Genetic Epidemiology 23:264–314 (2002)

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

NEW ORLEANS, LOUISIANA, USA
NOVEMBER 15–16, 2002

IGES-1

A genome-wide scan identifies a major locus predisposing to leprosy

A. Alcaïs[1], M.T. Mira[2], T. Pietrantonio[2], N.V. Thuc[3], M.C. Phuong[3], E. Schurr[2], L. Abel[1]
[1]Human Genetics of Infectious Diseases, INSERM U.550, Necker Medical School, Paris, France; [2]McGil centre for the Study of Host Resistance, McGill University Montrea, Canada; [3]Hospital for Dermato-Venereology, Ho Chi Minh City, Vietnam

Leprosy is a chronic infectious disease that is still a major global health problem with 700,000 new cases occurring each year. The disease presents itself in different clinical manifestations, ranging from single lesion paucibacillary to severe, multiple lesion multibacillary forms. In order to identify genomic regions linked with susceptibility to leprosy, we performed a genome-wide scan in families from South Vietnam. These families contain both paucibacillary and multibacillary cases allowing for the genetic analysis of susceptibility to leprosy "per se" and leprosy type. A total of 87 multiplex families were genotyped for 395 highly informative microsatellite markers spanning the entire human genome. Genetic model-free linkage analysis, using the Maximum Likelihood Binomial (MLB) method initially pointed to 11 genomic regions showing suggestive evidence for linkage with either leprosy per se or leprosy type (MLB lod score >1.5). Analysis of the high density marker maps (information content $>95\%$) revealed strong evidence for linkage between one chromosomal region and leprosy "per se" (MLB lod score of 4.21; $p = 5.10$–6). Microsatellite markers located in the same region have been used to replicate this result in a family-based association study using an independent sample of 208 simplex Vietnamese pedigrees ($p = 6.10$-5). In addition, the HLA region presented an MLB lod score of 2.62 ($p = 2.10$-4) when tested for leprosy "per se" and 2.64 when tested for leprosy type. These results strongly suggest the existence of at least one major locus for the control of susceptibility to leprosy "per se". They are also in accordance with the generally accepted idea of a role for the HLA region in controlling disease susceptibility and clinical form of leprosy disease.

IGES-2

A locus on chromosome 2 influences levels of tissue factor pathway inhibitor: results from the GAIT study

L. Almasy[1], J.M. Soria[2], J.C. Souto[2], A. Buil[2], M. Borrell[2], N. Sala[3], M. Lathrop[4], J. Blangero[1], J. Fontcuberta[2]
[1]Southwest Foundation for Biomedical Research, USA; [2]Hospital de la Santa Creu i Sant Pau, Spain; [3]Institut de Recerca Oncologica, Spain; [4]Centre National de Genotypage, France

Tissue factor pathway inhibitor (TFPI) is a protease inhibitor that inhibits fibrin clot formation through regulation of the extrinsic pathway of coagulation. Levels of TFPI have been associated with both atherosclerotic and thrombotic disease. We studied levels of TFPI in 397 individuals in 21 Spanish families participating in the Genetic Analysis of Idiopathic Thrombosis (GAIT) study. Twelve of these families were selected through a proband with idiopathic thrombosis and nine were ascertained without regard to phenotype. Subjects had a mean age of 37.7 years (range 1 to 88) with approximately equal numbers of males and females. A genome scan was performed using highly

Published online in Wiley InterScience (www.interscience.wiley.com). DOI:10.1002/gepi.10205

informative microsatellite markers spaced at approximately 10 cM intervals. Multipoint variance component linkage analyses included a correction for ascertainment in the 12 thrombophilic families. The additive genetic heritability of TFPI levels was 0.52 ($p<0.0001$), with no evidence for shared household effects. In the genome screen, only one LOD score >2 was observed. On chromosome 2q, the maximum multipoint LOD score was 3.52 near marker D2S1384. This is very close to the structural gene for TFPI which is located at 2q32. We are currently typing SNPs in the TFPI gene to investigate whether it may be the QTL responsible for this linkage finding.

IGES-3

The limits of fine mapping

L.D. Atwood, N.L. Heard-Costa
Dept. of Neurology, Boston University Medical Center, Boston, MA, USA

Fine mapping is the placement of additional markers near a significant linkage, and then repeating the linkage analysis with the goal of improving the location estimate. To test the utility of fine mapping we simulated an additive two-allele locus with minor allele frequency 0.1 located at 75cM on a 150cM chromosomal segment that had 16 polymorphic markers with 10cM spacing between markers. We simulated five quantitative trait models in which the proportion of variation due to the quantitative trait locus (QTL) was 0.20, 0.27, 0.40, 0.80, and 0.90. For each model, variance components linkage analysis (Genehunter) was performed until 1000 replicates were found with maximum lodscore greater than 3.0. For each of these significant replicates we simulated additional markers at three resolutions (2cM, 1cM, and 0.5cM) in a 20cM region centered on the maximum lodscore. Then we repeated the linkage for each of the three fine map resolutions. Results showed that, for the five models in which the trait accounted for 0.20, 0.27, 0.40, 0.80, 0.90 of the variation, fine mapping at 2cM reduced the average location error by 3%, 5%, 20%, 50%, and 53% respectively. Fine mapping at 1cM reduced the average location error by 3%, 7%, 20%, 61%, and 66% respectively. Fine mapping at 0.5cM did not differ appreciably from the 1cM results. This simulation study indicates that the proportion of variation due to the QTL strongly affects the utility of fine mapping. If the QTL accounts for a small proportion of the variation, as is the case for complex traits, then fine mapping has little value.

IGES-4

LD in recently isolated dutch population

Y.S. Aulchenko[1], P. Heutink[1], I. Mackay[2], A. Bertoli[1], J. Pullen[2], N. Vaessen[1], L.A. Sandkuijl[3], L. Cardon[4], J.J. Houwing-Duistermaat[1], B. Oostra[1], C.M. van Duijn[1]
[1]Erasmus MC Rotterdam, The Netherlands; [2]Oxagen Ltd, UK; [3]Leiden University MC, The Netherlands; [4]Wellcome Trust Center for Hum. Genetics, UK

The design and feasibility of whole-genome association studies are critically dependent on the extent and distribution of linkage disequilibrium (LD) across the genome. There is an ongoing debate about the distribution of LD both across the genome and between different (in particular genetically isolated) populations. We examined genome-wide LD in a recently isolated Dutch population and characterized four genomic regions using a dense marker map. We studied 58 spouses of patients identified in our disease oriented research studies. In these subjects, we genotyped 734 autosomal and 47 X-linked Short Tandem Repeat (STR) markers. We also analyzed a 11.9 Mb long telomeric region on chromosome 18p11 in more detail using 15 STR markers. An 1.6 Mb centromeric region on chromosome 3p12 was covered using 8 STR markers and a 12 Mb middle-arm region on chromosome 3p13 using 16 markers. For LD analysis we used a method, which results in sample-size independent estimates of LD and this allowed for comparison between earlier published studies. We find a significant ($P<0.0001$) relation between LD and genetic distance on a genome-wide scale. Distance alone explains 4% of total LD variation. More importantly, LD was still detectable at large distances up to 20 cM. Comparison with previous studies on LD in genetically isolated populations (Palau and the Central Valley of Costa-Rica) revealed notable similarity in genome-wide distribution of LD between these populations. For fine-mapping distance was explaining up to 39% of LD (chromosome 3p12). For chromosome 3p123-p13 and 18p11 we found very strong LD between markers separated by less then 1 Mb. Our data suggest that LD is substantial in our genetically isolated population.

IGES-5

Interpretation of European celiac disease linkage studies

M.C. Babron[1], F. Clerget-Darpoux[1], H. Ascher[2], P. Ciclitira[3], L. Sollid[4], J. Partanen[5], L. Greco[6], The European Genetics Cluster on Celiac Disease
[1]INSERM U535, Le Kremlin Bicêtre, France; [2]Göteborg U., Sweden; [3]St Thomas' Hospital, London, UK; [4]Oslo U., Norway; [5]Finnish Red Cross, Helsinki, Finland; [6]U. Federico II, Naples, Italy

The identification of genetic risk factors for multifactorial diseases is often carried out through systematic linkage analysis on the whole genome by different groups. Each study highlights regions of interest, but nevertheless rarely achieves the genome-wide level of significance. Meta-analysis methods which aim at globally interpreting linkage results, then appears a promising tool, as they may significantly increase the power of linkage studies. We applied a modification of a meta-analysis method, GSMA (Wise et al, 1999; Wise 2001) on

4 genome scans and several follow-up studies carried out by the partners of the European Cluster on Celiac Disease (CD). The principle of GSMA is to split the genome into bins of equal size, to rank the bins according to the value of linkage statistics obtained in each study, and then to evaluate the p-value associated with the summed rank by simulations. GSMA was modified to take into account the arbitrariness of bin cut-off points, as well as the sample size of each study. Besides the HLA region, already known to harbor a risk factor for celiac disease, this meta-analysis leaves no more doubt on the presence of a genetic risk factor in the 5q31-33 region. This region was suggested by several studies, but did not reach statistical values high enough to be conclusive when data sets were analyzed separately. Here, the significance is at the 0.1% level. Another region on chromosome 17, significant at the 5% level, might be of interest. This study was funded by the Commission of the European Communities (QLRT-1999-00037).

