 1, or 2) had a greater effect on the power of all programs to detect a QTL than did the proximity of the QTL to the nearest marker in the 10cM screen. This suggests that limited resources might better be used for fully genotyping parents, rather than increasing marker density beyond 10cM in a genome screen.

66

Testing Linkage Disequilibrium in Sibships Using Conditional Logistic Regression with Robust Variance Estimators

Peter Kraft, Kimberly D. Siegmund, and Bryan Langholz
Department of Preventive Medicine, University of Southern California, Los Angeles, CA

Conditional logistic regression (CLR) — the standard epidemiologic method for the analysis of matched case control studies — is a convenient tool for testing linkage disequilibrium between an observed marker and a disease locus using data on sibships. CLR can be applied to sibships of arbitrary size for testing linkage in the presence of allelic disequilibrium but is generally not valid for testing disequilibrium in the presence of linkage. Classical CLR assumes that disease is independently distributed in sibships, conditional on the marker data. This assumption is violated when the marker is linked to the disease locus. Classical CLR score and Wald tests for disequilibrium in the presence of linkage are liberal.

We present robust variance estimators that lead to valid CLR score and Wald tests in the presence of linkage. The variance estimator for the robust Wald test is the usual sandwich estimator.

We compare the power of these tests, the classical CLR tests, and the Sibship Disequilibrium Test (SDT) proposed by Horvath and Laird (1998) via simulation. The robust score and Wald tests had more power to detect linkage equilibrium than the SDT and classical CLR tests. For example, when the marker studied is the disease gene itself, we found that the sample sizes needed to attain 80% power for the robust score and Wald tests could be 15-25% smaller than those needed for the SDT. Furthermore, CLR can account for measured covariates, unlike the SDT.

67

Combinatorial Partitioning Reveals Interactive Effects of the *ACE I/D* and *PAI-1 4G/5G* Polymorphisms on Plasma PAI-1 Levels

J.M. Lamb, N.J. Brown, M.T. Baker, J.V. Gainer, K.R. Richardson, D. Connolly, A. Gupta, D.E. Vaughan, J.H. Moore
Program in Human Genetics, Vanderbilt University Medical School, Nashville, TN 37232-0700

There is accumulating evidence that plasma levels of plasminogen activator inhibitor 1 (PAI-1), a risk factor for thrombosis, may be influenced by biological interactions between the renin-angiotensin and fibrinolytic systems. The goal of this study was to determine whether there is statistical evidence for interactive effects of genes from these two biochemical systems on plasma PAI-1 levels in a sample of 49 African Americans and 106 Caucasians.

Plasma PAI-1 antigen levels were measured using an enzyme-linked immunosorbant assay. We measured the *insertion/deletion* (*I/D*) polymorphism in the *angiotensin converting enzyme* (*ACE*) gene and the *4G/5G* polymorphism in the *PAI-1* gene. Because analysis of variance (ANOVA) has limited power to detect gene-gene interactions in small sample sizes, we utilized the combinatorial partitioning method (CPM) of Nelson et al. (Genome Research, in press) to test for interactive effects of the *ACE I/D* and *PAI-1 4G/5G* polymorphisms on plasma PAI-1 levels. CPM is a data reduction technique that pools multilocus genotypes into equivalent classes thus gaining back degrees of freedom lost when using standard ANOVA.

Using CPM, we found statistically significant evidence for an interaction between the *ACE* and *PAI-1* genes in both African American females (R^2=0.223, P<0.05) and males (R^2=0.217, P<0.05) and Caucasian females (R^2=0.102, P<0.05) and males (R^2=0.107, P<0.05). These interactions were not detected when standard ANOVA was used. This study illustrates the importance of considering interactions between the renin-angiotensin and fibrinolytic systems for understanding the genetic architecture of plasma PAI-1 levels.

68

A Multipoint NPL Regression Analysis of a Genome Scan Data Set in Systemic Lupus Erythematosus

C.D. Langefeld[1], P.M. Gaffney[2], S.S. Rich[1], and T.W. Behrens[2]
[1]Wake Forest University School of Medicine,
[2]University of Minnesota Medical School

Systemic lupus erythematosus (SLE) is an autoimmune inflammatory disease that can affect various organ systems and is characterized by the production of pathogenic autoantibodies. Twin and other familial studies suggest a significant genetic predisposition with estimates of the relative risk for siblings of SLE patients in the range of 10-20. We analyzed a genome scan data set for 187 affected SLE sibling pair families using a multipoint NPL regression approach; this approach allows for the simultaneous evaluation of multiple loci, their interactions and their interactions with environmental or other phenotypic characteristics. Using the multipoint NPL(pairs) statistic in GeneHunter, we report the position-specific unconditional maximum LOD score (MLS) from the NPL regression model containing only one locus and the corresponding conditional MLS from the NPL regression model conditional on the following loci are in the model.

Position (nearest marker)	Uncond MLS	Cond MLS
Chr 2 at 74 cM (D2S337)	2.00	2.47
Chr 6 at 75 cM (D6S257)	4.15	3.47
Chr 7 at 165 cM (D7S798)	1.69	3.10
Chr 16 at 61 cM (D16S415)	3.77	3.17
Cha 18 at 23 cM (D18S452)	1.47	2.01

Thus, adjusting for the evidence for linkage at other loci markedly increased the magnitude of evidence for linkage on chr 2, 7 and 18, yielding three positions with MLS>3.0. In addition, we observed evidence for two significant interactions: 1) chr 7 at 165 cM and chr 18 at 23 cM (β=-0.63; p=0.0198), and 2) chr 16 at 61 cM and chr 18 at 23 cM (β=-0.62; p=0.0488). We are currently testing whether the evidence for linkage varies by ethnicity, renal disease or other phenotypic characteristics.

69

Detecting Linkage in the Presence of Heterogeneity for Affected Sib-Pair Data

Suzanne M. Leal and Bertram Müller-Myshok
Laboratory of Statistical Genetics, The Rockefeller University, New York, Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Hamburg

The affected sib-pair design is commonly used in the study of complex traits. "Model-free" methods have been used in the study of complex traits (i.e., diabetes, schizophrenia) since the underlying mode of inheritance is usually unknown. A feature of both mendelian and complex traits is linkage heterogeneity. The presence of heterogeneity can greatly reduce the power to detect linkage and bias the estimate of the recombination fraction. A likelihood method was developed to test for linkage in the presence of locus heterogeneity for affected sib-pair data. The amount of increase in power to detect linkage depends upon the underlying model. For example when affected sib pairs were generated under a threshold model where an individual is affected with penetrance of 0.35 if they inherited at least 3 disease alleles (allele 2), where the three possible genotypes are 22 & 12, 12 & 22 or 22 & 22. The disease allele frequency at the first locus

is 0.3 and at the second locus 0.1. For a data set with 500 sib-pairs where half of the affected sib-pairs are segregating these two loci (α=0.5) the power to detect linkage at the first locus was increased by 7%. When a dataset of 1000 affected sib-pairs was generated under an autosomal recessive model of inheritance with 10% of the families segregating the same disease locus (α=0.1) the power to detect linkage was increased by 32%. For both examples markers (H=0.8) that are completely *linked* to the disease loci (θ=0.0) were generated and a total of 1,000 replicates were generated and analyzed. This likelihood method increases the power to detect linkage in the presence of heterogeneity in a variety of situation.

70

Statistical Methods for Human Nondisjunction Data

Jinming Li[1], Stephanie L. Sherman[2], Wendy P. Robison[3], and Hongyu Zhao[1]
[1]Yale University School of Medicine, New Haven, CT; [2]Emory University School of Medicine, Atlanta, GA; [3]University of British Columbia, Vancouver

Nondisjunction refers to the presence of two copies of a chromosome in a gamete. Chromosomal nondisjunction may lead to uniparental disomy (UPD) or trisomy. For UPD individuals, the chromosome number is normal but both homologs of a chromosome pair have originated from a single parent. UPD15 is associated with Prader-Willi syndrome and Angelman syndrome. Trisomy 21 is the leading cause of mental retardation, and is the result of one parent transmitting a disomic gamete, and the other parent transmitting the usual monosomic gamete. Genetic studies of UPD and trisomy employing many markers have helped geneticists to gain a better understanding of the molecular mechanisms underlying nondisjunction. However, the valuable information from such studies has not been fully utilized by existing methods. We have developed a general approach to analyzing nondisjunction data that can simultaneously handle multiple markers (including missing and uninformative markers) and also can incorporate crossover interference. Under the assumption that there is at most one crossover within each marker interval, we will discuss how to use our approach to recover exchange patterns during meiosis. In addition, we will present a hidden Markov model (HMM) for nondisjunction data. The major advantage of this HMM model is that the amount of computation increases linearly with the number of markers analyzed. We have implemented our methods in computer programs and carried out extensive simulations to study the performance of our methods. We will summarize our simulation results and describe the applications of our methods to UPD 15 data and trisomy 21 data.

71

The Heterogeneity Lod Cannot be Used to Estimate the Population Proportion of Linked Families

Mark Logue and Veronica J. Vieland
University of Iowa, Iowa City, IA

The admixture model of Smith which underlies the ordinary heterogeneity lod (HLOD) is only strictly correct when both linked and unlinked forms of disease have the same gene frequencies and penetrances, a highly unrealistic assumption in most contexts. When the linked and unlinked forms of disease have different models, the HLOD is no longer proportional to the probability of the genotypic data given the phenotypic data, and therefore it is subject to ascertainment bias. We show using a simple sib-pair example that the max HLOD can yield severely asymptotically biased estimates of all constituent parameters (the admixture parameter α, the recombination fraction θ, and the penetrance(s) at the linked locus). Because the estimate of alpha is a function of ascertainment and the trait models at both the linked and unlinked loci it yields almost no information about the true population proportion of linked families. One important special case is a study design in which only a single phenotypic configuration is sampled, e.g., an ASP design. In this case, we show that the max HLOD yields an estimate of α that converges, not to the population proportion, but to the expected sample proportion of linked families, which is a function of ascertainment and both trait models; while additional parameters converge properly (if they can be estimated from the data). When multiple phenotypic configurations are sampled, however, the estimate of α does not even converge to the expected sample proportion, and it therefore yields absolutely no meaningful information about the trait. Moreover fixing the model, even at the correct value for the linked locus, will produce biased estimates of θ and may reduce power for detecting linkage. On the other hand, the HLOD appears to be a robust measure of the strength of the evidence for linkage even when the trait models differ, when the trait model is treated as a nuisance parameter and estimated.

72

A First Trial of Retrospective Collaboration for Positional Cloning in Complex Inheritance: Assay of the Cytokine Region on Chromosome 5 by the Consortium on Asthma Genetics (COAG)

C. Lonjou, N.E. Morton for COAG
University of Southampton, U.K.

The central problem of complex inheritance is to map an oligogene for disease susceptibility prior to sequencing, integrating linkage and association over samples that differ

in several ways. Combination of evidence over multiple samples supports loci contributing to asthma susceptibility in the cytokine region on 5q (maximum lod = 2.61 near IL4), but no evidence for atopy. In principle the problems with retrospective collaboration on linkage appear to have been solved, providing far more information than a single study. A multipoint lod table evaluated at commonly agreed reference loci is required for both collaboration and meta-analysis, but variations in ascertainment, pedigree structure, phenotype definition and marker selection are tolerated. These methods are invariant with statistical advances that increase the power of lods and are applicable to all diseases, motivating collaboration rather than competition. In contrast to linkage, positional cloning by allelic association has yet to be extended to multiple samples, a prerequisite for efficient combination with linkage and the greatest current challenge to genetic epidemiology.