IGES-6

Clinical heterogeneity in genomic scans of hereditary prostate cancer identified by logistic regression

M.D. Badzioch[1], J.L. Stanford[2], M. Janer[3], D.J. Schaid[4], S. Kolb[2], D. Friedrichsen[2], E.L. Goode[1], M. Gibbs[2], L. Hood[3], E.A. Ostrander[2], G.P. Jarvik[1]
[1]Univ. of Washington, USA; [2]Fred Hutchinson Cancer Research Center, USA; [3]Inst. For Systems Biology, Seattle, WA, USA; [4]Mayo Clinic, Rochester, MN, USA

Linkage studies of hereditary prostate cancer (HPC), a genetically heterogeneous disease, have had varied results. In addition to stratification by possible risk factors, methods to incorporate covariates into heterogeneity calculations are available. We report here results of a logistic regression procedure (Schaid 2001) used to assess lod score variation at 380 DNA markers in 94 HPC families for association with 5 potentially relevant clinical factors, age at diagnosis (A), disease aggressiveness (D), Gleason score (G), male-to-male transmission (M), and number affected (N). These factors were modeled with and without A. P-values for the covariates are based on the likelihood ratio test and assume the presence of heterogeneity. Significant results, $p<0.01$, were found at D1S1677 (A), D4S2366 (A+G), D11S2371 (M), D15S642 (A, A+G, A+M, and A+N) D16S764 (A+D), and DXS7132 (A) under the dominant model and at D2S1326 (A+D) D8S1119 (N), D10S1237 (N, A+N), D12S1024 (N, A+N), and D16S2624 (N) for the recessive. Prior to regression, only D12S1024 had a lod score >1. D8S1119 and D16S764 are at and D1S1677, D4S2366, and DXS7132 are 0-13cM from regions detected by Goddard (2001) with non-parametric covariate methods. DXS7132 is ~ 2 cM from the androgen receptor locus. Schaid DJ, et al., Am J Hum Genet 2001;68:1189. Goddard KAB, et al., Am J Hum Genet 2001;68:1197.

IGES-7

Evidence for a combined prostate-breast cancer susceptibility locus

A.B. Baffoe-Bonnie[1,4], B. Verhage[2], L.A.M. Kiemeney[2], T.H. Beaty[3], J.E. Bailey-Wilson[4]
[1]Division of Population Science, Fox Chase Cancer Center, Philadelphia, PA, USA; [2]Departments of Epidemiology & Urology, University of Nijmegen, The Netherlands; [3]The Johns Hopkins School of Hygiene and Public Health, Baltimore, MD, USA; [4]Inherited Disease Research Branch, National Human Genome Research Institute of the National Institutes of Health, Baltimore, MD, USA

Previous studies have shown evidence for a major gene that could help to explain the excess clustering of prostate cancer cases in families of breast cancer probands in the Icelandic population. This suggested that in addition to BRCA2, another gene(s) may confer excess risk to both breast and prostate cancer (Baffoe-Bonnie et al. in press). The most parsimonious model was a Mendelian codominant model, which could partially explain the familial aggregation of both cancers. We are testing the hypothesis of the combined breast-prostate cancer phenotype with a segregation analysis on 704 pedigrees ascertained through prostate cancer probands from the Netherlands. This may help to elucidate putative loci that possibly control these two common cancers. In these data, a clear inheritance model to explain the clustering of prostate cancer was not found, although a recessive gene could explain prostate cancer clustering in families of young probands (under the age of 66) with a non-screen detected tumor. Results of our current combined prostate and breast cancer analysis will be presented.

IGES-8

Similar genes effect the expression of 5HIAA and HVA in Vervet monkeys

J.N. Bailey[1], J.J. Mann[2], Y. Huang[2], M.J. Jorgensen[1], L.A. Fairbanks[1]
[1]Center for Primate Neuroethology, Neuropsychiatric Institute, UCLA, USA; [2]Department of Neuroscience, The New York Psychiatric Institute, USA

Brain monoaminergic activity is important in many behaviors and has also been related to impulsivity, aggression, and behavioral reactivity in nonhuman primates. In this study, the Vervet Research Colony was used to assess the heritability and genetic covariance of brain monoaminergic activity. The monkey colony contains a large, multigenerational pedigree that includes complex extended kin relationships across 7 generations. Monoaminergic activity was measured through cisternal cerebrospinal fluid (CSF) levels of the metabolites of serotonin (5 hydroxyindolacetic acid

or 5HIAA), dopamine (homovanillic acid or HVA) and norepinephrine (3-methoxy-4hydroxy-phenylglycol or MHPG) on 576 vervet monkeys. Cerebrospinal fluid was removed within 20 minutes of capture and reflects baseline or normal expression. All of the monoamine metabolites had a significant genetic heritability; the heritability of 5HIAA was 0.70 ($p<0.007$), HVA was 0.66 ($p<0.001$), and MHPG was 0.60 ($p<0.001$). As expected age and sex contributed to the variation for all three monoamines. Bivariate analysis suggests that the expression of 5HIAA and HVA are due to similar genes that have a pleiotrophic effect. The genetic correlation is high ($\rho = 0.83$) and statistically significant ($p = 0.008$). There was no genetic relationship seen between 5HIAA and MHPG ($\rho = 0.32$, $p = 0.29$) nor between MHPG and HVA ($\rho = 0.18$, $p = 0.57$). The Vervet Research Colony offers a great opportunity to study the genetics of complex in vitro systems.

IGES-9

Quantitative trait analysis of angiotensin-converting enzyme (ACE) levels in relation to haplotypes in the ACE gene

J.H. Barrett[1], G.I. Rice[2], P.J. Grant[2]
[1]Cancer Research UK Genetic Epidemiology Division, University of Leeds, UK; [2]Academic Unit of Molecular Vascular Medicine, University of Leeds, UK

Methods have recently been developed for testing for association between a quantitative trait and haplotypes where phase is unknown. One approach is to estimate haplotype frequencies and then apply a weighted linear regression to test for association with the quantitative trait (1), while another approach is based on score tests (2). We apply these methods to investigate the relationship between plasma ACE levels and 8 biallelic polymorphisms in the ACE gene. A total of 740 unrelated subjects have been genotyped and ACE levels are available on 422 subjects. After log transformation plasma ACE levels are approximately normally distributed. The polymorphisms considered span a 10 kb region and are in tight linkage disequilibrium (LD). Using the E-M algorithm on all the data it was estimated that only 7 distinct haplotypes (out of a possible 256) account for 92% of all haplotypes. Two polymorphisms were excluded from further analysis since they were in complete LD with a third polymorphism. A further two polymorphisms show no association with phenotype (log transformed ACE levels). Analysis based on four-locus haplotypes indicates that two of the polymorphisms account for the strong association between genotype and phenotype. Detailed results and comparisons between the different methods of analysis will be presented. (1) AP Mander. The Stata Journal 2:65–70, 2002 (2) DJ Schaid et al., Am J Hum Genet 70:425–434, 2002.

IGES-10

Allele-sharing methods in large pedigrees

S. Basu[1], E.M. Wijsman[2], E.A. Thompson[1]
[1]Dept. of Stat, Univ of WA, USA; [2]Div. of Med. Genet, Univ of WA, USA

Allele-sharing methods provide a robust approach to linkage detection for complex traits using pedigree data. Affected related individuals have increased probability of sharing genes identical by descent (ibd) at trait loci, and hence also at linked marker loci at which they therefore show increased allele-sharing over that predicted by their relationship alone. Relatives of discordant phenotype have decreased probability of sharing genes ibd at trait loci and hence have decreased similarity at linked markers. Many ibd allele-sharing statistics have been developed for small pedigrees, but for large pedigrees with multiple related affected individuals, these either do not make efficient use of pedigree information available or else become computationally intractable. We propose a new ibd-statistic for use in large multi-sibship pedigrees. The statistic uses data on both affected and unaffected individuals and is computationally tractable. Since exact computation of ibd probabilities conditional on marker data at multiple linked markers is infeasible on large pedigrees, we use Markov chain Monte Carlo methods to provide a consistent estimate of the statistic. A resampling technique provides an upper bound on the variance of the statistic under the null. The effect of this over-conservative standardization is discussed and illustrated and a permutation approach is proposed as an alternative. These methods will be presented together with a comparison of ours and other ibd statistics. Examples of application on both simulated and real data will be presented.

IGES-11

Exploring haplotype sharing analysis with SNPs in candidate genes using the GAW 12 simulated data

L. Beckmann[1], C. Fischer[2], G. te Meerman[3], P.L. Majoram[4], J. Chang-Claude[1]
[1]German Cancer Research Center, Heidelberg, Germany; [2]Institute of Human Genetics, University of Heidelberg, Germany; [3]Department of Medical Genetics, University of Groningen, The Netherlands; [4]Department for Preventive Medicine, University of Southern California, Los Angeles, CA, USA

We employed haplotype sharing analysis to successfully map candidate gene 1 with direct influence on affection using highly polymorphic microsatellites. We now use this method to investigate the association of affection status with SNP haplotypes in more than 70 SNPs in gene 1 in the isolated and the general population of the GAW 12 simulated data. Haplotype sharing analysis depends heavily on reliable haplotype construction. Using Genehunter haplotypes strong evidence was found for most SNPs in the large sample of the isolated

population and thus provide evidence for an involvement of this gene, but the maximum −log10(p) values for the HSS test statistic did not correspond to the location of the true variant. The pattern of linkage disequilibrium between the causal variant and other SNPs within gene 1 reflects the relative short history of the isolated population. Without sufficient haplotype decay, haplotype sharing analysis cannot contribute to further fine mapping within the gene. In this example, TDT analysis appears to perform better than HSS in identifying the disease-causing variant using SNPs within a candidate gene in an outbred population. The dependency of HSS fine mapping characteristics on population history and haplotyping method will presented by simulation results using an ancestral recombination approach with SNPs.