73

Power to Detect Linkage Using Sex-Averaged vs. Sex-Specific Lod Scores

Elizabeth A. Ludington, Veronica J. Vieland, Jian Huang
The University of Iowa

We used simulation studies involving general pedigrees to investigate the increase in power to detect linkage when allowing for differences between the male and female genetic maps. We considered a broad range of generating models that included both rare and common diseases, reduced penetrance, phenocopies, and less than full information at the marker locus. We found that maximizing lod scores over sex-specific recombination fractions did not result in substantial increases in power compared to lod scores maximized over a single sex-averaged recombination fraction. For example, for a dominant model with 50% penetrance and true male and female recombination rates of 5% and 20% respectively, the sex-specific lod score had a maximum increase in power of only three percentage points at a test size of 0.0001. For several generating models, the sex-averaged lod score and the sex-specific lod score had the same power, within sampling variability. For example, the sex-averaged lod score was only one percentage point more powerful than the sex-specific lod score for a common recessive disease with 50% penetrance. Overall power curves show that across a variety of simple and complex genetic models, the sex-specific and the sex-averaged lod scores are quite similar at all test sizes. We conclude that, even in regions where male and female map lengths differ appreciably, using sex-specific lod scores rather than sex-averaged lods is unlikely to substantially increase power to detect linkage.

74

Heritability of Resting Metabolic Rate in a Lean Nigerian Population

A. Luke, X. Zhu, A. Adeyemo, C. Rotimi, and R. Cooper
Loyola University, Maywood, IL; University of Ibadan, Ibadan, Nigeria; and Howard University, Washington, DC

The prevalence of obesity is greater among the populations of the African diaspora living in the western hemisphere than in West Africans. While on the individual level obesity is the result of an imbalance between energy intake and expenditure, the causes of population differences in prevalence are not clear. Resting metabolic rate (RMR) is the single largest component of total daily energy expenditure, comprising up to 70%, and is not thought to be subject to environmental influences. While the literature is inconsistent with regard to the relationship between RMR and obesity, we previously demonstrated that Nigerian women with a low relative RMR, adjusted for body composition, gained more weight over a three-year period than those with higher RMR. In an attempt to define both the genetic and environmental determinants of obesity and its intermediate phenotypes in black populations, we measured body composition, using bio-electrical impedance analysis, and RMR in 387 Nigerians from 61 pedigrees. Heritability was estimated for body composition and anthropometric traits using SOLAR, with age and sex as covariates, and for RMR with age, sex and fat-free mass as covariates. In this lean population, mean (SD) body mass index was 21.6 (4.2), heritability estimates for body mass index, fat-free mass, fat mass, % body fat, height and weight ranged from 40% to 60% - comparable to the literature. The heritability estimate for RMR was 30.3% (SE 0.10). In conjunction with results from Native American and white populations, these data indicate that RMR may be an important heritable trait related to obesity

75

Transmission/Disequilibrium Strategy in Complex Traits

Fabio Macciardi[1] and Alberto Morabito[2]
[1]CAMHealth, University of Toronto, Toronto, Canada; [2]Department of Medicine, Surgery and Dentistry (San Paolo), Milan, Italy

Detecting susceptibility genes for complex trait is difficult given the high degree of phenotypic and genetic heterogeneity of these traits. The linkage approach may not yield robust results, because this strategy has limited power to detect genes that confer a moderate risk or susceptibility. An alternative is represented by the gene-

association strategy, where comparisons for genetic polymorphisms are made between patients and controls matched by age, race and sex. Positive associations are thought to result from either true association of the specific polymorphisms with disease, or from the polymorphism studied being in linkage disequilibrium (LD) with a nearby polymorphism or mutation which confers risk. However, many confounding factors hamper the interpretation of these studies and today within-family association designs are preferred. The Transmission/ Disequilibrium Test is a test for Linkage Disequilibrium (LD) based on the detection of unequal parents-child transmission of high- versus low-risk alleles, and is a useful alternative for mapping genes with a modest effect on risk. LD mapping using haplotypes is even more powerful than analyzing LD with a single locus at a time, and recently methods have been proposed for a haplotype TDT strategy (H-TDT, Clayton and Jones, 1999). If LD mapping with haplotypes greatly enhances our ability to detect genes in complex disorders, it is nonetheless essential to preliminary evaluate the specific LD across any combination of the different polymorphisms at the candidate gene/region. A proposed procedure for LD mapping would require: (i) evaluation of the LD between the various polymorphisms (SNPs) across the region, using genotypes from unrelated parents; (ii) testing for the presence of a locus-disease linkage disequilibrium using the TDT for each individual marker, and finally (iii) testing the multiple marker haplotype transmission disequilibrium (H-TDT).

76

Genetics of Attention Deficit Hyperactivity Disorder: A Meta-Analysis

B.S. Maher, M.L. Marazita, R.E. Ferrell, R.E. Tarter, and M.M. Vanyukov
University of Pittsburgh, Pittsburgh, PA

The dopamine system may play a major role in the development of Attention Deficit Hyperactivity Disorder (ADHD). We applied a random effects model meta-analysis to TDT-based studies (several published and one unpublished, with probands from the Center for Education and Drug Abuse Research [CEDAR]) of association between ADHD and dopamine system genes DRD4, DRD5 and DAT1. A statistical test of heterogeneity was conducted for each group of studies. The meta-analysis of DRD4 included data from 5 studies with a total of 240 informative meioses. The pooled odds ratio estimate was 2.372 (1.419-3.967; $p<0.001$), demonstrating positive association. There was no evidence of heterogeneity between the DRD4 studies ($p=0.32$). For DRD5, two studies (167 informative meioses) were combined yielding a significant odds ratio of 2.725 (1.750-4.242; $p<0.001$). No evidence for heterogeneity between the studies was found ($p=0.83$). Five studies examining DAT1 with a total of 627 informative meioses yielded a non-significant pooled odds ratio estimate of 1.889 (0.934-3.824; $p=0.077$). Again, there was no support of heterogeneity between the studies. Overall, the meta-analyses support the involvement of the dopamine system genes in the ADHD liability variation and suggest the need for studies examining interactions between these genes.

77

A Major Locus Underlies the Inheritance of the Ratio FEV_1/FVC, Which is a Determinant of Airways Obstruction

A. Malhotra, A.P. Peiffer, D.T. Ryujin, R.E. Kanner, M.F. Leppert, and S.J. Hasstedt
University of Utah, Salt Lake City, UT

Diseases affecting pulmonary function are prevalent in the population. Diseases such as chronic obstructive pulmonary disease (COPD) and small airway disease have led to a high morbidity and mortality rate of individuals. There are a number of factors that are associated with lung disease. Cigarette smoking is one of the most important environmental factors affecting pulmonary function. However, a significant number of patients suffering from lung disease are non-smokers. Therefore, this suggests a role of genetic factors in the etiology of this disease.

Pulmonary function is studied by measuring the level of airway obstruction. Spirometry is used to measure the pulmonary function determinants FEV_1 (Forced Expiratory Volume in 1 second) and FVC (Forced Vital Capacity). The FEV_1/FVC ratio is reduced in individuals suffering from airways obstructions. Genetic analysis of this phenotype will increase our understanding of the role of genetic factors in the development of pulmonary function defects.

FEV_1 and FVC were measured on 159 members of 16 Utah pedigrees, originally collected for the Centre d'Etude du Polymorphisme Humain (CEPH) project. A segregation analysis of the FEV_1/FVC ratio was performed and has led to the conclusion of a major locus underlying the inheritance of this phenotype ($\chi^2=3.95$ $p=0.27$). The environmental model was rejected ($\chi^2=11.47$ $p=0.0094$). A linkage analysis will be performed on these data to localize the gene underlying FEV_1/FVC inheritance.

78

Estimation of Linkage and Association from Allele Transmission (TDT) Data

James D. Malley[1], Richard A. Redner[2], Thomas A. Severini[3], Judith A. Badner[4], and Joan E. Bailey-Wilson[5]

[1]CIT, NIH, Bethesda, MD, [2]University of Tulsa, OK, [3]Northwestern University, IL, University of Chicago, IL, and [5]NHGRI, NIH

The transmission/disequilibrium test (TDT) is widely used in genetics, but it does not estimate model parameters when departure from the null model is detected. On the other hand, it is a broadly accepted tenet of statistical practice that, whenever possible, one should calculate estimates of model parameters and derive associated standard errors and confidence intervals. Thus, we derive, for the first time, estimates of linkage and association from allele transmission (TDT) data. We present a parametric bootstrap procedure for generating confidence intervals for association, linkage recombination fraction, θ, and marker allele frequency, p. The confidence intervals provide separate tests of association and linkage, and have been validated in simulations. When a given marker allele is known to have a positive association with the disease gene, then we can use a likelihood ratio test for linkage. In general our procedures have at least as much statistical power as the usual TDT. In the presence of weak linkage ($0.3 \leq \theta \leq 0.5$) our methods can be as much as ten times more powerful than the TDT for detecting strong association.

79

Effects of Misspecification of Allele Frequencies on the Power of Haseman-Elston Sib-pair Linkage Method in Quantitative Trait Models

D.M. Mandal[1], A.F. Wilson[2], and J.E. Bailey-Wilson[2]
[1]LSU Health Sciences Center, New Orleans, USA;
[2]NHGRI/NIH, Baltimore, USA

It is well known that the Haseman-Elston (H-E) sib-pair linkage method does not assume that the genetic model underlying the trait phenotype is known without error although this assumption is made for marker loci. However, misspecification of allele frequencies at the marker locus decreases power when some or all parental genotypes are unknown. In this study, the power of the H-E sib-pair methods were compared for different types of traits when some or all parental data were missing and allele frequencies at the marker loci were misspecified.

Data were generated for a quantitative trait and marker loci in nuclear families using G.A.S.P. (V3.3). Three types of traits were simulated with two equifrequent alleles with a random environmental effect (50%, 70%, 90%). The simulated data were analyzed using (i) one of the parent?s marker data, and (ii) no parental marker data, with both correct and incorrect marker allele frequencies. This test is found to be robust in most of the situations considered except for a slight decrease in power when sample size is small and when the marker locus is not very polymorphic.

80

Prostate Cancer Study in African-American Population: Problems and Successful Strategies with Minority Recruitment

D.M. Mandal[1], T.N. Turley[2], L.B. Borrouso[1], J. Trent[3], J.E. Bailey-Wilson[3], and W. Rayford[4]
[1]Department of Medicine, [2]Department of Biometry and Genetics, [4]Department of Urology, LSU Health Sciences Center, New Orleans, LA; [3]NHGRI/NIH, Baltimore, MD

Prostate cancer is the most common malignancy in men in the United States, with the age-adjusted incidence of prostate cancer in African-American men being approximately 50% greater than in Caucasian-American men. Genetic factors may be important in explaining this variable incidence rate. There are data mostly on Caucasian families to support the existence of four potential prostate cancer susceptibility genes: three on chromosome 1 and the fourth one on the X-chromosome. Studies have shown that 1q24-25 (HPC1, which is estimated to account for probably less than one third of hereditary prostate cancer families) may contribute to the clustering of prostate cancer in some African-American families. However, there are not enough data on African-Americans to confirm the linkage. We are identifying African-American families with history of ≥ 3 prostate cancer cases through the prostate cancer screening program which is already in effect at the LSUHSC Department of Urology. So far, 51 of those families who mentioned a family history of prostate cancer at the time of screening have been contacted. Of these, 42 families agreed to participate in the study. The remaining 9 families could not be contacted thus far. Pedigree history questionnaires have been sent to six families with at least three cases of prostate cancer in blood relatives. Description of the data resource will be presented and the critical issues involving study of a minority population will be discussed. We will also discuss strategies to be used in enrolling African-Americans from a wide socioeconomic background into a genetic study.

81

Case/control Family Study Design to Investigate the Genetics of Autonomic Nervous System Dysfunction

M.L. Marazita[1], B.S. Maher[1], M.E. Cooper[1], J.M. Silvestri[2], A.D. Huffman[2], S.M. Smok-Pearsall[2], M.H. Kowal[2], and D.E. Weese-Mayer[2]
[1]University of Pittsburgh, Pittsburgh, PA, USA; [2]Rush Children's Hospital at Rush-Presbyterian-St. Luke's Medical Center, Rush University, Chicago, IL, USA

Many human disease traits appear to be familial but do not segregate in a clearly Mendelian fashion, in part

because the disease phenotype is not the proximate expression of the genes involved. These diseases may be lethal or have high morbidity, making it difficult to assemble families with affected family members. Study of associated phenotypic features may allow identification of family members who are not ?affected?, but who are carriers of relevant genes, thus improving the feasibility of genetic mapping studies for the disease. Idiopathic congenital central hypoventilation syndrome (CCHS) is a rare syndrome of disordered respiratory control with high morbidity/mortality. Hypothesizing that CCHS is the most severe manifestation of general autonomic nervous system dysfunction (ANSD), we applied a case/control family study design to investigate the genetics of ANSD phenotypic features in the families of CCHS probands. 52 confirmed CCHS probands were identified in the U.S., and matched on age, race and gender to controls. A scripted questionnaire characterized ANSD features in the cases, controls, and their family members. The frequency of each ANSD feature was compared between cases and controls, and between case-relatives versus control-relatives. Most ANSD symptoms were found to be more likely in cases and their relatives than controls ($p<0.05$). We performed major locus segregation analysis utilizing regressive models. Case families were consistent with transmission of a major effect, control families were not ($p<0.0001$).

82

Genotype by Smoking Interaction for Leptin Levels in the San Antonio Family Heart Study

L.J. Martin[1], S.A. Cole[1], J.E. Hixson[2], J. Blangero[1], and A.G. Comuzzie[1]
[1]Southwest Foundation for Biomedical Research, San Antonio, TX; [2]Human Genetics Center, University of Texas-Houston Health Science Center, Houston, TX

Recent studies have reported a marked inverse effect of smoking on serum levels of leptin (an adipocyte derived protein), offering a possible explanation for variation in body weight between smokers and non-smokers. The goal of this study was to examine the genetic architecture of the response to smoking in leptin levels using data from the San Antonio Family Heart Study. We employed a variance decomposition analysis using maximum likelihood methods to model genotype by smoking interactions for leptin levels and to examine the impact of the exclusion of smokers in a subsequent linkage analysis. We found significant evidence (p-value = 0.003) for different genes influence variation in serum leptin levels in smokers and non-smokers. In the subsequent linkage analysis with smokers excluded, we obtained a maximum LOD score of 3.1 near D8S1102. Interestingly, in a previous linkage analysis of leptin levels that included both smokers and non-smokers, we detected suggestive evidence of linkage (LOD = 2.1) near D8S1110. In both analyses the peak LOD spans a 95% confidence interval containing the β3 adrenergic receptor gene (*β3AR*), a strong candidate gene for obesity. Given these results it can be hypothesized that a quantitative trait locus in this vicinity of chromosome 8 may have a differential effect on the expression of leptin in smokers versus non-smokers.

This research was supported by NIH grant HL45522.

83

Evaluating Epistasis By Testing Change In Regression Slope Using Stratified Relative-Pair Analysis

R.A. Mathias[1], E.W. Pugh[1], Y.Y. Shugart[1], A.J.M. Sorant[2], K.C. Barnes[1], T.H. Beaty[1], and A.F. Wilson[2]
[1]Johns Hopkins University, Baltimore, MD, USA; [2]NHGRI, NIH, MD, Baltimore, USA

The detection of linkage to single loci using relative-pair regression analyses for quantitative traits is usually straightforward. However, the presence of gene-gene interaction, i.e., where the effects of one locus are modified by effects at a second locus, can substantially alter the power to detect linkage. Once evidence for linkage has been found to a given locus, stratification of relative-pairs for evidence of excess sharing of marker alleles identical by descent (IBD) at this locus may reveal other loci whose effects are dependent on the sharing of marker alleles IBD at the first locus. Testing for a difference in regression slope for the second locus using the full and stratified sample may provide a measure of interaction between the loci. This method was used to analyze total serum IgE in a set of Afro-Caribbean families. An interaction was found between two unlinked chromosomal regions. This finding was subsequently replicated in an independent Caucasian population. Currently, we are performing simulation studies to evaluate the statistical power and type I error rates of this procedure to detect interaction between two loci that jointly control a quantitative trait.

84

Parent-of-Origin Effects of the Specific Response to Allergens in the French EGEA Study

F. Meunier[1], M.H. Dizier[2], M. Guilloud-Bataille[2], M.P. Oryszczyn[3], J. Hochez[2], J. Maccario[3], J. Feingold[4], F. Kauffmann[3], and F. Demenais[1]
[1]INSERM EPI 00-06, [2]INSERM U535, [3]INSERM U472, [4]INSERM U155, Paris, France

Parent-of-origin effects are suspected to play an important role in atopy. These effects were investigated

by segregation and linkage analyses of the specific response to allergens (SRA) in 212 families ascertained through one asthmatic offspring (segregation) and 107 families with at least two asthmatic sibs (linkage), as part of the EGEA study. The following SRA phenotypes were considered: skin-prick test to at least one allergen (SPT), Multi-Rast Phadiatop® test, specific SPTs to Timothy Grass Pollen (TGP) and *Dermatophagoides Pteronyssinus* (*Der p*). Segregation analyses did not show evidence for major gene effects but familial dependences of all phenotypes, except *Der p*, were highly significant. The patterns of familial transmission include father-child ($p=0.005$) and mother-child ($p<10^{-4}$) dependences for Multi-Rast, mother-child dependence for SPT ($p<10^{-5}$) and father-child dependence ($p=0.004$) for TGP. Among the six regions detected by our genome screen (2q for *Der p*, 4q for Multi-Rast, 10p and 17q for SPT, 11p and 12q for TGP), we found significant differences between paternal and maternal contributions to linkage for Multi-Rast ($p=0.05$) with paternal linkage to 4q and SPT ($p=0.04$) with maternal linkage to 10p. Considering separately the parental meioses increased highly the significance for linkage to these two regions. Taking into account parent-of-origin effects appears of importance for gene identification.

Supported by Convention INSERM-MSD and FRM.

85

A Cellular Automata-Based Pattern Recognition Approach to Identifying Gene-Gene and Gene-Environment Interactions

Jason H. Moore and Lance W. Hahn
Program in Human Genetics, Vanderbilt University Medical School, Nashville, TN 37232-0700

It is increasingly clear that gene-gene and gene-environment interactions play an important role in determining risk of common complex diseases with multifactorial etiologies. We have developed an approach to identifying gene-gene and gene-environment interactions in discordant sib-pair study designs that combines the power of cellular automata (CA) for pattern recognition and genetic algorithms for machine learning.

CAs are dynamic systems that consist of an array of cells, each with a finite number of states that are updated at discrete time steps according to specific rules. The state of a cell at the next time step is determined by the current states of the neighboring cells. CAs can be used to perform computations by taking advantage of features such as massive parallelism. We have adapted a CA to accept an array of genotypes and/or environmental classes as input and produce an array of information as output that can be used to classify sibs as affected or unaffected (e.g. f(010110|affected) = 111111 or f(100010|unaffected) = 000000). In addition, we have adapted the genetic algorithm machine learning methodology and cross-validation to identify combinations of genotypes and/or environmental classes and CA parameters necessary to correctly classify affected and unaffected sibs.

Using simulated data, we have demonstrated that this approach can identify high-order interactions in small sample sizes in the absence of marginal or main effects. The results of this study suggest that pattern recognition approaches will be useful for the identification of genes that influence susceptibility to common complex diseases only through their interaction with other genetic and/or environmental factors.

86

Familial Resemblance of Some Intermediate Phenotypes of Hypertension in Two U.S. Sub-Populations

Swapan K. Nath[1], Aravinda Chakravarti[2], Richard Cooper[3], Sue Lewis[2], Audrey Lynn[2], Hemant Tiwari[1], Alan Weder[4], and Nicholas J. Schork[1]
[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH; [2]Department of Genetics, Case Western Reserve University, Cleveland, OH; [3]Department of Preventive Medicine, Loyola University, Chicago, IL; [4]Department of Medicine, University of Michigan Medical School, Ann Arbor, MI

Variation of quantitative phenotypes results from the combined effects of genes, environmental factors, and their interactions. To assess the familial resemblance of a quantitative phenotype, we employed familial correlation based model. Heritability of the quantitative phenotype was then calculated from the familial correlation. In order to assess differences in the heritability of different phenotypic measures across different populations, we estimated and compared the heritability of several correlated anthropomorphic, metabolic, and hemodynamic intermediate quantitative phenotypes of hypertension between a rural U.S. Caucasian population and an urban African-American population. Each phenotype was adjusted for the effects of covariate (e.g., age and gender). In addition, some phenotypes (e.g., BMI) were transformed to approximate normality via Box-Cox power transformations. The pattern of familial correlation and estimated heritability of some phenotypes were significantly varies across sub-populations. We discuss the implications of our findings for gene discovery initiatives. The data for this study was generated by GenNet Network of the FBPP and funded by the NHLBI.

87

Clustering Methods Applied to the Allele-Sharing Data in an Alzheimer's Disease and a Simulated Dataset

Rosalind J. Neuman, Nancy L. Saccone,
Peter Holmans, John P. Rice, and Lingwei Sun
Washington University, St. Louis, MO

Traditional allele-sharing linkage analysis methods have played an important role in the search for disease susceptibility loci in complex diseases. One limitation of these methods is their inability to include non-genetic factors directly in the analysis. We applied 2 clustering techniques, latent class analysis (LCA) and cluster analysis (CLA), to allele-sharing in 2 sets of data: A real data set consisting of 507 sibling pairs from families with late-onset Alzheimer's disease (AD) and sibling data from one of the GAW11 simulated populations. The AD data contained 282 ASP and 225 discordant pairs (DSP); the GAW11 consisted of 184 ASPs and 801 DSP. Briefly, the basic strategy was to identify subsets of pairs with a large proportion of ASP, followed by examining those chromosomal regions where these pairs share an excess number of alleles IBD.

Markers from chromosomes 1, 6, 10, and 12 were used to analyze the AD data. Some of the markers were chosen to be located near the regions of interest previously identified; others were chosen at random. The analyses also included the APOE status of each pair: whether both, neither, or only 1 sib carried the e4 allele. LCA and CLA identified subsets of the data with an excess number of ASP sharing IBD in the same regions where signals had been detected in a previous analysis of these AD data using ASP methods.

LCA and CLA were applied to GAW11 markers chosen to be near the trait loci and at unlinked random locations. Both methods identified those markers linked to all 3 disease loci either with or without the inclusion of the environmental risk factor.

These results indicate the feasibility of using classification methods to clarify the relationship between the disease phenotype and individual locus information.

88

Does Accounting for Gene-Environment (G×E) Interaction Increase the Power to Detect the Main Effect of the Gene?

J.M. Norris[1,2], H. Selinger-Leneman[1], E. Genin[1],
M. Khlat[3]
[1]INSERM U535, Le Kremlin-Bicetre, France;
[2]University of Colorado Health Sciences Center,
Denver, USA; [3]Institut National d'Etudes
Demographiques, Paris, France

Despite tremendous efforts, few genes involved in the susceptibility for complex disorders have been identified. One explanation for this is that these disorders are a result of an interaction between genes and environment, and under such conditions, it may be difficult to measure the true genetic effect without accounting for the interaction. Likelihood-based analyses allow for the study of the joint effects of genotype and environment using case-parent trios. Prior studies have focused on estimating the power of this method *to detect the G×E interaction*. In this study, we explored under which conditions accounting for G×E interaction enhanced one's ability *to detect the main effect of the gene*. Using simulated data and asymptotic power calculations, we investigated the power to detect the gene effect over varying exposure frequencies and several models of G×E interaction. In general, for a given sample size, interaction model and allele frequency, the gain in power while accounting for the interaction is dependent on the exposure frequency, with the largest gain seen for the smallest exposure frequencies. For example, considering a gene with a disease allele frequency of 0.2 that has no effect in the absence of exposure and a 10-fold increase in risk among exposed homozygotes, the power to detect the gene effect in 200 trios is 78.8% when accounting for the interaction and 31.4% when not, for an exposure frequency is 10%. For an exposure frequency of 20%, the respective power calculations were 97.5 and 82.2%. When designing a study to determine the main effect of the gene in the presence of G×E interaction, one must consider the several factors, including the exposure frequency.