IGES-12

Covariate adjustment for haplotype sharing analysis using mantel statistics

L. Beckmann[1,2], P. Marjoram[2], D.C. Thomas[2]
[1]German Cancer Research Center, Heidelberg; [2]Dept. of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

For complex traits, adjustment for disease covariates may improve the power of Haplotype Sharing Statistics for localising disease susceptibility genes. Mantel (Cancer Research, 27:209–220, 1967) introduced a class of statistics for the purpose of testing space-time clustering of disease, based on a test of the form $T = \Sigma_{ij} X_{ij} Y_{ij}$, where X_{ij} denotes the "similarity" of a pair of cases in space and Y_{ij} their similarity in time, and used a permutation test to determine the significance of the observed clustering. Here, we apply this general approach to haplotype sharing in case-parent trios, taking $X_{ij}(x)$ to be the genetic similarity of a pair of haplotypes at a marker position x and Y_{ij} a measure of phenotypic similarity. The basic idea is that individuals who have similar phenotypes also have similar haplotypes around a disease causing variant. A number of different variants are considered, depending on which pairs are included in the summation, and the definition of phenotypic similarity, incorporating covariates in different ways. By permuting the phenotypes among the individuals, we derive an empirical null distribution for the observed test statistic T. Our results with simulated data show that Mantel statistics provide us with a flexible tool. Summing over all pairs of 119 transmitted haplotypes, the statistic gives $p < .022$ at the correct marker position. Alternatively summing over all pairs 118 haplotypes from individuals who are exposed of $p < .001$ at the correct marker position.

IGES-13

Simultaneous localization of two linked disease susceptibility genes

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

Simultaneous consideration of multiple susceptibility genes may increase the power to detect genes involved in the predisposition to complex disorders. Although several authors have proposed methods to test for the existence and/or interaction of secondary genes taking into account the presence of primary genes, most of these methods have assumed that the second disease gene is not linked to the first one. Furthermore, the focus has been on testing for potential interactions between two genes, rather than localizing them. In some genome scan linkage analyses broad peaks are observed. Such peaks could be the result of multiple susceptibility genes found in the same chromosomal region. Extending the work of Liang et al. (Human Heredity 51, 2001, 64–78), we have developed a model for the simultaneous localization of two susceptibility genes in one region. We derived an expression for expected allele sharing in affected sib pairs across a chromosomal segment containing two susceptibility genes. With this formulation, using information on marker IBD sharing, the generalized estimating equation (GEE) approach can be used to estimate the locations of both disease genes simultaneously. The main difficulty with this approach arises from the complexity of the variance function. We compare approaches to dealing with this difficulty. The proposed method may aid in separating large peaks into two when they are not too close, however our findings suggest that minor genes located near a major gene would be nearly undetectable.

IGES-14

An alternative approach to the A-test for estimating linkage parameters in the presence of heterogeneity

S. Biswas, S. Lin
Department of Statistics, Ohio State University, USA

Admixture test (A-test) is a widely used method for analyzing linkage data when locus heterogeneity is suspected. It has already been shown that estimates obtained under such formulation are biased when the linked and unlinked families do not have the same linkage information according to a certain measure of informativeness based on LOD score. We show that the estimates can be biased even if the two kinds of families have the same linkage information according to the measure. We consider an alternative approach where we form groups of families that have the same distributions. For each group we use a different parameter to denote the level of heterogeneity in that group. After formulating the mixture likelihood using this approach we apply three maximization algorithms to obtain the maximum likelihood estimates. These algorithms are Expectation Maximization (EM), Classification Expectation Maximization (CEM) and Stochastic Expectation Maximization (SEM). Our simulation study shows that SEM performs best, EM a little worse while CEM performs poorly. These simulations as well as analytical analyses show that when different types of families have different levels of

heterogeneity, this approach gives unbiased estimates in contrast to the A-test approach. Our approach can be used for two-point as well as multi-point analysis. We also obtain standard errors of the EM and SEM estimates based on methods available in the literature.

IGES-15

Genome-Wide linkage analysis of Apo-A1 and ApoB in the sedentary state and the response to training: the HERITAGE family study

I.B. Borecki[1], M.F. Feitosa[1], T. Rankinen[2], J.P. Després[3], A.S. Leon[4], J.S. Skinner[5], J.H. Wilmore[6], C. Bouchard[2], D.C. Rao[1], M.A. Province[1]
[1]Washington Univ, MO, USA; [2]Pennington Biomed Res Center, LA, USA; [3]Laval Univ, Canada; [4]Univ Minnesota, MN; [5]Indiana Univ, IN, USA; [6]Texas A & M Univ, TX, USA

Elevated plasma levels of apolipoprotein B (ApoB) and depressed plasma levels of apolipoprotein A1 (Apo-A1) are associated with increasead risk of cardiovascular disease and premature atherosclerosis. In order to identify regions that are likely to harbor QTLs, we conducted an autosomal genome scan for Apo-A1 and ApoB using a variance-components linkage analysis (SEGPATH). Apo-A1 and ApoB were assessed both at baseline and after 20 weeks of endurance-exercise training in 99 White and 101 Black families from the HERITAGE Family Study. Phenotypes were adjusted for age, sex, BMI, fasting insulin, hormone intake and smoking; responses to training were further adjusted for baseline levels. The scan, involving 509 markers, was conducted separately by race. There was strong evidence for linkage on chromosome 13q11 and chromosome 17p11.1 with Apo-A1 in Black families at baseline and in response to training, respectively, with the maximum multipoint lod scores (LOD) of 3.0 ($p = 0.00009$) at marker D13S141 and LOD of 4.6 at marker D17S1294 ($p = 0.000001$). For ApoB at baseline there was evidence of linkage on chromosome 1q42 with LOD of 2.2 ($p = 0.00081$) at marker D1S2860 in Black families; and on chromosome 8q24 with LOD of 2.3 ($p = 0.00059$) at marker D8S1179 in White families.

IGES-16

Data mining of QTL by covariate interactions influencing BMI to identify linked subgroups using a tree linkage approach

I.B. Borecki, M.F. Feitosa, A. Kraja, M.A. Province
Div. Biostatistics, Washington Univ. Schl. Med.St. Louis, MO, USA

Obesity is a risk factor for CHD and is associated with dyslipidemia, hypertension, glucose intolerance, and hyperinsulinemia. A commonly studied measure of adiposity is the body mass index (BMI). There is evidence of both genetic and environmental factors, and studies to find the relevant genes have produced candidates on virtually every human chromosome. While consistency of results is emerging from genome scan studies for some loci, there is still a fair degree of variability in results among studies. It is likely that different gene by environment interactions play an important role. We set out to explore interactions between each of two QTLs influencing BMI that we have identified on chromosomes 7 and 13 via a linkage study, and a series of covariates including demographic measures, anthropometrics, other cardiovascular risk factors, and lifestyle indicators (diet, physical activity, and smoking). We applied a novel tree-based linkage method in which a sample of sib pairs is partitioned according to strength of evidence of linkage in groups defined by dichotomizations of the covariate distributions. Results for both QTLs indicated that the linkage signal is maximized in subgroups defined by middle-aged sib pairs, suggesting that modeling age by QTL interactions could improve the linkage signal and the precision of localization and that these families should be the focus of follow up studies. This tree linkage method is an efficient exploratory tool that is fast and can easily capture even complex models of interaction.

IGES-17

Familial aggregation of asthma-associated quantitative phenotypes in 335 French EGEA families: is there evidence for common genetic determinants?

E. Bouzigon[1], A.S. Carpentier[1], M.H. Dizier[2], M.P. Oryszczyn[3], J. Maccario[3], F. Kauffmann[3], F. Demenais[1]
[1]INSERM EMI 00-06, Evry, France; [2]INSERM U535, Kremlin-Bicêtre, France; [3]INSERM U472, Villejuif, France

Asthma is a complex disease, associated with physiological phenotypes including immunoglobulin E (IgE) levels, skin prick tests to allergens (SPTQ), eosinophil counts (EOS) and force expiratory volume in one second (FEV1). We investigated patterns of familial correlations (FC) of these phenotypes in 335 French families ascertained through one asthmatic proband (212 offspring and 123 parents) using regressive models. These traits were adjusted for covariates and position of family members to correct for ascertainment. Estimates of spouse, father-offspring, mother-offspring and sib-sib correlations were consistent with polygenic inheritance for 3 traits: IgE (heritability, $H = 49\%$), SPTQ ($H = 28\%$) and FEV1 ($H = 44\%$). FC were homogeneous between the 2 family sets ascertained through offspring and through parents. Familial aggregation of EOS showed a complex pattern: polygenic inheritance and familial environment in the whole sample and heterogeneity between the

2 family sets. To search for common determinants between these phenotypes, each phenotype was adjusted for each of the other three. This led to a small reduction of FC, by at most 20%. These data are consistent with the presence of substantial genetic determinants of asthma-associated phenotypes which appear largely independent. These results have implications for gene identification in the EGEA study. This abstract is funded by INSERM-MSD & IDS grant.