89

Genotype Interacts with Glucose Tolerance Status to Influence Obesity Traits in American Indians: The Strong Heart Family Study

K.E. North[1], J.W. MacCluer[1], B.V. Howard[2],
M.S. Marion[3], E.T. Lee[4], and R.R. Fabsitz[5]
[1]Southwest Foundation for Biomedical Research, San
Antonio, TX; [2]Medstar Research Institute,
Washington, DC; [3]Aberdeen Area Tribal Chrmn's
Health Board, Rapid City, SD; [4]Univ Oklahoma
Health Science Center, Oklahoma City, OK; [5]NHLBI,
Bethesda, MD

The relationship between obesity related traits and glucose tolerance is well established. Previous research in the Strong Heart Family Study (SHFS) has demonstrated significant heritabilities for obesity related traits and has implicated glucose tolerance status as an important correlate in American Indian tribes in Arizona (AZ), Oklahoma (OK), and the Dakotas (DA). The purpose of the present study was to determine whether still-unidentified genetic effects on obesity related phenotypes

in the SHFS differed in individuals with impaired glucose tolerance (IGT) or diabetes as compared to those with normal glucose tolerance (NGT).

Approximately 900 individuals, age 18 or older, in 32 extended families, were examined between 1997 and 1999 (12 families in AZ, 11 in OK, 9 in DA). Genotype x glucose tolerance status interaction was estimated for body mass index (BMI), fat mass, waist?to-hip ratio (WHR), and weight, using a maximum likelihood variance decomposition technique implemented in FISHER. We found that the genes influencing BMI (p=0.004), fat mass (p=0.001), and weight (p=0.04) are differentially expressed in individuals with IGT and diabetes, compared to NGT individuals. These findings will be pursued in analyses of data from a genome scan now in progress.

Supported by NHLBI grant U01 HL41642.

90

A Quantitative Estimate of Individual Genetic Vulnerability for a Complex Trait: Application to Bipolar Illness

J.I. Nurnberger, Jr, T. Foroud, E.T. Meyer, K.L. Hu, L. Flury, J. Su, H. Edenberg, and Collaborators from the NIMH Genetics Initiative Bipolar Group
Indiana University, Departments of Psychiatry, Medical and Molecular Genetics, and Biochemistry and Molecular Biology

We have developed an algorithm that may be applied to genomic survey data to derive estimates of genetic vulnerability for each person in a pedigree. The method utilizes NPL scores from GENEHUNTER PLUS to identify loci of interest. A 'sharing score' among affecteds is computed for each allele at a given locus for each pedigree. An individual's 'locus vulnerability score' is the sum of scores for each allele multiplied by the NPL score at the locus. The total 'vulnerability score' is then corrected for the genotypic information available for the individual at the loci of interest. This method was applied to the initial NIMH Genetics Initiative Bipolar dataset (N = 537 persons). We used the output from a GENEHUNTER PLUS analysis with a diagnostic model in which SA/BP, BPI, BPII, and UPR are affected. Thirty-three loci with Z scores $\geq$ 1.0 and p values $\leq$ 0.1 were included in the analysis; loci were separated from each other by at least 20 cM. A bimodal distribution of vulnerability scores was produced, with affected individuals generating higher scores than unaffected. The difference between affected and unaffected was outside the range of differences observed in 100 randomly generated datasets using the same pedigrees. A threshold score predicted affected status with 75.2% sensitivity and specificity. We suggest that this algorithm, while necessarily employing a crude approximation to the underlying genetics of bipolar illness, is nevertheless capable of producing clinically meaningful results. Refinement of the method is in progress, including additional simulations to assess the distributional properties of the test statistic, and the development of a version that includes information regarding alleles shared IBD.

91

Multipoint IBD Estimates for General Pedigrees

Jeffrey R. O'Connell
University of Pittsburgh, PA

Multipoint identity by descent (IBD) probabilities are used in a variety of genetic analyses, including nonparametric linkage (NPL) statistics and quantitative trait locus (QTL) analyses. For small to medium-size pedigrees exact IBD probabilities using all the multipoint marker data can be efficiently calculated at any map position using the Lander-Green algorithm as implemented into programs such as GENEHUNTER and ALLEGRO. For pedigrees beyond the computational limits of the Lander-Green algorithm, approximate multipoint IBD probabilities can be computed using Markov chain Monte Carlo (MCMC) methods as implemented into SimWalk2. An alternative method is introduced that uses the Elston-Stewart algorithm to compute multipoint IBD probabilities at a map position by using a dummy locus at the position together with nearby flanking markers, where at the dummy locus each founder is assigned a distinct pair of alleles and non founders set as unknown. To compute the IBD probabilities for a set of individuals P, the program PROFILER, computes the joint probability of each genotype vector for P at the dummy locus conditional on the flanking markers. In general, PROFILER uses the likelihood engine of VITESSE and efficient recursive genotype elimination algorithms to compute joint probabilities of genotype vectors for any subset of individuals in the pedigree and any subset of markers conditional on the full pedigree and marker data. The multipoint IBD probabilities are approximate since the number of flanking markers that are computationally feasible in the Elston-Stewart algorithm is limited. Accuracy and timing results are compared between PROFILE and both ALLEGRO and SimWalk2.

92

Power of Genomewide Association Studies of Complex Disease Genes: Statistical Limitations of Direct and Indirect Approaches

Jun Ohashi and Katsushi Tokunaga
University of Tokyo, Tokyo, Japan

Genomewide association studies using a dense map of single nucleotide polymorphism (SNP) markers seem to enable us to detect a number of complex disease genes. In the genomewide screening, whether susceptibility genes can be detected is dependent not only on the degree of linkage disequilibrium between the disease variant and the SNP marker but also on the difference in their allele frequencies. These factors as well as penetrance of the disease variant influence the power of indirect genomewide association studies, which will be actually performed for many complex diseases in the near future, whereas the statistical limitation of such approaches is not well understood. We calculated the required number of samples in case-control studies for any penetrances, allele frequencies, and linkage disequilibrium. The simple formula obtained in this paper will help us to design case-control studies for complex human diseases. Using the formula, we found that a remarkable reduction in statistical power of indirect studies, compared to direct ones, is unavoidable in the genomewide screening of the complex disease genes. Also, SNP alleles suitable for genetic markers were found to be very rare, based on the theory of population genetics. These results suggest that indirect genomewide association studies may miss disease variants with modest contributions even if a large number of SNP markers are available.

93

Powerful LD Mapping with SNPs by Data Mining Methods

P. Onkamo[1], V. Ollikainen[2], P. Sevon[3], H.T.T. Toivonen[4], K. Vasko[1], H. Mannila[4], and J. Kere[2]

[1]Rolf Nevanlinna Institute, University of Helsinki, Finland; [2]Finnish Genome Center, University of Helsinki, Finland; [3]Department of Computer Science, University of Helsinki, Finland; [4]Nokia Research Center, Helsinki, Finland

Linkage disequilibrium (LD) mapping with SNPs has been questioned because of the huge multiple testing problem, limited marker information content, and errors in genotyping process. Previously, we have introduced a new, data-mining based approach for LD mapping, Haplotype Pattern Mining (Toivonen et al., AJHG 66(7), 2000). In HPM the algorithm finds the disease-associated haplotype patterns, which are observed strings of alleles in neighbouring marker loci, possibly containing gaps. All patterns found are ordered by their strength of association to the phenotype, and the patterns exceeding a given threshold level are used for prediction of gene location. We apply this model-free method to data sets containing large numbers of densely spaced SNPs. To accommodate environmental effects we allow, as a new feature of the method, the haplotype patterns to contain an arbitrary number of polychotomized covariates. We evaluate the performance of HPM with simulated genomewide SNP data sets taken from a population founded 1,000 years ago by a group of 10,000 individuals that was allowed to grow exponentially to its current size of one million inhabitants. The markers were placed at intervals of 1 cM, and the frequency of the major allele was between 0.5 and 0.7. We analyzed the performance of HPM with various sample sizes and disease models with high numbers of phenocopies. The method is shown to be capable of analyzing SNP data with missing alleles, haplotyping ambiguities and erroneous data. The limited marker information content is overcome as HPM effectively combines the information from nearby marker loci. HPM also allows correcting for multiple testing by a standard permutation procedure. The statistical power is roughly the same as that of microsatellites with one third of the density of SNPs.

94

A Retrospective Collaboration on Chromosome 5 by the International Consortium on Asthma Genetics (COAG): Localization of a Novel Gene Regulating Lung Function to the 5q Cytokine Cluster

L.J. Palmer[1,2], P.R. Burton[3], W.O.C.M. Cookson[4], J.W. Holloway[5], K.B. Jacobs[1], and R.C. Elston[1]

[1]Case Western Reserve University, Cleveland, USA; [2]Harvard University, Boston, USA; [3]Leicester University, UK; [4]Oxford University, UK; [5]Southampton University, UK

The Consortium On Asthma Genetics (COAG) was established in 1999 as an international effort to define the candidate regions and genes for asthma and atopy with greater precision and reliability than can be achieved by smaller individual studies. A pilot retrospective study has focused on combination of evidence for the best investigated asthma candidate region, the cytokine cluster on chromosome 5q31-33. The percent predicted forced expiratory volume in one second (FEV_1), a physiological phenotype closely associated with asthma, was measured in 3 of the COAG datasets (n=309 randomly ascertained Caucasian families; 702 sib-pairs; 15 markers typed). Novel analytical approaches included the use of a new variance components implementation of the multipoint Haseman-Elston linkage statistic and the use of Gibbs sampling to 'pre-process' the phenotype prior to linkage. Initial analyses found a novel area of possible linkage in the cytokine cluster ($P<0.01$). Use of sigma-squared-A-random effects (SSARs) and major gene focusing strengthened evidence of linkage to this region

(P<0.00001) and narrowed the peak. In all cases, results based upon pooled study-specific regression coefficients were consistent with, although more conservative than, results based upon the analysis of pooled datasets. Future fine mapping studies will hopefully define the specific 5q31 gene regulating lung function. This study illustrates the benefits of cooperation among groups studying a complex human disease, that data from diverse studies and populations can be effectively pooled, and that new and important information can result from the analysis of large, pooled datasets.

95

Haseman and Elston Revisited: The Effects of Asertainment and Residual Familial Correlations on Power to Detect Linkage

L.J. Palmer[1,2], K.B. Jacobs[1], and R.C. Elston[1]
[1]Case Western Reserve University, Cleveland, OH, USA; [2]Harvard University, Boston, MA, USA

A new implementation of the Haseman-Elston (H-E) sib-pair linkage test for quantitative traits that uses the mean-corrected product of the sib's trait values has been recently described; this test has been shown to be more powerful than the original squared-trait difference H-E test in the case of a random population sample (Elston, R., Buxbaum, S., Jacobs, K. & Olson, J. Haseman and Elston Revisited. *Genetic Epidemiology* [in press]). Under a variety of models simulating single ascertainment, strength of ascertainment and concomitant deviation of the sample mean away from the true population mean was found to have deleterious effects on the power to detect linkage to a polymorphic marker. In contrast, the empirical power of the original H-E squared-trait difference linkage test is insensitive to the mean and was found to be better than the cross-product linkage test for many values of the mean. Unexpectedly, the power of the cross-product test was not maximal at either the population mean or the sample mean, but at some point between these two possible values. In further simulations, we found that the squared-trait difference test is consistently more powerful than the cross-product test when strong residual familial correlations are present, but consistently less powerful than the cross-product test in the absence of such correlations. Thus, further investigation has indicated that, although the new H-E cross-product linkage statistic is often more powerful than the original H-E squared-trait difference linkage statistic, there are some situations where this statistic should be applied with caution. These involve situations of strong ascertainment and situations where the variance of the continuous phenotype being investigated has substantial familial components not due to the marker being linked.

96

Power to Reconstruct the Genotype of a Missing Individual

Nathan Pankratz, Daniel L. Koller, and Tatiana Foroud
Department of Medical and Molecular Genetics, Indiana University, Indianapolis, USA

To examine the power of various numbers and types of relatives to infer the genotypic data of a missing individual, we used GASP (Wilson et al., 1996) to generate three-generational pedigrees consisting of a pair of siblings, their deceased parents, and one of the siblings' spouse and three offspring. Individuals in the upper 10% of the simulated quantitative trait distribution were considered affected, and all others were assigned an unknown phenotype. Markers with heterozygosity of 80% were simulated at 10 cM intervals. The gene was located in the middle of the marker map, 2 cM from the nearest marker. For each replicate, 500 families containing a pair of affected siblings in the second generation were retained for analysis. Genehunter-Plus (Kruglyak et al., 1996; Kong and Cox, 1997) was used to analyze the 500 families in each of 1,000 replicates.

Qualitative linkage analyses were performed with genotypic data assumed to be available for both affected siblings, the situation in which both members of the sibpair have DNA. This was our reference LOD score result, to which all subsequent analyses were compared. We then performed a series of steps in which we removed the affected sibling with the offspring and reanalyzed the sample, attempting to infer the missing individual's genotype using a sequentially smaller number of relatives. Six conditions were tested: 1) three offspring, with and without DNA on the deceased individual's spouse; 2) two offspring, with and without the deceased individual's spouse; and 3) one offspring, with and without DNA on the deceased individual's spouse. The lod score was recomputed under each condition in the 1,000 replicates consisting of 500 affected sibling pairs each. When the pedigree consisted of two or three offspring and the spouse or three offspring without the spouse, the analysis retained over 70% of the optimum lod score. When one or two offspring and no spousal DNA or one offspring and the spouse were included, nearly 50% of the optimum lod score was still obtained.

97

Locus Heterogeneity Based Upon Age of Diagnosis in Type 1 Diabetes

A.D. Paterson
Department of Genetics and Genomic Biology, The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada

The age at diagnosis (AOD) of type 1 diabetes (T1DM) varies considerably. Twin and animal studies suggest that AOD may be under genetic control. Similar to other complex diseases, we hypothesized that by using T1DM families who have similar AOD may identify novel loci and increase the power for fine-mapping. To test this, we used data from a genome scan for linkage using 356 affected sibpair families with T1DM from the UK (Mein et al., 1998). The mean and median AOD is 10 years in these families. We thus subdivided the families into three groups: those where both sibs were diagnosed ≤ 10 years (Early); both > 10 years (Late); and those with one sib in each category ('Discordant'). Linkage analysis was performed separately in each group across the genome. No significant differences in linkage to HLA (6p21.3) were observed between the three groups. However, linkage derives predominantly from Early families at *IDDM10*, on 10p13-q11. Two linked loci, at 4p16.3 and 4p16.1 demonstrate linkage in Late and Early families respectively. These linkages are of interest because they include regions containing genes which cause diabetes. Huntington's disease patients (HD, 4p16.3) are at increased risk for late onset diabetes and HD transgenic mice develop diabetes. Wolfram syndrome patients (WFS1, 4p16.1) suffer from diabetes with AOD 6 years. Finally, a marker on 4q26-27 (D4S430) appears to demonstrate evidence for linkage (excess non-sharing) in families with 'discordant' AOD. This region potentially overlaps with a locus for Type 2 diabetes, and is < 5 cM from IL2, the human homolog of the murine gene thought to be Idd3. Relative homogeneity of AOD appears to identify novel loci for T1DM and allows for narrowing confidence intervals for the position of the underlying susceptibility loci.

98

Hierarchical Agglomerative Nesting of Gene Expression Levels from cDNA Microarrays

Leif E. Peterson
Department of Medicine, Baylor College of Medicine, Houston, TX USA

Expression of genes involved in cellular processes and response to environmental stimuli can be assessed on a large-scale with complimentary DNA (cDNA) microarrays. Clinical applications with cDNA microarrays compare gene expression levels in patient tissue with gene expression levels in cell lines, or changes in expression arising from treatment with pharmaceuticals. A hierarchical agglomerative nesting procedure for gene expression levels from cDNA microarrays is described. A two-step nesting procedure is implemented which first agglomerates arrays as objects and genes as attributes, followed by agglomeration of genes as objects and arrays as attributes. Dispersion matrices of the raw or standardized input data are adjusted for missing values and genes are selected based on acceptance criteria for correlation coefficients. Principal components analysis is performed after hierarchical agglomerative nesting to identify patterns of expression for unique sets of genes. Results are provided for simulated data sets involving 2,500, 5,000 and 10,000 genes and a varying number of arrays (10, 25, 50, and 100). Future analyses will look at the role of exploratory factor analysis and rotation before agglomeration procedures are applied.

Supported by NCI grant CA78199-02.

99

A Genome Screen for Atopy

S.S. Rich[1], M.N. Blumenthal[2], A. Florance[1], T. Beaty[3], E. Bleecker[1], C. Ober[4], S. Banks-Schlegel[5], and C.D. Langefeld[1] for the CSGA
[1]Wake Forest University; [2]University of Minnesota; [3]Johns Hopkins University; [4]University of Chicago, [5]NHLBI

Atopy is the adverse immune reaction involving IgE antibodies, is associated with rhinitis, asthma and dermatitis and is known to aggregate in families. In the Collaborative Study on the Genetics of Asthma (CSGA), families were collected having two or more siblings with asthma. All members were characterized clinically, including atopy (defined by the presence of asthma and/or rhinitis with one or more positive skin tests to common allergens). We report here on the results of a genome scan for atopy.

A total of 146 Caucasian, 128 African-American and 40 Hispanic American families were included in the genetic analyses. In total, multipoint lod scores greater than one were observed for chromosomes 11 (lod=1.96, D11S1986), 20 (lod=1.80, D20S473/D20S604), 19 (lod=1.29, D19S519) and 21 (lod=1.07, D21S1270). Using the subsetting approach of Kong & Cox, no evidence for interaction was observed between chromosomes 11 and 20; there was modest evidence for interaction between chromosomes 20 and 1. Most of the evidence for linkage on chromosome 19 disappears when conditioned on 20. These results suggest that the genetic contribution to atopy is complex with both primary genetic effects and interaction effects. Future analyses will be focused on the interaction effects using the NPL regression approach (Langefeld & Boehnke).

100

Detection of High-Order Interactions Among Estrogen Metabolism Genes in Sporadic Breast Cancer

M.D. Ritchie, N. Roodi, L.R. Bailey, W.D. Dupont, W.D. Plummer, L.W. Hahn, F.F. Parl, and J.H. Moore
Program in Human Genetics, Vanderbilt University Medical School, Nashville, TN 37232-0700

Mammary metabolism of estrogen to carcinogenic catechol estrogens is thought to increase the risk of sporadic breast cancer. An important question is whether breast cancer risk can be predicted by knowledge about variation in genes from the estrogen metabolism pathway. Previous studies by our group and others suggest that genes from this pathway do not have main or independent effects. The goal of this study was to determine whether interactions among variations in genes from this pathway are associated with sporadic breast cancer.

Our sample included 414 age-matched Caucasian women: 207 cases and 207 controls. We measured eight polymorphisms in five genes from the estrogen metabolism pathway: *CYP1A1, CYP1B1, COMT, GSTM1*, and *GSTT1*. We employed a multifactor data reduction (MDR) method to evaluate each of the possible interactions among the eight polymorphisms. For each particular combination of loci, we identified multilocus genotypes that are more common in cases than controls and pooled these into a single group. This reduces the dimensionality of the data to one. We assessed the average prediction error or misclassification rate of each combination of loci and their optimal multilocus genotype grouping using nine-fold cross validation.

The most parsimonious model had an average misclassification rate of 15.2% and included five polymorphisms from the *CYP1A1, CYP1B1, COMT* and *GSTM1* genes. Permutation testing indicated the probability of observing this misclassification rate given the null hypothesis of no association was less than 0.05. Adding more loci did not significantly influence the average misclassification rate. This study illustrates the importance of considering high-order genetic interactions in studies of complex diseases such as sporadic breast cancer when there are no main or independent genetic effects.

101

Confirmatory Evidence of Linkage to 7Q for Autism Based on Combined Analysis of Three Independent Data Sets

Susan Santangelo[1] for CLSA, Allison Ashley-Koch[2], Margaret Pericak-Vance[2], Joseph Buxbaum[3]
[1]Tufts University School of Medicine, Boston MA; [2]Duke University, Durham NC; [3]Seaver Autism Research Center (SARC), Mount Sinai School of Medicine, New York, NY

In order to follow up on an initial report of evidence of linkage for autism to 7q [MLS=2.5 @ 145 cM, IMGSAC, 1998], we jointly analyzed chromosome 7 data from our three independent genome screens: CLSA (N=75), SARC (N=89), and Duke (N=83); each data set comprises primarily ASPs. The Duke sample gives a max multipoint MLS [Risch, 1990, with the possible triangle constraint of Holmans, 1993] of 2.5 @ 129 cM, while both the CLSA and SARC samples provide much weaker evidence (1.4 @ 100 cM; 1.2 @ 155 cM respectively). When averaged across the studies, the MLSs are quite small across the entire region, which might suggest little confirmatory evidence from the combined data. However, recent work [Vieland, Huang, & Wang, this meeting] has shown that when levels of heterogeneity vary across data sets, averaging the scores without adequate allowance for inter-sample heterogeneity will tend to dramatically underestimate the true evidence for linkage when linkage exists. By contrast, in ASP data the *sum* across data sets of maximum MLSs is a robust approximation to a correct analysis allowing for inter-sample heterogeneity. Summing the max MLSs from the three studies at 1cM intervals along a 68 cM stretch of 7q yields a total max summed MLS of 2.6 @ 130 cM with two additional peaks of 2.5 @ 100 cM and 2.4 @ 150 cM and with positive signal across the entire span. The p-value corresponding to a max MLS of 2.6 based on three data sets is approximately 0.004. Thus while the averaged MLSs did not appear to confirm linkage, basing our inference instead on the sum of max MLSs does appear to confirm a growing body of evidence for an autism gene in this region of 7q [see e.g.,Ashley-Koch et al., 1999]. However, precise localization of the gene within this 50cM region remains to be resolved.

102

Linkage Heterogeneity: Extending the Mixture Likelihood to Include Pedigree Covariates

D.J. Schaid and S.M. McDonnell
Mayo Clinic/Foundation, Rochester, MN, USA

Linkage heterogeneity is often evaluated by computing the A-Test (e.g., HOMOG software), which is based on a likelihood ratio for a mixture of two types of pedigrees: a fraction (α) of pedigrees that are linked to a chromosomal region, and the remaining fraction ($1-\alpha$) that are not linked. To further discriminate linked from non-linked pedigrees, it is not unusual to perform linkage analyses within subsets defined by features of the pedigrees, such as young vs. old age of onset. However, subset analyses are prone to weaker power and inflated false-positive rates, and they do not facilitate more complex analyses, such as interaction of pedigree features. To provide more flexible linkage heterogeneity analyses, we present a new linkage heterogeneity regression model, for which the pedigree characteristics are used as covariates (X), and the probability that the pedigree is a

linked-type is modeled by logistic regression, $\log(\alpha/[1-\alpha) = \beta_{o+} \beta X$. Because the information on whether a pedigree is linked or not is missing, the Expectation-Maximization algorithm is used to maximize the likelihood. These new methods are demonstrated by application to linkage data for prostate cancer families (AJHG 66:945-957, 2000).

103

The Relative Efficiency of Using Related Cases in Case-Control Studies of Candidate Genes

D.J. Schaid and S.N. Thibodeau
Mayo Clinic/Foundation, Rochester, MN, USA

Case-control studies of candidate genes can be performed using available families sampled for genetic linkage analyses. While planning a candidate-gene study for hereditary prostate cancer, we were faced with either randomly sampling one case per family, or using multiple cases per family and accounting for the dependence of cases in power calculations. Recognizing that controls must be carefully selected (i.e., matched on ethnicity, or family-based controls), it is was not clear how much information genetically-related cases would offer. We show that the relative efficiency of using n affected sibs versus n unrelated cases is $1/\{1+(n-1)\rho\}$, where ρ is the correlation coefficient which depends on the candidate-gene allele frequency, the genotype relative risks, and whether allele counts or carrier status is compared between cases and controls. For power calculations, the relative efficiency can be translated to an "effective" number of independent cases per family, $n^{eff} = n/\{1+(n-1)\rho\}$, which can then be used in the usual power formulas to compare two proportions. For common alleles with small relative risks, ρ is close to 0.5, implying that n^{eff} would not be greater than 2, no matter how many affected siblings are analyzed.