IGES-18

An empirical Bayes approach to assessing potential bias in segregation studies due to stoppage effects

M. Brimacombe[1], W. Zahorodny[2]
[1]Dept of Preventive Medicine and Community Health, NJMS-UMDNJ, USA; [2]Dept of Pediatrics, NJMS-UMDNJ, USA

The potential bias related effect of stoppage in familial segregation studies is examined here using a two stage empirical Bayes approach. First, the family size distribution underlying the definition of the relevant likelihood function is parameterized and fit using an empirical Bayes approach. This empirical estimate of the family size distribution parameters is then used in the overall likelihood function and maximum likelihood estimates are obtained which reflect the underlying fitted family size distribution. Several assumed parametric family size distributions are examined and applied from this perspective. The empirical stability of the overall maximum likelihood estimates and associated bias in relation to family size distribution is investigated by varying the first stage estimates over their confidence or credible interval and examining the related variation in the estimated bias and maximum likelihood estimates. Maximum potential levels of bias and the effects of various family size distributions are discussed. This work is conducted in collaboration with the NJ Autism Registry

IGES-19

Genetic epidemiology study of complex clinical phenotypes in isolated populations

K.B. Bulayeva[1,2], T.A. Pavlova[1], R.M. Kurbanov[2], O.A. Bulayev[1]
[1]Dept. of Hum. Genet, Vavilov Inst. Gen. Genetics, RAS, Russia; [2]Dept. of Ethnic Population Genetics, Dag. Center RAS, Russia

We will present results of genetic epidemiology study in Daghestan (Northern Caucasus, Russia) highland isolates. Daghestan contains 26 distinct indigenous ethnic groups. The origin of such extraordinarily high ethnic diversity in Daghestan is still unknown. Our results of genetic diversity study showed that Daghestan populations are clearly close to European ethnic populations. The genetic data support the picture of these as ancient highly isolated populations with rate of the endogamy 85–97% and $F = 0.01–0.015$. We found that many of Daghestan populations have very high prevalence of certain complex diseases such as cardiovascular illnesses, cancer, schizophrenia, mental retardation, and progressive muscular dystrophy. Such isolated populations are the basis of a powerful approach to detect genes for complex diseases because it is likely that most disease alleles are derived from limited number of common ancestors. As part of an ongoing study several extended schizophrenia and cardiovascular kindreds have been ascertained in these isolates. The results showed that inbred subjects experienced greatly increased prevalence of the complex diseases. We will present a comparative genome-wide scan multipoint linkage analysis of the extended multiplex Daghestan pedigrees of different Daghestan isolates that allow us to distinguish between true and false linkage signals and to refine the location of liability genes of certain complex diseases.

IGES-20

SNPs random forests and asthma susceptibility

A. Bureau, K. Falls, J. Dupuis, B. Hayward, T. Keith, P. Van Eerdewegh
Dept. of Human Genetics, Genome Therapeutics Corp, Waltham, MA, USA

A random forest is a collection of classification trees constructed using a random subset of predictors on bootstrap samples of observations. A class prediction for each observation is obtained from the votes of the trees constructed on the bootstrap samples from which the observation was excluded. The importance of a variable, such as a genetic polymorphism in a case-control study, is measured by the increase in misclassification when the values of the variable are randomly permuted among all observations. Random forests are able to detect the equal importance of highly correlated variables such as SNPs in linkage disequilibrium since each of those variables will appear in different trees. We applied the random forest classification method to identify SNPs predictive of asthma in ADAM33, a previously identified asthma susceptibility gene on 20p13. We analyzed 37 SNPs typed on 131 asthma cases from unrelated Caucasian families linked to 20p13 and 217 unrelated controls from the same populations. Five ADAM33 SNPs decrease the average margin between the votes for the true and false classes by at least 10%, the SNP V-3 reaching 30%. V-3 was not significantly associated with asthma by itself in a case-control test, but formed with ST + 4 the most associated 2-SNP haplotype ($p = 0.00004$). The random forest correctly classifies 97% of the controls and 64% of the cases. Our results highlight the usefulness of predictive models to identify key SNPs in the complex interactions between the phenotype and genotype in a disease susceptibility gene.

IGES-21

SNP subset selection for genetic association studies

M.C. Byng[1], J.C. Whittaker[2], A.P. Cuthbert[1], C.G. Mathew[1], C.M. Lewis[1]
[1]Division of Medical and Molecular Genetics, Guy's, King's and St. Thomas' School of Medicine, London, UK; [2]Department of Epidemiology and Public Health, Imperial College School of Medicine, London, UK

Association studies are widely used to identify susceptibility genes for complex traits. The high degree of linkage disequilibrium in the human genome implies that a subset of SNPS can be genotyped in a candidate gene study, without compromising the power of the study. We propose several strategies to reduce the number of SNPs to be genotyped. These methods include applying nearest and furthest neighbour clustering algorithms to pairwise linkage disequilibrium measures and estimating the proportion of haplotypes each SNP subset identifies. In addition we show how power calculations, based on the average power to identify a SNP as the disease susceptibility mutation, can be used to choose SNP subsets. Examples of how to calculate the average power are given for haplotype-based and logistic regression based statistical analyses. All these methods provide a ranking method for subsets of a specific size, but do not provide criteria for overall choice of SNP subset size. We therefore develop such criteria by incorporating power calculations into a decision analysis, where the choice of SNP subset size depends on the genotyping costs and the perceived benefits of identifying association. These methods are illustrated using eleven SNPs in the MMP2 gene.

IGES-22

Meta analysis of association of SNPs in HPC2 (ELAC2) and prostate cancer

N.J. Camp[1], S.V. Tavtigian[2]
[1]Genetic Epidemiology, Dept of Medical Informatics, University of Utah, USA; [2]Myriad Genetics, Utah, USA

Recently Tavtigian (2001 NatGen 27:172–80) identified HPC2 on chromosome 17p as a prostate cancer susceptibility gene. Two segregating mutations were found in extended Utah pedigrees, and two missense variants (Ser217Leu and Ala541Thr) were found to associate with prostate cancer across the resource. Following this discovery several groups attempted to confirm the finding with varying success (Rebbeck 2000 AJHG 67:1014–9; Xu 2001 AJHG 68:901–11; Vesprini 2001 AJHG 68:912–7; Suarez 2001 CanRes 61:4982–4; Wang 2001 CanRes 61:6494–9; Rokman 2001 CanRes 61:6038–41). We have performed a meta analysis, including and excluding the original data, to consolidate evidence for association of the variants S217L and A541T in HPC2 and prostate cancer. We constructed a Mantel-Haenszel meta-analysis with 3 categories: Familial prostate cancer vs Low risk controls (FL); All cases vs Low risk controls (AL); All Cases vs All controls (AA). Both variants were analyzed for carriage of the rare allele. Data from Rokman could not be included since genotypes were not given. Excluding Tavtigian data, the association of the S217L and A541T variants was confirmed in the FL analysis ($p = 0.017$ 1.37[1.06, 1.78]; $p = 0.0059$ 2.01[1.22, 3.29]). The A541T variant was also significant in the AL analysis ($p = 0.0023$) and borderline in AA ($p = 0.073$). Including Tavtigian data, A541T was significant in all analyses (FL $p = 0.00024$; AL $p = 0.00011$; AA $p = 0.012$) and S217L was significant in FL and AL ($p = 0.0053$; $p = 0.036$) and borderline in the AA analysis. A multi-locus analysis considering carriage at both vs neither variant was highly significant for both FL and AL ($p = 0.00011$ 2.44[1.55, 3.83]; $p = 0.00012$ 2.05[1.42, 2.96]). The AA group could not be analyzed as the data were not available. In conclusion our meta-analyses indicate that the total available data is supportive for a role of HPC2 (ELAC2) in susceptibility to prostate cancer.

IGES-23

Analyses of an SLE genome scan: prioritizing regions for fine mapping

R.M. Cantor[1], J.Y. Yuan[1], J.M. Grossman[2], B.H. Hahn[2] B.P Tsao[2]
[1]Dept. Human Genetics; [2]Div. Rheumatology, UCLA School of Medicine, USA

SLE is an autoimmune disorder with a substantial and complex genetic component. Fine mapping in our sample of 145 affected sibpairs provided confirmatory evidence for linkage from independent scans and identified an interaction of 1q23 with 16q12. A full genome scan in 65 of these families with 79 affected sibpairs was analyzed by multipoint npl scores using Genehunter 2.0, and means test empiric p-values using SAGE 4.0. Agreement in prioritized regions highlighted by the linkage statistics was observed but not complete. An npl score of 2.74 with a means test empiric p-value of .02 was seen on 19q13, while an npl score of 2.5 and a means test empiric p-value of .0002 was seen on 18q23. Regions on 3p22, 13q33 and 16q12 were less significant and showed agreement (npl scores > 2.0 and means test p-values $= .02$). 18q23 was again highlighted when all 463 sibpairs in these families were analyzed by variance component linkage using SAGE 4.0 ($p < .004$). But, Haseman-Elston analysis by Genehunter provided no evidence supporting 18q23. On 19q13 variance component linkage analysis with SAGE 4.0 ($p < .02$) and Haseman-Elston multipoint analysis with Genehunter ($lod = 1.0$) were in agreement but marginal. All results on 10q24 were in agreement ($npl = 2.26$, means $p < .006$, variance components $p < .006$, and Haseman Elston $lod = 2.0$) but initially less significant than 18q23 and 19q13. Factors including consistency of linkage information from unaffected sibling pairs and agreement among multiple analyses can help prioritize regions for fine mapping studies.