104

Conditional Predictive Ordinates for Generalized Linear Mixed Models Fitted to Censored Survival Phenotypes Using Gibbs Sampling

K.J. Scurrah[1,2,3], M.L. Hazelton[2], and P.R. Burton[1, 3]
[1] TVW Telethon Institute for Child Health Research, Perth, Australia; [2] University of Western Australia, Perth, Australia; [3] University of Leicester, Leicester, United Kingdom

We have recently described methods for fitting generalized linear mixed models (GLMMs) to censored survival data collected from nuclear and extended families using the MCMC technique of Gibbs sampling as implemented in the BUGS software. However, once models have been fitted, model determination, checking and selection should be performed. We must decide whether a given model fits the observed data adequately, and which of two or more possible models is 'best'. Conditional predictive ordinates (CPOs) are a useful tool in this process. These can easily be calculated in BUGS using a Monte Carlo integration, and they enable models to be compared both locally, by checking which observations favour each model, and globally, by using the pseudo-Bayes factor (the product of the CPOs). The use of both graphical and numerical methods can assist in this process. A particular advantage of using CPOs for genetic data is that the model fit for different classes of individuals, such as founders or specific pedigrees, can be compared. Another advantage is that the models compared need not be nested; for example, the fit of models assuming a Weibull distribution for survival times can be compared with piecewise exponential models. We will describe how to calculate CPOs for our models, and present some graphical and numerical results for a variety of models which have been fitted to real and simulated data.

105

Evaluating the Genetic Contribution to Change in Quantitative Traits: the Framingham Heart Study

Yujun Shao, Mark Glickman, Anita L. DeStefano, Ralph B. D'Agostino, Sr., Richard H. Myers, and L. Adrienne Cupples
Boston University Schools of Public Health and Medicine and the Department of Mathematics and Statistics, USA

A wide variety of methods have been developed for evaluation of the genetic contribution to levels of a quantitative trait. Few methods have been developed for the evaluation of the genetic contribution in the expression of the change in a quantitative trait over time. We present a variance component model which examines the genetic contribution to a quantitative trait measured repeatedly over time, combining classical variance components with longitudinal data analytic methods to focus specifically upon change in the trait. The model includes components for both the mean level of the trait as well as the slope of the trait over time. Random effects in the model represent deviations for each individual from the population mean and slope. The variance of these random effects is partitioned into genetic and environmental factors, with familial correlations determining the genetic component and individual correlations determining the environmental covariance structure. This approach can be readily extended to non-linear change over time. We apply this method to low lipoprotein density levels measured repeatedly in the Framingham Heart Study over 20 years of follow up.

106

A New Weighted Mean Proportion of Marker Alleles Identical by Descent Test for Linkage in Affected Sib Pairs

S. Shete and C. I. Amos
Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, TX

In affected sib pair methods for linkage it has been shown that the mean proportion of marker alleles shared identical by descent by affected sibpairs is most powerful test for detecting linkage among several other affected sibpair tests. However, it is not always possible to obtain the ibd sharing information between the sibs unambiguously even when the parental genotypes are known. Here we define a measure of informativity for obtaining ibd sharing based on the parental mating and sibpair genotypes. This measure reflects the precision with which one can obtain the ibd sharing information between sibs based on family data. Then we define a new weighted mean proportion of alleles shared ibd by sibpairs test statistics. In this we assign precision for estimating the ibd sharing as weights to the proportion of alleles shared ibd by affected sibpairs. The simulation results and its implications are discussed.

107

Vagal Hypertonia: Mode of Inheritance and Relation with Sudden Infant Death Syndrome

T. Shojaei-Brosseau[1], C. Bonaïti-Pellié[1], S. Lyonnet[2], J. Feingold[3], and V. Lucet[4]
[1]Inserm U521, Villejuif; [2]Inserm, U393, Paris; [3]Inserm U155, Paris; [4]Centre de Cardiologie Infantile du Château des Côtes, Les Loges-en-Josas, France

Vagal Hypertonia (VH) is caused by dysfunction of vegetative nervous system. It is defined by the occurrence of breath holding spell and repetitive syncope in specific circumstances, without any cardiac or neurological disease. The main objective of this study was to look for a possible role of VH in sudden infant death syndrome (SIDS) and to evidence genetic factors accounting for the observed familial aggregation of VH. Ninety-one children who came for medical advice in the cardiology center of Loges-en-Josas (France) were registered between July 1998 and February 1999. The family history could be investigated in 67 of them through a self-questionnaire completed by a telephone interview. Using standardized incidence ratio (SIR), we compared the frequency of SIDS among relatives of probands to the general population. Segregation analysis was performed using the unified model (POINTER software) for familial aggregation of VH and the parameters were re-estimated under a single locus model of inheritance using a classical method of segregation analysis. An excess of SIDS was observed among relatives of probands compared to the general population ($p<0.00001$). Segregation analysis was consistent with dominant inheritance of VH. The penetrance was estimated to be 62 % and the proportion of sporadic cases was estimated to be 30 %. These results suggest that VH may be a risk factor of SIDS and that a single dominant gene may explain familial aggregation of VH.

108

The Effect of Non-Random Ascertainment on the Power of the Variance Components Method for Linkage Analysis

Yin Y. Shugart[1], Alexa J.M. Sorant[2], Rasika A. Mathias[1], Elizabeth W. Pugh[1], Alexander F. Wilson[2]
[1]CIDR, Johns Hopkins University; [2]NIH/NHGRI, Baltimore, MD

Although the variance components (VC) approach to linkage analysis is now widely used for mapping quantitative traits, the statistical properties of the method are not yet completely characterized. We recently evaluated the type I error, power and estimated size of the modeled effect in randomly selected nuclear families using the VC approach as implemented in GENEHUNTER 2. In the present simulation study, we investigated the effect of violating the assumption of random sampling. We considered two different ascertainment methods, including (1) selecting families with at least two sibs with phenotypic values exceeding a given threshold and (2) selecting extremely discordant sib pairs (one sib whose phenotype exceeds a high threshold and one whose phenotype is below a low threshold). For each ascertainment method and for different trait heritabilities, 2000 samples of 100 nuclear families with sibship size 4 are simulated and analyzed. Preliminary results indicate that non-random sampling by extreme discordant sibling pair (EDSP) selection reduces the power of the VC method and inflates the type I error rate.

109

Ascertainment Bias in Family Studies for Estimating Gene-Environment Interaction Effects

Kimberly D. Siegmund and Bryan Langholz
University of Southern California, Los Angeles, USA

Position (nearest marker)A study of gene-environment interaction using family matched case-control data is susceptible to ascertainment bias. Selection of related controls from outside the case ascertainment region can lead to biased parameter estimates if the prevalence of

the environmental exposure differs across the regions from which the cases and unaffected relatives are selected. Suppose we are studying the effect of sun exposure on melanoma. All cases are ascertained in Los Angeles, California, a region with high sun exposure, and all matched sibling controls are selected from outside Southern California in regions with much lower sun exposure. As a result of ascertainment, region will be a confounding variable related to both sun exposure and disease occurrence. Using conditional logistic regression, the standard analysis for matched case-control data, we study the effect of this (biased) sampling design on estimating genetic and environmental main effects and gene-environment interaction effects. Although this sampling will yield large biases (100%+) in estimates of the main environmental effect, we find that if genetic and environmental exposures are independent, biases in estimates of the genetic main effect and gene-environment interaction effect can be small (1-2%). However, in a selected (extreme) example we overestimated the interaction relative risk (RR) ratio by 19% (estimated RR ratio = 3.2 vs true RR ratio = 2.7). As a result, we recommend selecting unaffected sibling controls from the same region as the case whenever possible.

110

The Effect of "Stoppage" on Segregation Ratio Estimation

S.L. Slager[1,2] and S.E. Hodge[1,2]
[1]Columbia University, New York; [2]NY State Psychiatric Institute

Standard segregation analysis assumes that the observed family size distribution (the distribution of the number of offspring among nuclear families) is not affected by trait parameters such as the segregation ratio, p. We investigate a situation, called "stoppage", that violates this assumption. Brookfield et al. (*J. Med. Genet.* 25:181,1988) defined stoppage in terms of a probability, d, that a family will stop having children after the birth of an affected child. Extreme stoppage occurs when d = 1. We demonstrate how stoppage can be viewed as a form of sequential "within-family" sampling and thus a complication of the ascertainment issue typically encountered in segregation analysis. Under "random" ascertainment (i.e., all families have equal probability of being ascertained, even families with no affected children), we prove that the presence of stoppage does not bias estimates of p. However, for other ascertainment schemes, we turn to simulations. We simulated 200 datasets, each with 200 nuclear families having anywhere from 1 to 11 offspring. The true p was set to 0.1, 0.25, or 0.5, and the ascertainment scheme was either "single" (i.e., $\pi>0$) or "truncate" ($\pi = 1$). The table below shows our results, based on likelihoods that assume the correct ascertainment scheme, but ignore stoppage.

Estimates of p for selected true values of p, d, and π

	p = 0.1		p = 0.25		p = 0.5	
d	$\pi\to0$	$\pi=1$	$\pi\to0$	$\pi=1$	$\pi\to0$	$\pi=1$
0	0.100	0.101	0.251	0.250	0.502	0.499
0.2	0.081	0.083	0.187	0.213	0.385	0.440
0.8	0.034	0.036	0.091	0.093	0.220	0.241

Our results show that ignoring the effect of stoppage can drastically bias the segregation ratio estimate when ascertainment is not "random".

111

Multipoint Mapping of QTL in Arbitrary Pedigrees Accounting for Polygenic Effects

B.R. Southey[1] and R.L. Fernando[2]
[1]University of Illinois, Urbana, IL, USA; [2]Iowa State University, Ames, IA, USA

Detection of genes influencing quantitative traits in outbred populations is complicated by complex pedigree structures, unknown phase and unknown mode of inheritance. A method that combines the techniques of complex segregation analysis and multipoint mapping was developed for arbitrary pedigree structures. The computation of the likelihood is based on the Elston-Stewart algorithm. Polygenic sources of variation are incorporated using two alternative approximations for efficient computation.

Simple and complex pedigree structures of approximately 650 individuals over three generations were simulated using a monogenic additive genetic model. A single QTL of varying magnitudes was simulated in the center of an interval of four markers located 10cM apart. Similar results were obtained for both types of pedigrees. In general, parameter estimates were similar to the true values when the simulated QTL had a large effect. With QTL of smaller effect (less than 10% of the phenotypic variance) QTL location and genotypic means were poorly estimated indicating insufficient power. In addition, a simple pedigree scenario was simulated with a polygenic trait being controlled by ten unlinked QTL. Estimates from the polygenic approximations were closer to the simulated values than those from the monogenic model especially for the residual variance. This was expected since the monogenic model cannot account for the presence of the additional QTL. The precision of the QTL location estimate from the polygenic approximations was not improved over the monogenic model. The developed method provides a flexible approach for QTL detection that can be easily applied to a wide range of situations typically found in outbred populations.

112

Genetic Epidemiology of Diabetes with Onset within One Year of Life

M.A. Stazi[1], D. Iafusco[2], R. Cotichini[1], F. Prisco[2], M. Cotellessa[3], M.E. Martinucci[4], M. Mazzella[3], V. Cherubini[5], F. Barbetti[6], F. Cerutti[7], and Italian Study Group on Early Onset Diabetes
[1]Istituto Superiore di Sanità, Rome; [2]University of Naples, [3]University of Genoa, [4]University of Florence, [5]University of Ancona, [6]University of Milan, [7]University of Turin

Insulin dependent diabetes mellitus (IDDM) diagnosed in the first few months of life is a rare condition. Its pathogenesis is heterogeneous and, in the majority of cases, unknown. Clinical, genetic and epidemiological features of 111 patients (56% males, 44% females) who developed permanent IDDM and started insulin therapy in their first year of life were studied in order to determine possible different markers between non-autoimmune likely forms and autoimmune IDDM. This is up to now the largest population based cohort of diabetic infants ever collected. The epidemic curve by age at onset of diabetes showed a clear minimum at 180 days: these two sets of patients were compared in the analysis. Mothers of children with early onset IDDM (within 180 days of life) reported a higher percentage of threaten abortion (OR=4.4, 95%CI 1.2-16.7). Small for date weight at birth was also more common in the early onset group (OR=10.8, 95%CI 2.8-45.0). Patients were molecularly typed for HLA class II DQB1 and DQA1 loci; *nonAsp* at position 57 of DQB1 or *Arg* at position 52 of DQA1 were considered at high risk for IDDM susceptibility. Seventy percent of patients with IDDM diagnosed within 180 days showed a "protective" HLA genotype (0 or 1 *nonAsp/Arg* etherodimers) for autoimmune diabetes (OR=0.05, 95%CI 0.0-0.6). Italy is the European country with the highest difference of IDDM incidence between two areas: Sardinia, 33/100,000 per year and continental Italy, 5/100,000 per year. Of the 36 children with early onset IDDM, only one was born in Sardinia (OR=0.08, 95%CI 0.01-0.7). In our opinion these results support the hypothesis that early onset IDDM cases differ from those with later onset and they are likely to be pictured in a different pathogenic scenario.