IGES-24

Two single nucleotide polymorphisms in the ACE2 locus are associated with cardiovascular disease

A.A. Chen, G. Barnes, A. Foti, E. Nolin, S. Lewitzky, J. Metivier, J. Meyer, A. Parker, and E. Topol
Millennium Pharmaceuticals, Cambridge, MA, USA

Cardiovascular disease is responsible for 17 million deaths each year. In the United States, it has been the number one cause of death each year for over a century. Environmental influences such as smoking, inactivity, and poor diet are well characterized risk factors, however many genetic elements remain to be elucidated. We have identified an association between cardiovascular endpoints in a Caucasian population ($n = 301/304$ cases/controls) and two single nucleotide polymorphisms (SNPs) in the angiotensin converting enzyme 2 (ACE2) locus. Cardiovascular endpoints studied include percutaneous transliminal coronary angioplasty, coronary artery bypass grafting, myocardial infarction, and catheterization with greater than 70% stenosis. A G/A substitution within intron three of ACE2 is positively associated with cardiovascular endpoints ($p = .02$) in normotensive individuals of both genders. An insertion/deletion SNP within intron nine is associated with cardiovascular events ($p = .03$) in hypertensive females. These two SNPs are in strong linkage disequilibrium ($D' = .93$, $p < .0001$); further investigation will be required to clarify their individual contributions to risk of cardiovascular disease. These results indicate that ACE2 may be an appropriate target for development of novel therapies for cardiovascular disease.

IGES-25

Using semiparametric association test to detect genetic association in case-control design under structured population

H-S. Chen[1], S. Zhang[1], X. Zhu[2], H. Zhao[3]
[1]Department of Mathematical Sciences, Michigan Technological University, USA; [2]Department of Preventive Medicine and Epidemiology, Loyola Medical School, USA; [3]Department of Epidemiology and Public Health, Yale University School of Medicine, USA

Case control studies tend to have a higher false positive rate in the presence of population stratification. Recently, some methods have been proposed using genomic markers to control for population stratification. However, in continuous admixture population, these methods often do not perform well. We propose a semiparametric association test to detect genetic association between a candidate marker and a qualitative trait of interest in case-control design. The performance of the test is compared to other existing methods. The results show that our method gives a correct type I error rate and has higher power compared to other methods, especially in continuous admixture population.

IGES-26

Single nucleotide polymorphism genome scan localization of a novel glomerulocystic kidney disease locus to chromosome 11p15

J.S. Collins[1], W. Pan[2], B. Tanriover[2], R.C.P. Go[2], L.M. Guay-Woodford[2]
[1]Greenwood Genetic Center, USA; [2]University of Alabama at Birmingham, USA

We previously characterized a large, three-generation African-American family with autosomal dominant Glomerulocystic Kidney Disease (GCKD) and excluded linkage with PKD1, PKD2, and the human orthologue of mouse *jcpk* (Sharp et al., 1997 J Am Soc Nephrol 8:77–84). We performed a whole-genome scan using the Affymetrix GeneChip HuSNP™ Mapping Assay on 8 GCKD affected individuals, 2 obligate carriers, and 4 phenotypically unaffected family members. Of the 1494 single nucleotide polymorphism (SNP) markers on this chip, 1163 were assigned to chromosome (chr) localizations using physical mapping information. Our scan was conducted at an average interval of 4 cM with an average largest gap per chr of 22.5 cM. African-American SNP allele frequencies, which are not publicly available, were calculated from 29 unrelated African-American individuals. Pairwise analyses using the FASTLINK program yielded LOD scores of 1.6 and 1.4, respectively, for SNP markers WIAF-3839 and WIAF-3053 ($ELOD = 1.8$) at chr 11p15.4. Further genotyping in these 14 individuals as well as 6 additional phenotypically unaffected family members was performed with six microsatellite markers spaced at 2 cM intervals across the candidate interval. The highest two-point LOD score of 2.7 ($ELOD = 2.7$) was obtained for marker D11S1338 and multipoint analyses yielded a maximum LOD score of 3.0. Haplotype analyses refined the candidate region to an 11.1 cM interval telomeric to marker D11S909.

IGES-27

Hierarchical modeling of candidate gene markers and haplotypes

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

With numerous markers available within a candidate gene, one often evaluates both marker- and haplotype-level associations. Marker-specific estimates aim to elucidate the precise functional variant, whereas haplotype effects may provide more power for the detection of association. For both approaches, however, conventional estimation methods for multiple exposures may lead to unstable and biased estimates due to sparse data. In addition, issues of multiple comparisons arise when performing

several tests of significance across a chromosomal region. As a unifying approach for marker- and haplotype-specific estimates, and to address problems associated with the evaluation of multiple exposures, we propose a hierarchical model incorporating haplotype-level information. Specifically, the variant arrangement on haplotypes provides higher-level information that allows marker-specific estimates to ''borrow'' information from each marker with a variant in the same haplotype. This accomplishes three main objectives: 1) stabilization of unreliable estimates, 2) providing regional support through the second-stage haplotype estimates, and 3) strengthening of marker-specific estimates by bolstering them with the underlying haplotype effects. We present the model, provide a hypothetical example to outline its properties, and apply it to existing data. We demonstrate the ability to obtain both marker- and haplotype-level associations and the potential refinement in estimation by using the underlying haplotype structure to incorporate the correlations between the markers.

IGES-28

The association of loss of heterozygosity (LOH) in select regions of suspected tumor suppressor genes and survival in squamous cell carcinoma of the head and neck

S.W. Coon[1], A.T. Savera[2], R.J. Zarbo[2], M.S. Benninger[3], G.A. Chase[1], B.A. Rybicki[1], D.L. Van Dyke[4]
Depts. of [1]Biostatistics & Research Epidemiology; [2]Surgical Pathology; [3]Otolaryngology; [4]Medical Genetics, Henry Ford Health System, USA

Loss of heterozygosity (LOH) in chromosomal regions of tumor tissue that harbor tumor suppressor genes may lead to decreased survival time of cancer patients with squamous cell carcinoma of the head and neck (HNSCC). We studied eight regions often lost in HNSCC in 152 patients with a primary diagnosis of HNSCC in the Henry Ford Health System, followed for up to 75.1 months (median = 37.8 months). Tumor and normal tissue DNA specimens were typed for microsatellite repeat markers in each of the eight regions. The association between LOH and death from HNSCC was investigated, weighted by number of informative markers in a region, and adjusted for age, race, tumor stage, and current smoking status. LOH was significantly associated with reduced survival at three of the examined chromosomal regions. A greater risk for cancer mortality was observed for LOH at 3p13, Hazard Ratio (HR) = 1.75 (95% confidence interval (CI): 1.05–2.89, P = 0.03), at 8p22-p23, HR = 1.92 (95% CI: 1.12–3.29, P = 0.02) and at 9p21-p24, HR = 2.28 (95% CI: 1.38–3.76, P = 0.001). Patients with LOH at 10p13-pter had decreased cancer mortality, HR = 0.51 (95% CI: 0.25–1.03) but was not statistically significant at the 5% alpha level. Our results indicate that LOH at several chromosomal sites may offer additional prognostic information to the more traditional indicators such as tumor stage and age.

IGES-29

Case/pseudo-control analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions and parent-of-origin effects.

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

Estimation and testing of genetic effects such as genotype relative risks is often performed conditionally on parental genotypes using data from case-parent trios. This strategy avoids estimating nuisance parameters such as parental mating frequencies and confounding causes of association such as population stratification. For effects at a single locus, the resulting analysis is equivalent to conditional logistic regression using the case and three 'pseudo-controls' derived from the untransmitted parental alleles. We have previously shown that a similar approach can be used for analysing genotype and haplotype effects at a set of closely linked loci, but with a required adjustment to the conditioning argument that results in varying numbers of pseudo-controls depending on the disease model that is to be fitted. Here we extend this method to the analysis of epistatic effects (gene-gene interactions) at unlinked loci and allow additional incorporation of gene-environment interactions. The conditional logistic approach provides a natural and flexible framework for incorporating these additional effects. By relaxing the conditioning on parental genotypes to allow exchangeability of parental genotypes, we show how similar approaches can be used to study parent-of-origin effects at one or more loci. This approach provides limited power to distinguish between parent-of-origin effects and interaction between genotypes of mother and child.

IGES-30

Alzheimer pathogenesis for men and women: the 3-year acceleration in tangle stage found for women and jump in amyloid stage in late middle age are mediated independently by APOE4

E.H. Corder[1], E. Ghebremedhin[2], M. Taylor[1], D.R. Thal[4], T.G. Ohm[5], H. Braak[2]
[1]Center for Demographic Studies, Duke Univ., USA; [2]Dept. of Clin. Neuroanatomy, J.W. Goethe-University, Germany; [3]Dept. of Sociology, Duke Univ.; [4]Inst. for Neuropathology, University of Bonn Medical Center, Germany; [5]Dept. of Anatomy, Med. Faculty Charite of the Humboldt Univ. Germany

Epidemiologic studies place women at a higher risk of Alzheimer's disease. We describe age-related gender differences in Alzheimer pathogenesis and the role of apolipoprotein E (APOE) polymorphism. Brains from 3165 men and 2450 women were investigated; age range, 20 to 105 years. Neuropathologically, six neurofibrillary tangle (NFT) stages and three senile plaque (SP) stages were differentiated. Linear regression models were used to predict stages based on age, gender, and APOE polymorphism. Women had appreciably higher mean NFT stage from age 45 onward, age 55 for SP stage. There was a 3-year acceleration of NFT pathology for women at ages 65, 75, and 85 in relation to men. The gender gap was largest at age 65 when women had +13% higher NFT stage ($p = 0.01$) and +33% higher SP stage ($p = 0.001$). Women had a more rapid spread of tangles from the initial transentorhinal location to adjacent areas – fuelled by APOE e4 and associated with increased SP stage ($p = 0.006$) especially for e4+ women at age 65 ($p = 0.0002$). At oldest ages women with limited tangle pathology frequently had SP in primary cortical fields. The 3-year acceleration in allocortical tangle pathology found for women and large increases in plaque distribution at ages 65 and 75 were most evident for women who carry the e4 allele for APOE. These pathologic differences, not more frequent isocortical tangles, may usually account for the observed higher rates of Alzheimer-like dementia among women.

IGES-31

Racial differences in elevated cancer risk among relatives of early-onset lung cancer cases

M.L. Cote, A.J. Spindler, A.G. Schwartz
Wayne State University School of Medicine and Karmanos Cancer Institute, Population Studies and Prevention Program, USA

African Americans are at higher risk for lung carcinoma than Caucasians in the United States, especially among those with early onset of disease. This ongoing study evaluates whether relatives of lung cancer cases are at increased risk of lung and other cancers. Families were identified through 547 population-based lung cancer cases diagnosed prior to 50 years of age and 637 population-based healthy metropolitan Detroit residents selected through random digit dialing. Data were collected for 2852 relatives of cases and 2998 relatives of controls. Data regarding spouses and offspring were excluded from this analysis. After adjusting for each relative's age, race, sex and smoking status, first-degree relatives of cases were at 2.7-fold (95% CI, 1.8–4.0) increased risk for lung cancer, with African American relatives at greater risk ($RR = 4.1$, 95% CI, 2.0–8.4) than Caucasian relatives ($RR = 2.1$, 95% CI, 1.3–3.4). Similar disparities were seen for tobacco related cancers, with African American relatives at more than twice the risk compared to Caucasian relatives ($RR = 4.8$, 95% CI 2.4–9.7 and $RR = 1.7$, 95% CI, 1.1–2.6, respectively). Relative risk estimates were also elevated, but not significantly, for breast cancer in African American female relatives ($RR = 2.0$, 95% CI 0.9–4.3); risk was not elevated in family members of Caucasian cases. These findings of racial differences in familial aggregation for cancers of the lung, tobacco-related sites and breast may suggest that the role of common susceptibility genes for cancer differs by race. Several susceptibility genes are currently under investigation in this population.

IGES-32

Association studies in candidate genes: strategies to select SNPs to be tested

E. Cousin[1], E. Genin[2], S. Ricard[1], S. Mace[1], C. Chansac[1], J.F. Deleuze[1]
[1]Aventis Pharma, Evry Genetics Center, France; [2]INSERM U535, Le Kremlin-Bicêtre, France

When numerous Single Nucleotide Polymorphisms (SNPs) have been identified in a gene candidate to be a disease risk factor, a relevant and still unanswered question is to determine how many and which of these SNPs are needed to be tested for association with the disease. Testing them all is expensive and often unnecessary. Indeed, alleles at different SNPs may be associated in the population because of the existence of linkage disequilibrium so that knowing the alleles carried at one SNP could provide exact or partial knowledge of alleles carried at a second SNP. Different strategies have been developed to determine the most appropriate SNP subset to be genotyped in a candidate gene. We compare here two methods, a method that identifies the SNPs defining the most frequent haplotypes present in the population (haplotype tag SNPs) (Johnson et al., Nature Genetics 2001, 29: 233–237) and a method based on power computations that we have developed (Genin, Genet Epidemiol. 2001, S1: 614–619). We show that the method based on power computations is simpler and especially more efficient. It avoids the exclusion of frequent SNPs that may be present in only rare haplotypes and it also estimates the loss of power according to the subset of SNPs selected and the number of tests performed.

IGES-33

Within-family analyses of polymorphism effects adjusted for familial factors, with application to CYP17 and breast cancer

J. Cui, A.B. Spurdle, G.S. Dite, M.C. Southey, D.J. Venter, M.R.E. McCredie, G.G. Giles, G. Chenevix-Trench, J.L. Hopper
Centre for Genetic Epidemiology, University of Melbourne, Australia

We investigated the effect of a CYP17 MspA1 polymorphism on breast cancer risk using three different within-family analysis methods: the Class A regressive logistic model, Age of Onset Regressive Logistic Model, and Proportional Hazards Familial Model. All methods were fitted using the software MENDEL. We applied these methods to 1575

three-generation population-based families ascertained through incident cases of breast cancer identified by cancer registries for the Australian Breast Cancer Family Study (ABCFS), in part funded by the NIH Cooperative Family Registry for Breast Cancer Studies (CFRBCS). Half of the case probands were diagnosed before the age of 40 years and the remainder between the age of 40 and 59 years. Four-two case families in which a deleterious mutation in BRCA1 or BRCA2 has to date been identified were excluded from these analyses (see Cui et al., 2001, Am J Hum Genet). The estimates for relative risks associated with the CYP17 genotypes were adjusted for mother's disease status, and if affected her age at onset, and the proband's age at onset. All methods were consistent in showing that a recessive model gave the best fit to the data. Under this model, the relative risk for homozygotes was 1.47 [95% confidence interval (1.27, 1.70)]. This suggests that CYP17 polymorphism is associated with an increased risk of breast cancer in families in which at least one woman has been diagnosed with breast cancer at a relatively young age.

IGES-34

Linkage to candidate genes in the Cleveland Colon Neoplasia Sibling Study: the N-acetyltransferase locus and epistatic effects

D. Daley, G.L. Wiesner, S. Iyengar, S.D. Markowtiz, R.C. Elston
Case Western Reserve University, Cleveland, OH, USA

The N-acetyltransferase locus comprises two closely linked genes, NAT1 and NAT2. These genes code for acetylation enzymes, important in the metabolism of drugs and other possible carcinogens. Recent evidence suggests that the NATs may play an important role in the development of colorectal cancer (CRC) in carriers of the mismatch repair genes (HNPCC). We have identified joint linkage to the NAT region and a second colon cancer candidate gene (locus 2) in a subset of 60 families in the Cleveland Colon Neoplasia Sibling Study. In order to be classified as affected a diagnosis before the age of 65 of colon cancer, high grade dysplasia (HGD) or adenomatous polyps is required; if diagnosed with HGD or adenomatous polyps the lesion must be $\geq$1cm or diagnosed before the age of 51. Unaffected individuals must have been screened negative with sigmoidoscopy or colonosopy at an age $\geq$ the oldest affected sibling. All other individuals are classified as unknown. Genotyping was performed for NAT1, NAT2, and D8S261, a closely linked microsatellite (MS) 2cM away. We have recently published our linkage results to this region (Daley et al., 2002). At locus 2, genotyping was completed for 1 highly polymorphic MS located $\sim$.3cM away and a common SNP located in its coding region. Evidence for linkage to the NAT region and locus 2 is provided by the Haseman-Elston regression method, with evidence suggesting a possible dominant by dominant epistatic component of variance. Sequencing of locus 2 is currently underway and full results will be presented.

IGES-35

Multivariate linkage analysis using phenotypes related to the insulin resistance-metabolic disorder

M. De Andrade[1], C. Olswold[1], S.L.R. Kardia[2], E. Boerwinkle[3], S.T. Turner[4]
[1]Health Sciences Research and [4]Hypertension, Mayo Clinic, USA; [2]Human Genetics, Univ. of Michigan, USA; [3]Human Genetics Center, UT Health Science Center, USA

Genome-wide linkage analyses for complex disorders are challenging. Although these disorders aggregate in families, they usually do not segregate in a Mendelian fashion, since they are influenced by multiple genes, each of which is polymorphic and has small-to-moderate effects. In this study we performed multivariate linkage analyses for phenotypes associated with the insulin resistance-metabolic syndrome, using the Rochester Family Heart Study data consisting of 279 extended pedigrees (2494 individuals). This metabolic syndrome is identified by abnormalities in three or more of the following phenotypes: abdominal girth (measured by waist circumference), triglycerides, HDL cholesterol, systolic blood pressure, and fasting glucose. Multivariate linkage analyses were performed on all possible combinations of fasting insulin and these 5 traits taken 3 at-a-time. Significant LOD scores were observed on 5q (LOD = 5.10) and 6q (LOD = 3.68) for the combination of triglycerides, fasting insulin and fasting glucose. In contrast, the univariate LOD results for these 3 traits on 5q were 1.28, 0.71, and 0.28, and for 6q were 2.24, 0.54, and 0.00, respectively. Thus, using multiple traits that are correlated with the insulin resistance-metabolic syndrome provided an enhanced ability to localize responsible genetic factors, which in turn will lead to better understanding the genetic basis of this complex disorder.

IGES-36

A polymorphism of intron 1 interferon-gamma gene (IFNG) influences immunoglobulin E (IgE) levels in 107 French EGEA families

F. Demenais[1], M. Boussaha[1], M.H. Dizier[2]
[1]INSERM EMI 00-06, Evry, France; [2]INSERM U535, Kremlin-Bicêtre, France

Genome-wide scans of asthma and asthma-related phenotypes have consistently reported linkage to 12q. This region contains numerous candidate genes influencing allergic inflammation, including the IFNG gene. Fine mapping of 12q13-q23 region in Barbados families indicated the best evidence for linkage of asthma to a CA repeat polymorphism in the first intron of IFNG, although no association

was detected (J Allergy Clin Immunol, 1999, 104:485–91). We have investigated the role of IFNG intron 1 polymorphism in asthma and asthma-related phenotypes in 107 French EGEA families with at least 2 asthmatic siblings, using a family-based association approach. Genotyping of IFNG intron 1 polymorphism in 212 parents and 279 offspring of the EGEA families revealed 4 alleles with frequencies ranging from 2.5% to 53%. To search for linkage and association of this variant with asthma related-phenotypes, we used the Transmission Disequilibrium Test for binary traits (asthma, Skin Prick Test positivity, Multi-RAST Phadiatop) and combined segregation-linkage analysis based on regressive models including linkage disequilibrium for continuous traits (IgE levels, eosinophil counts). No significant result was obtained for binary phenotypes or eosinophil counts. Segregation-linkage analysis of IgE showed that carriers of rare alleles (14 or 15 CA repeats) of IFNG intron 1 have significantly lower IgE levels than non-carriers ($p < 0.001$). This polymorphism accounts for 4% of IgE variation.