113

Meta Analysis of Relative Penetrance Rank Order Statistics with Application to HLA DR-DQ Genes and Type 1 Diabetes

G. Thomson
University of California, Berkeley, CA, USA

The direct involvement of the HLA DR-DQ genes (DRB1, DQA1, and DQB1) in type 1 diabetes is well established. These genes display a complex hierarchy of predisposing, intermediate, and protective effects at the genotype and haplotype levels. The ratio of the observed frequency of a genotype (haplotype) in patients over the frequency in controls, referred to as the P/C (patient/control) ratio, is an MLE of the relative genotype (haplotype) penetrance values. A novel test has been developed to compare the relative rank orders of predisposing through protective P/C ratios of genotypes and haplotypes across ethnic groups. The algorithm developed to determine the probability of an observed rank order in a population compared to a putatively "known" rank order, allows for the fact that not all genotypes (haplotypes) will be found in every population. The key to development of the algorithm is use of a recurrence relationship to determine the probabilities when an additional genotype (haplotype) is added to the analysis. Meta analysis of the resulting p values across populations is weighted by sample size. Consistency in rank order of P/C ratios of HLA DR-DQ genotypes and haplotypes in type 1 diabetes is seen across ethnic groups. This allows investigation of the specific amino acids at the HLA DR-DQ genes involved in type 1 diabetes. The rank order method developed is applicable to other genetic regions besides HLA.

114

The Effect of Allele Frequency Misspecification on Affected Sibpair Linkage Studies

H.K. Tiwari[1], B. Thiel[1], D. Fallin[1], and N.J. Schork[1,2,3]
[1]Case Western Reserve University, Cleveland, OH, USA; [2]The Genset Corporation, La Jolla, CA; [3]Harvard University School of Public Health, Boston, MA

Affected sibpair (ASP) linkage analyses are often used to identify the rough genomic location of disease predisposing loci. Although of tremendous scientific value and sampling convenience, ASP designs can be problematic when parental genotype information is not available. This is due to the fact that in order to compute relevant allele sharing probabilities in the absence of parental genotype information, one must rely on accurate marker allele frequencies. In this paper we investigate the effect of marker allele frequency misspecification on ASP test statistics using simple analytic derivations. We consider the case of a single biallelic marker locus for two situations. The first situation involves general allele frequency misspecification. The second situation involves allele frequency misspecification of the type that would inevitably arise in samples of sibpairs manifesting cryptic stratification. We show that the false positive rate for ASP tests that make use of misspecified marker allele frequencies can be substantial in certain situations. We

discuss the implications of our results and also comment on the degree to which multipoint strategies and the use of multiallelic marker loci can overcome misspecification problems.

115

Affected Sibpair Linkage Analysis in Inbred Populations with Missing Parental Genotypes

A.A. Todorov[1], E. Génin[2], M. Corbex[3], Z. Xie[1], D. Tomov[1], and A.C. Heath[1]
[1]Washington University School of Medicine, St. Louis USA; [2]INSERM U535, Le Kremlin-Bicêtre, France; [3]LGN, CNRS, Paris, France

Testing for linkage in inbred populations may be more powerful than in outbred ones since not only the affected sibs but also their parents would tend to share more alleles identical by descent (IBD) at the susceptibility locus. It is thus of interest to estimate the IBD state probabilities of the parents of the ASP and to account for this information in a test of linkage.

We describe a method for testing for linkage with affected sibpairs (ASP) in inbred populations, when the sample consists of nuclear families with incomplete genotypic information on parents. The approach does not assume prior knowledge of inbreeding coefficients. The relevant parameters (probabilities of IBD states for parents, probability that an ASP is in a particular IBD state, given the parents') are estimated using the EM algorithm. The linkage test then consists of testing whether the transmission probabilities significantly depart from what would be expected under the hypothesis of no linkage.

Second, we quantify the loss in power due to the inclusion of families with an ASP and several additional sibs, but no available parents. For the models we considered, accepting families with no parents but 4+ unaffected offspring was acceptable as long as the study does maintain a substantial proportion of families with two parents.

116

Longitudinal Quantitative Genetic Analysis of Childhood Skeletal Maturation

B. Towne[1], J.S. Parks[2], M.R. Brown[2], R.M. Siervogel[1], A.F. Roche[1], and J. Blangero[3]
[1]Wright State University School of Medicine, Dayton OH; [2]Emory University School of Medicine, Atlanta, GA; [3]Southwest Foundation for Biomedical Research, San Antonio, TX

Analysis of longitudinal familial data offers the potential to explicitly model age-related changes in the genetic regulation of a trait over time. We incorporated parametric correlation functions in the multivariate variance components-based SOLAR package to model the heritability of skeletal maturity in children aged from 3 to 15 years, and the genetic and environmental correlations between skeletal maturity assessments across this age range. Mixed cross-sectional and longitudinal data consisting of 6,893 radiographic skeletal age (SA) assessments of 807 children in 192 kindreds in the Fels Longitudinal Study were simultaneously analyzed. Models positing a stable heritability of SA from chronological ages 3 to 15 years, and constant genetic and/or environmental correlations between SAs across this age range, were rejected. The heritabilities of SA at each chronological age were: 3=0.71, 4=0.73, 5=0.77, 6=0.93, 7=0.78, 8=0.77, 9=0.73, 10=0.63, 11=0.45, 12=0.39, 13=0.34, 14=0.23, and 15=0.11. The genetic correlation matrix showed gradually decreasing correlations between SA at different chronological ages as age differences increased, while the random environmental correlation matrix showed a pronounced pattern of decreasing correlations. These results show a high heritability of SA through early puberty, and suggest that skeletal maturation at different stages of development is influenced by different sets of genes and environmental factors.

117

Summed vs. Averaged Lod Scores: Which Represents the True Evidence for Linkage Based on Multiple Independent Data Sets?

V.J. Vieland, J. Huang, and K. Wang
University of Iowa, Iowa City, IA

For complex disorders, it may be impossible to achieve definitive evidence of linkage in a single data set, raising the question of how best to measure linkage evidence across multiple independent data sets (D_i's). One intuitively appealing approach is to pool all D_i's and calculate a single summary statistic at each locus, e.g., a heterogeneity lod (HLOD; Smith, 1961). This procedure essentially measures the *average* evidence (HLOD-A) across D_i's which might seem like a good method for eliminating any false positive evidence having arisen through sampling variability in the smaller individual D_i's. An alternative is to *sum* maximum HLODs across D_i's (HLOD-S), which might seem problematic since the HLOD-S can never be less than the largest individual HLOD obtained from the separate D_i's, even when there is really no linkage. In spite of these considerations, however, we show that for ASP data, in the presence of inter-sample heterogeneity HLOD-A tends to dramatically underestimate linkage evidence; while HLOD-S provides a greatly superior approximation to a correct LR statistic allowing for inter-sample heterogeneity, with considerably

better power than HLOD-A. Asymptotic results and simulations show that the HLOD-S accumulates just over 1 d.f. per data set, so that its null distribution remains close to 0^2 with *i* d.f. even in realistic sample sizes (N=100 ASPs). We also show that the distribution of HLOD-S is asymptotically the same as the distribution of the sum of maximum MLS statistics [Risch, 1990; with the possible triangle constraint of Holmans, 1993]. Extensions to other forms of the likelihood and to larger, more informative pedigree structures remain to be explored. Nevertheless, these results show that when multiple studies yield different results at the same locus, *summing* the individual maximum lods rather than *averaging* them can provide a far better overview of the true overall strength of evidence.

118

On the Maximization Procedure of the Heterogeneity LOD in GENEHUNTER

K. Wang
The University of Iowa, Iowa City, Iowa, USA

GENEHUNTER is a popular genetics program. However, it sometimes generates misleading parameter estimates for parameters in the heterogeneity LOD score, and the maximized heterogeneity LOD scores are sometimes slightly negative. These observations cast doubt on the reliability of the maximized heterogeneity LOD score produced by GENEHUNTER. In this report, these problems are addressed by investigating the properties of the maximization method adopted by GENEHUNTER. It turns out that misleading parameter estimates are attributable to the maximization method used in GENEHUNTER. Examples indicate this method can generate misleading parameter estimates for data containing little linkage information, as well as data containing strong linkage information. Even though the parameter estimates can be misleading, the maximized heterogeneity LOD score is still close to the true maximum. Performances of several maximization methods are assessed through examples.

119

Algorithmic Improvements to Markov-Chain Monte-Carlo Pedigree Analysis

Geoffrey C. Wedig, Kevin B. Jacobs, and Robert C. Elston.
Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA

Algorithms that directly compute pedigree likelihoods using many linked genetic markers are currently limited by either the number of markers examined simultaneously (O'Connel and Weeks, Nat. Genet. 1995; 11:402-408) or by pedigree size (Lander and Green, Proc. Natl. Acad. Sci. USA 1987; 84:2363-2367), (Idury and Elston, Hum. Hered. 1997; 47:197-202). However, simulation methods have been used effectively when direct computation has proved infeasible, in particular, the Markov-chain Monte Carlo methods proposed by Sobel and Lange (Am. J. Hum. Genet. 1996; 58:1323-1337). Their algorithm utilizes descent graphs to rapidly update pedigree likelihoods for descent patterns permuted after each state transition in the Markov-chain. The speed of any such algorithm is a function of both the fixed costs associated with computing each transition and the methods by which transitions are chosen. The choice of transitions has been well examined. However, little work has been done to refine the basic descent-graph algorithm used to update the pedigree likelihood. We explore several algorithmic improvements, which reduce the time required for each step.

120

Accuracy of the "Mendell-Elston" Approximation Depends on Pedigree Arrangement

Jeff T. Williams[1], Paul Van Eerdewegh[2], and John Blangero[1]
[1]Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, TX;
[2]Genome Therapeutics Corporation, Waltham, MA, USA

The accuracy of the classical "Mendell-Elston" approximation to the pedigree likelihood for qualitative traits depends on the order in which the affected and unaffected individuals within the pedigree are "peeled". For disease prevalences less than 50%, the approximation is most accurate if affected individuals are peeled before affected individuals. In either case, the relative error decreases as disease prevalence increases, and larger pedigrees incur a greater total error. With the large pedigrees used in many quantitative genetic studies, inferences regarding heritability and linkage may be seriously biased simply by suboptimal arrangement of the pedigree members. We illustrate the problem and its effect on parameter estimation using simulated and real data sets.

121

Genomic Adjustment for Population Stratification in Association Studies

J.S. Witte and D.V. Conti
Case Western Reserve University, Cleveland, OH, USA

Population stratification may occur in association studies if subjects come from sub-populations with different rates of disease and allele frequencies at unlinked loci. In

this situation, cases will be more likely than controls to arise from the sub-population(s) with the higher disease rate(s). Furthermore, cases and controls will have different non-causal allele frequencies, possibly leading to false positive or negative associations. The potential for population stratification bias in association studies remains controversial: some argue that it is essentially a non-issue, while others maintain that family-based designs will always be required to assess causality. As a compromise, we propose here using genomic information to evaluate and adjust for this bias. This approach probabilistically clusters individuals into sub-populations, and then incorporates this information as a covariate in a logistic regression model to adjust for potential confounding. Using an agent-based simulation (GenERA), we investigate the use of this approach across a number of different scenarios, and compare it to more traditional approaches, such as ignoring potential stratification, adjusting for ethnicity, and adding to the regression model a panel of unlinked markers. Our results indicate that when population stratification exists: 1) adjusting for "ethnicity" may not be adequate; 2) adjusting for a panel of markers helps address the bias, but can give poor precision; and 3) the genomic adjustment approach appears to work well. Finally, our findings also suggest that population stratification may only be a major concern in relatively extreme situations.