This abstract is funded by INSERM-MSD and IDS grants.

IGES-37

Genetic variation at the scavenger receptor class B type I (SR-BI) gene locus interacts with diabetes in determining plasma lipids: The Framingham Offspring Study

S. Demissie[1], L.A. Cupples[1], D. Corella[2], D. Osgood[2], J.M. Ordovas[2]
[1]Dept. of Biostatistics, Boston Univ., USA; [2]Lipid Metabolism Laboratory, Jean Mayer-USDA Human Nutrition Research Center on Aging, Tufts Univ., USA

The scavenger receptor class B type I (SR-BI) is a key component in the reverse cholesterol transport pathway. In previous studies we have found three common polymorphisms (in exon 1, intron 5 and exon 8) associated with plasma lipids and body mass index (BMI). We hypothesized that diabetic status may interact with these polymorphisms in determining lipid concentrations. We evaluated this hypothesis in 2289 non-diabetic (47% men) and 176 diabetic (64% men) participants in the Framingham Offspring Study. SR-BI genotype, apolipoprotein E (apoE) genotype, anthropometric, clinical, biochemical, and life style variables were determined. Statistical analyses were performed using SAS Proc Genmod (the generalized estimating equation) to adjust for familial correlations. After adjustment for age, sex, smoking, alcohol intake, apoE genotype, and BMI, we found statistically significant interaction between SR-BI exon 1 genotypes and diabetes indicating that diabetic subjects with the less common allele (allele A) have significantly lower lipid levels. For low-density-lipoprotein cholesterol (LDL-C), the adjusted means ($\pm$standard error) were 127.9 ± 1.0 and 127.0 ± 1.5 for G/G and G/A-or-A/A non-diabetics, respectively, compared to 123.0 ± 3.9 and 106.0 ± 5.7 for G/G and G/A-or-A/A diabetics (P for interaction $= 0.03$). Similar, but weaker, results were obtained for total cholesterol and high-density-lipoprotein cholesterol (HDL-C). Interaction effects were not significant for intron 5 and exon 8 polymorphisms. In conclusion, diabetes status may modulate the effect of the SR-BI gene variation on the lipid profile and affect the cardiovascular risk.

IGES-38

Meta-analysis of linkage studies for complex diseases: a simulation study

A. Dempfle[1], S. Loesgen[2]
[1]Institute of Medical Biometry and Epidemiology, University of Marburg, Germany; [2]LoesGen, Oberbözberg, Switzerland

A common strategy to identify susceptibility genes for diseases are linkage studies. In complex diseases with several genetic and environmental factors, the power of a genome scan to detect genes with only moderate effect is often poor with the usual sample sizes of a few hundred affected sib pairs. Meta-analysis of several studies is a promising approach. Appropriate data scenarios are getting more realistic. Samples on common disorders (like asthma, diabetes, ...) are collected by several groups, often under standardized or similar study protocols. By now the initial analyses of many genome scans are published and data could be made available for combined analysis. There are many differences between studies that must be considered for meta-analysis, like different study populations (with possible genetic heterogeneity), the definition of relevant phenotypes, genetic markers and sample sizes. We investigate the effect of different genetic markers and the resulting loss in information content. One possibility is to calculate multipoint linkage scores for each study separately and properly combine them. This can be done by rank statistics or weighted averages, with weights incorporating systematic differences between studies. We propose a new meta-analysis method that explicitly allows for variable information content and compare this to existing methods (Genome Search Meta-Analysis (a), Fisher's p-value combination and Truncated Product Method (b)) in a simulation study. (a) Ann. Hum. Genet 63 (3) 1999:263–272. (b) Genet Epidemiol 22 (2) 2002:170–185.

IGES-39

Exploring the phenotype among alcoholic families associated with *GABRA2*

D.M. Dick[1], H.J. Edenberg[1], B. Porjesz[2], H. Begleiter[2], T. Foroud[1] and the COGA collaborators
[1]Indiana University School of Medicine, Indianapolis, IN, USA; [2]State University of New York, Brooklyn, NY, USA

Linkage analyses in the Collaborative Study of the Genetics of Alcoholism (COGA) sample

(Reich et al., 1998) and in a sample of Southwestern American Indians (Long et al., 1998) have previously identified a region on chromosome 4 near the $GABA_A$ gene receptor cluster that is linked to alcoholism. EEG activity has also been linked to the $GABA_A$ receptor region on chromosome 4 in COGA (Porjesz et al., 2002). Subsequent family-based association analyses found significant evidence of association between alcohol dependence and SNPs in *GABRA2*, which encodes the alpha2 subunit of the $GABA_A$ receptor. Significant association was observed with nine SNPs across *GABRA2* that are in high linkage disequilibrium with each other. We have determined that the evidence of association comes from approximately 80 families (31% of the sample). We are currently conducting analyses to better delineate the alcoholism phenotype among these families and to determine how this phenotype differs from that of families who do not show association with *GABRA2*. COGA has collected a variety of measures among alcoholic individuals, including a polydiagnostic interview allowing for the evaluation of comorbid disorders, personality questionnaires assessing characteristics such as novelty-seeking, harm avoidance, and reward dependence, and electrophysiological data-such as EEG and ERP. These studies should clarify both the phenotype(s) of alcoholism and the linkage and association data.

IGES-40

Linkage and genetic heterogeneity indicated for asthma and atopy on chromosomes 8p and 12q in 107 French EGEA families

M.H. Dizier[1], H. Quesneville[2], C. Besse-Schmittler[3], M. Guilloud-Bataille[1], H. Selinger-Leneman[1], F. Clerget-Darpoux[1], F. Demenais[3]
[1]INSERM U535, Kremlin-Bicêtre; [2]Inst. J Monod, Paris; [3]INSERM-EMI 00-06, Evry. France

Using the 107 French families with at least 2 asthmatic siblings, as a part of the EGEA study, we have investigated linkage to asthma (or atopy) and genetic heterogeneity according to the presence/absence of atopy (or asthma) using: 1) the Triangle Test Statistic (TTS), which considers the IBD (identical by descent) distribution among affected sib-pairs discordant for an associated phenotype and 2) the predivided sample test (PST) which compares the IBD distribution of marker alleles between affected sib-pairs concordant and discordant for the associated trait. Two regions, 8p and 12q, already reported to be linked to both asthma and atopy, were examined here. Twenty asthmatic sib-pairs discordant for atopy and 24 atopic pairs discordant for asthma were analyzed by both TTS and PST methods and 97 pairs with atopic asthma by PST. Some evidence of linkage was observed for two markers in the 8p region, D8S504 and D8S503 for asthma and atopy respectively with genetic heterogeneity according to the presence/absence of the associated trait. For 12q chromosome, there were also similar evidence of linkage with genetic heterogeneity to two markers, D12S83 and D12S95, for atopy and asthma respectively. Provided the small distance between the two markers on either 8p (16cM) or 12q (21 cM), it is unclear whether one or two genetic factors are involved in either region.

IGES-41

Testing linkage and gene × environment interaction: comparison of different affected sib-pair methods

M.H. Dizier, H. Selinger-Leneman, E. Genin
INSERM U535, Le Kremlin Bicêtre. France

The aim of this study was to compare under different models of gene-environment ($G \times E$) interaction, the power to detect linkage and $G \times E$ interaction for different tests using affected sib-pairs. Methods considered here were 1) the 'maximum likelihood lod-score' (MLS) based on the distribution of parental alleles identical by descent (IBD) in affected sibs; 2) the sMLS, the sum of the MLS calculated in affected sib-pairs, with 2, 1 or 0 sib(s) exposed respectively; 3) the Predidived Sample Test (PST) which compares the IBD distribution between affected sib-pairs with 2, 1 or 0 sib(s) exposed; and 4) the TTS which uses the IBD distribution among affected sib-pairs with one sib exposed. Both the MLS and the sMLS allow to detect linkage, however only the sMLS accounts for a possible $G \times E$ interaction but without testing it. In contrast, the PST and the TTS allows to detect both linkage and $G \times E$ interaction. Results showed that when exposure to E cancels the effect of G or change the direction of this effect (i.e. the protective allele becomes the risk allele), the PST and sMLS may provide greater power to detect linkage than did the MLS. Under models where E changes the direction of the effect of G, the TTS test may also be more powerful than the PST and the sMLS. Under the other models, the MLS remains the most powerful test to detect linkage. However, only the PST and TTS allow the detection of $G \times E$ interaction.

IGES-42

Analytic results for a model of genotypic assortive mating

Y. Dong and R. C. Elston
Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

Lange [Math. Biosci. 29: 49–57. 1976] modeled genotypic assortive mating in a population as a mixture of random mating and complete genotypic assortive mating. He showed that at a single locus there exists a stable equilibrium if the fraction of complete genotypic assortive mating is small. Here

we start by showing that the fraction of complete genotypic assortive mating at a diallelic locus is equal to the correlation coefficient between the genotypes of the two mates when these are quantified as the numbers of alleles of each kind. We then set up this genotypic assortive mating pattern as a model for both founders and non-founders in a typical pedigree under study and calculate the joint probability of the genotypes of the two mates. In this way we are able to calculate the conditional probability of a specific genotype of a mate, given the type of his or her mate and any anterior pedigree information. Hence we derive an algorithm to calculate the likelihood function for a pedigree under this assortive mating model using the nuclear family-connector approach for traversing a pedigree. This algorithm can handle pedigrees with both mating clusters and mating chains. We restrict our discussion to pedigrees with no loops and to the two-allele case only for the sake of simplicity.