122

Extent of Linkage Disequilibrium in Human Populations and Their Implication in SNP Mapping

Momiao Xiong and Li Jin
Human Genetics Center, University of Texas-Houston Health Science Center, Houston, TX 77225

The extent of linkage disequilibrium (LD) in a population is dictated by its demographic history which is defined by the interplay of several factors including mutation, admixture, migration, and genetic drift. In this paper, we explore the extend of LD in general population, isolated population, admixed population and inbred populations using traditional population genetics as well as simulation-based approaches. Among those aforementioned parameters, the age of mutation (including natural and migration-introduced mutations) and the level of initial LD of the populations generally determine the magnitude of the level of LD. In this study, we carefully chose the parameters based on our current knowledge of modern human populations of various kind and the estimation derived from to the studies of real populations including those we have been recently worked on in East and Southeast Asia. These results demonstrate that, in contrary to the estimation of Kruglyak that in human population, detectable levels of LD will rarely extend beyond 3 kb, the regions in which detectable LD extends about several hundreds base-pairs in "general" populations such as Europeans, which implies that approximately 15,000 SNPs will be required for whole-genome scan studies. In the recent admixed populations such as African-Americans and Mexican-Americans, we show that the detectable level of LD can generally extend over 1cM, which implies that several thousands of SNPs markers are enough for whole-genome studies. In isolated populations, the initial level of LD is dictated by the number of founders. It is likely to ascertain isolated populations in which the detectable LD extends over several cM. Our recent experiences in East and Southeast Asia have shown that the populations that were founded by several founders 10-20 generations do exist.

123

Multipoint Linkage Disequilibrium Fine-Scale Mapping of Quantitative Trait Locus

Momiao Xiong, Jinying Zhao, Li Jin, and Eric Boerwinkle
Human Genetics Center, University of Texas-Houston Health Science Center, Houston, TX 77225

With increasing popularity of QTL mapping in humans, the need for statistical methodology for fine-scale QTL mapping becomes increasingly urgent. Although comparable advances in both methodology development and practice have been made in linkage disequilibrium fine-scale mapping for qualitative traits there is increasing recognition that statistical methods for fine-scale mapping of QTL should be developed. The ability to refine localization of the genes influencing quantitative traits depends on the number of recombination events. One way to increase the recombination events is to use historical recombination. In this paper we develop a general mathematic framework for multipoint linkage disequilibrium analysis of QTL simultaneously using multiple nearby markers which can be used as bases for fine-scale mapping of QTL in randomly sampled individuals and for joint linkage and linkage disequilibrium analyses in general pedigrees. We extend simple regression, interval mapping and composite interval mapping of QTL for plant and animals to humans and demonstrate that simple regression, interval and composite interval mapping of QTL are special cases of the proposed multipoint linkage disequilibrium mapping of QTL. We unify the analysis of mapping QTLs in humans and, in plant and animals. To implement multipoint linkage disequilibrium mapping of QTL we present population genetic models for the haplotype frequencies. We investigate and extensively compare the statistical properties (power, sample size and support interval) of the proposed tests for multipoint linkage disequilibrium mapping of QTLs under various genetic models and parameters. The proposed method is applied to a study of loci influencing interindividual systolic blood pressure variation in a Chinese population.

124

Physiological and Statistical Genetic Models of Epistasis

Momiao Xiong, Jinying Zhao, Christopher Amos, and Eric Boerwinkle
Human Genetics Center, University of Texas-Houston Health Science Center, Houston, TX 77225

Epistasis plays an important role in genetic architecture of a trait and resurface from time to time in the literature. However, most widely used statistic methods for quantitative trait locus (QTL) analysis have focused on identifying individual QTLs and their effects, and ignore the interaction between the trait loci. In this report, we present physiological and statistical genetic models for epistasis and develop computational methods for the estimation of physiological additive, dominance and epistasis values for the physiological model, and genetic additive, dominance effects and additive × additive, additive × dominance, dominance × dominance effects for the statistical genetic model.

We establish relationship between parameters for the physiological model and parameters for the statistical genetic model. We carry out the power comparison of the tests for the presence of epistasis based on physiological and statistical genetic model and demonstrate that the physiological model has higher power to detect epistasis than the statistical genetic model. We extend the physiological model and statistical genetic model for epistasis at the trait loci to the marker loci and demonstrate that the epistatic effects at the marker locus for the statistical genetic model almost vanish even for the markers with the mild genetic distance from the trait locus. We apply the developed physiological and statistical genetic models to mapping QTL and propose general multi-locus physiological and statistical genetic models for QTL analysis. We investigate the power of single-locus and multi-locus physiological and statistical genetic models for mapping QTL. Finally, we give examples to illustrate the application of the proposed models to QTL analysis.

125

Linkage Disequilibrium Mapping Using Genotype Data

Shuanglin Zhang and Hongyu Zhao
Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT

Linkage disequilibrium mapping (LDM) has proven a powerful tool for fine-mapping disease genes. Statistical methods that use multiple markers simultaneously are more powerful than those using markers separately. However, existing multilocus methods all assume that haplotype information is available for statistical analysis. But haplotypes are often difficult to obtain, especially for studies involving many genetic markers and relatively few individuals from each pedigree. To overcome this apparent discrepancy between statistical methods and genetic data that are available to geneticists, we have developed a multilocus approach for LDM using genotype data. Because disease models are usually unknown for complex traits of interest, our model introduces a heterogeneity parameter to capture differences among different disease models. Because this heterogeneity parameter is estimated together with other parameters (e.g. disease gene location and age of mutation) in our analysis, our approach does not depend on any prior knowledge of the disease model. In our maximum likelihood estimates of the model parameters, we treat the unobservable haplotypes as complete data, and the observable genotypes as incomplete data, and use the expectation-maximization algorithm to estimate the model parameters. For haplotype analysis, we employ the Decay of Haplotype Sharing (DHS) method proposed by McPeek and Strahs [1999] and extended by us (Zhang and Zhao, 2000). We assess the performance of our methods through extensive simulations under a variety of disease models, population history, sample sizes, and genetic marker spacing. Our simulation results suggest that the methods are very robust for locating disease genes, even for small sample sizes.

126

Transmission/Disequilibrium Tests Using Multiple Tightly Linked Markers

Hongyu Zhao[1], Shuanglin Zhang[1], Kathleen R. Merikangas[1], Deiter B. Wildenauer[2], Fengzhu Sun[3], and Kenneth K. Kidd[1]
[1]Yale University School of Medicine, New Haven, CT; [2]Department of Psychiatry, University of Bonn, Germany; [3]University of Southern California, CA

Transmission/disequilibrium tests have attracted much attention in genetic studies of complex traits because of the possibility that they may have greater power than other linkage methods to detect genes having small to moderate effects, and they are robust against population stratification. Highly polymorphic markers have recently become available throughout the genome and many such markers can be studied within short physical distances in the human genome. Studies using multiple tightly linked markers are more informative than those using single markers. However, such information has not been fully utilized by existing statistical methods, resulting in possibly substantial loss of information in identifying genes underlying complex traits. In this talk, we will discuss several novel statistical methods to analyze multiple tightly linked markers. Simulation results suggest that our methods are more powerful than existing methods. We then apply the proposed methods to study genetic linkage between the dopamine D2 receptor locus and alcoholism.

127

Disease Causality of Missense Mutations: a Hierarchical Approach

X. Zhou[1], G. Parmigiani[1,2], and E. Iversen [1]
[1]Duke University, Durham, NC, USA; [2]Johns Hopkins University, Baltimore, MD, USA

Missense mutations of disease genes pose a challenging classification problem because of the uncertainty associated with their implications to the risk of disease. Assessing the risk implications is often complicated by small sample size and lack of an appropriate functional assay. For large genes such as BRCA1 and BRCA2, it is common to infer risk implications from pedigree data. It is typical to have a relatively small sample size for each mutation, and to only have pedigrees of individuals who are selected because of a high disease rate in the family. This selection mechanism is likely to overstate the mutation's contribution to risk of disease. In this study, we develop a Bayesian hierarchical methodology which classifies missense mutations as deleterious or non-deleterious based on pedigree data from a high risk clinic. Our work extends currently available techniques in several ways. It assumes an age-dependent penetrance model with multiple cancer sites and accounts for imperfect sensitivity of genotyping. The hierarchical structure enables the systematic comparison of the effects of different mutations and the study of the mutations as a group. To avoid overestimating the risk effects, the model calibrates the classification by using pedigrees identified through probands that are either negative or have known deleterious mutations. We apply this model to the study of a sample of BRCA1 and BRCA2 missense mutations. We provide the probability that each mutation is deleterious and estimate the mutation-specific penetrance functions.

128

Connection Between the Haseman-Elston Method and the Weighted Pairwise Correlation Statistic

Andreas Ziegler
Institute of Medical Biometry and Epidemiology, Philipps-University of Marburg, Bunsenstr. 3, 35033 Marburg, Germany

In this presentation, I discuss similarities and differences of the traditional Haseman-Elston [Behav Genet 1972, 2:3-19] method, the new Haseman-Elston method [Elston et al., Genet Epidemiol 2000, in press] and the weighted pairwise correlation (WPC) method [Commenges, Genet Epidemiol 1994, 2:189-200]. All three approaches can be applied to a wide range of traits including binary, ordered categorical, quantitative or censored phenotypes. They are all based on the principle of similarity. Genetic similarity is preferably measured by marker identity by descent. The measure of phenotypic similarity, however, differs greatly between the methods. Furthermore, they are derived from completely different statistical models. Nevertheless, a connection between the traditional Haseman-Elston method and the WPC statistic has been established. In this presentation, a relationship between the new Haseman-Elston method and the WPC approach will be derived for a sample of independent sib pairs, a diallelic additive trait locus, standardized phenotypes and standardized IBD values.

129

Linkage and Association Analysis of Angiotensin I-converting Enzyme (ACE) Gene Polymorphisms with ACE Activity and Blood Pressure

X. Zhu[1], R. Cooper[1], A. Adeyemo[2], R. Ward[3], N. Boureki[3], A. Luke[1], G. Chen[1], and C. Rotimi[4]
[1]Loyola University of Chicago, Maywood, IL, USA; [2]University College Hospital, Ibadan, Nigeria; [3]University of Oxford, Oxford, UK; [4]Howard University, Washington DC, USA

The angiotensin I-converting enzyme gene is a candidate susceptibility gene for cardiovascular disease. Conflicting evidence exists, however; this could result from using linked markers like the insertion/deletion (I/D) motif rather than functional polymorphisms or tightly linked variants. Access to all variants improves the precision of this analysis. To address the ACE-BP relationship we examined 1,343 rural Nigerians from 342 families and genotyped 13 polymorphisms in the ACE gene. To localize the genetic effect we first performed linkage and association analysis of all the markers with ACE activity. In multipoint variance components analysis this region was significantly linked to ACE activity (LOD=7.2). Likewise, all the polymorphisms were significantly associated with ACE activity ($P < 0.003$), except for those where the less common allele had a frequency < 0.1. The two most highly associated polymorphisms were A-240T and G2350A ($p < 2.6H10^{-10}$), which accounted for 4% and 19% of the variance, respectively. Both polymorphisms were significantly associated with ACE activity after adjusting for the effect of the other variant ($P < 10^{-6}$). Because these markers are separated by 12 kb at least two separate ACE QTL loci are likely to exist. Using this localized genetic information about the ACE phenotype, similar analyses were performed with systolic and diastolic blood pressure and no significant relationships were found. If the effect of this gene operates through circulating ACE activity, it is unlikely to be a susceptibility gene for high blood pressure.