IGES-43

Estimation of genotype relative risks from TDT data with an application to asthma and the ADAM33 gene

J. Dupuis, B. Hayward, A. Bureau, T. P. Keith, P. Van Eerdewegh
Dept. of Human Genetics, Genome Therapeutics Corp, Waltham, MA, USA

Risch and Merikangas (Science (1996) 273:1516–7) proposed to perform genome wide association studies using a trio design, where affected individuals and their parents are collected. This type of design is usually analyzed using the Transmission Disequilibrium Test (TDT). Once a gene or mutation is found to be associated with the disease, it is often of interest to estimate the gene effect, typically in terms of genotype relative risks (GRRs). Ideally, the GRRs would be estimated from an unascertained sample. For a rare disease and a case/control design, the odds ratio provides a reasonable estimate of GRR. However, for a complex trait, when the disease may not be rare, the odds ratio does not provide a good approximation. We present a partial likelihood method to estimate the GRRs from genotype transmissions to affected individuals. The method does not require genotypes to be in HWE nor allele frequency estimates, makes minimal assumptions on the mode of inheritance for the disease, and provides variance estimates. We applied this method to ADAM33, a recently identified asthma susceptibility gene. For ADAM33 SNPs and SNP haplotypes that were most significantly associated with asthma, the GRR to heterozygote carrier of the at-risk allele/haplotype ranged from 1.02 to 1.44, while the GRR to carriers of two copies were in the range of 1.24 to 1.87. The estimates were consistent with a multiplicative model for the GRRs, where the risk to an individual carrying two copies of the at-risk allele is the square of the risk to a heterozygote carrier.

IGES-44

Familial aggregation of several common complex diseases in a Dutch recent genetically isolated population

R. El Galta, J.J. Houwing-Duistermaat, M. Dekker, C.M. Van Duijn, T. Stijnen
Dept of Epi & Biostat, Erasmus MC Rotterdam, The Netherlands

Relatedness of 48 patients with Parkinson's Disease, of 161 patients with Alzheimer's Disease, of 117 patients with type 2 diabetes, and of 43 patients with type 1 diabetes were compared with relatedness of controls. Patients and controls were ascertained from an isolated population of 20,000 inhabitants in the Netherlands. This population was founded by 150 individuals around 1700. A genealogical database of a significant part of this population is available. For each disease, a control set was constructed by random selection of individuals matched with respect to age and sex. As measure of relationship between two individuals we used the kinship coefficient (KC). For each disease, pairewise KC were calculated for all pairs of patients and similarly for all pairs of controls. To test the null hypothesis of no familial aggregation, the difference between the averages of the log of KC of cases and controls divided by the square root of the sum of the estimated variances was used as statistic. The variances were calculated taking into account the correlation between different pairewise KC's. The geometric average KC for Parkinson's patients was 0.58e-4 (0.2e-4, 0.16e-3) and for the corresponding control group it was lower 6.77e-6 (1.23e-7, 0.37e-3). However the difference was not significant ($p = 0.15$). For the other studied outcomes results will be presented. Based on this suggested statistic not much evidence exists that genetic factors contribute to Parkinson's Disease in this population.

IGES-45

Calpain-10 gene, physical inactivity and type 2 diabetes mellitus: a possible gene-environment interaction

C.D. Engelman[1], M.M. Barmada[2], R.E. Ferrell[2], J.M. Norris[1]
[1]Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, CO, USA; [2]Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, USA

Epidemiological studies have found physical inactivity to be an environmental determinant in type 2 diabetes (T2DM). In addition, the calpain-10 (CAPN10) gene has been implicated in T2DM by some linkage and association studies. Despite the importance of both environmental and genetic factors, few studies have examined these factors together. We applied two analytic methods, the transmission disequilibrium test (TDT) and the family based association test (FBAT), to 439 Hispanics with T2DM from 123 pedigrees and 201

non-Hispanic whites with T2DM from 149 pedigrees. Lifetime leisure and occupational physical activity prior to diabetes diagnosis was summarized as average metabolic equivalents task units (METs) per week. We performed an analysis overall and stratified by the upper and lower tertiles of physical activity (i.e. METS) of the affected individual. Overall, we found no evidence for LD between the individual polymorphisms (SNP-44, -43, -19, and -63) and T2DM, nor between the haplotypes defined by these SNPs and T2DM. Stratification by the upper and lower tertiles of physical activity revealed evidence for LD between the 1121 haplotype and T2DM in the least active tertile ($p = 0.03$ and 0.05 in the Hispanics and non-Hispanic whites, respectively), but not in the most active tertile. This result suggests a gene-environment interaction between CAPN10 and physical inactivity, where the association between CAPN10 and T2DM is seen only in those with a sedentary lifestyle.

IGES-46

Effect of box-cox transformation on power of variance components and Haseman-Elston tests to detect quantitative trait loci

C.J. Etzel[1], S. Shete[1], T.M. Beasley[2], J.R. Fernandez[2,3,4], D.B. Allison[2,3,4], C.I. Amos[1]
[1]Department of Epidemiology, University of Texas, M.D. Anderson Cancer Center, Houston, TX, USA; [2]Department of Biostatistics, Section on Statistical Genetics, The University of Alabama at Birmingham, Birmingham, AL, USA; [3]Department of Nutrition Sciences, Division of Physiology and Metabolism, The University of Alabama at Birmingham, Birmingham, AL, USA; [4]Clinical Nutrition Research Center, The University of Alabama at Birmingham, Birmingham, AL, USA

Non-normality of the phenotypic distribution can affect power to detect quantitative trait loci in sib pair studies. In our previous investigations (Fernandez et al.; Hum Hered 2002;53:59–67), we observed that Winsorization of the sib pair phenotypes increased power of QTL detection with Haseman-Elston least squares tests and variance component analysis. Phenotype Winsorization led to a slight increase in type 1 error in Haseman-Elston tests and a slight decrease in variance components analysis. In this investigation, we considered transforming the sib pair phenotypes using the Box-Cox family of transformations. Data were simulated for normal and non-normal (skewed and kurtotic) distributions. Phenotypic values were replaced by Box-Cox transformed values. Ten thousand replications were performed for the three Haseman-Elston tests of linkage and the likelihood ratio test (LRT), the Wald test and other robust versions based on the variance components method. We observed that transformation of the phenotypes by the Box-Cox method resulted in lower empirical type 1 error rates among LRT and Wald linkage tests, but transformations had little effect on the H-E linkage tests as compared to Winsorized and non-transformed phenotypes. For example, at the 0.01 alpha level, empirical Type 1 error rates were 0.0074, 0.0469 and 0.0827 for the Haseman-Elston test, LRT and Wald test, respectively, for non-transformed, skewed phenotypes. The empirical Type 1 errors were 0.0082, 0.0299 and 0.0537 after Winsorization and 0.0113, 0.0153 and 0.0304 after the Box-Cox transformation. Further, power (adjusted for empirical type 1 error) at the 0.01 alpha level ranged from 5 to 17% across all tests using the non-transformed, skewed phenotypes, 8 to 22% after Winsorization of the phenotypes and 10 to 30% after Box-Cox transformation.

IGES-47

Rapid ordering of multiple loci leading to framework maps

C.T. Falk
The New York Blood Center, New York, NY, USA

Mapping disease susceptibility genes using multipoint methods, requires a framework map of correctly ordered loci. With the large number of closely linked loci presently available, generating the correct order for a large set of loci becomes difficult. Using Thompson's [1987, J Math Appl Med Biol 4:93-108] concept of minimizing the obligatory crossover count in phase-known families, we present a "minimum break" algorithm that provides a method for reliable, rapid ordering of a large number of (unordered) loci in 3-generation families. Ordering is based on minimizing the total number of crossovers ("breaks") in a set of haplotypes, where (grandparental) allele origin can be assigned to each haplotype, but locus order is unknown. Haplotypes are determined with the "chrompic" option of CRIMAP [Green et al., 1990, CRIMAP v 2.4]. A simulated annealing algorithm described by Falk and Falk [1995, Phys Rev E 52:895–900] is then used to generate the best order(s). This is followed by a "binning" step, where all loci not separated by recombination are grouped within a single bin, resulting in a framework locus order. CRIMAP is then used to finalize the map. The method is illustrated with chromosome 7 data from the Marshfield Web Site (Broman and Weber 1999, Genet Epidemiol 16:337–343). First, the 25 markers used by Broman and Weber are used to create a framework map, resulting in a map equivalent to theirs. To illustrate the power of the method to rapidly generate a framework order for a larger number of loci, all chromosome 7 loci from Marshfield with more than 150 informative meioses were ordered. A framework map of 61 loci was produced, using less than 2 minutes of computer time. The estimated map length was in good agreement (in both length and order) with the Marshfield map for the same segment of chromosome 7. The major advantage of the method is its ability to rapidly order large sets of loci from which a reliable framework map can be generated.

Supported by the National Institutes of Health (GM29177).

IGES-48

Association studies of QTL for multi-allele markers by mixed models

R.Z. Fan, J.S. Jung
Texas A&M University, USA

In this report mixed models are used for association study between a quantitative trait locus (QTL) and a multi-allele marker. One may collapse a multiple allele marker to be a bi-allelic marker in association study. In certain circumstances, this may not be a good method since much information may be lost by collapsing different alleles to be new alleles. Moreover, different ways to collapse a multiple allele marker can lead to different results which makes the interpretation of results a complicated task. In this paper, mixed models are proposed for two types of nuclear family data. The first type of data is the genetic data of offspring of heterozygous parents, and the second type of data is the genetic data of offspring of both homozygous and heterozygous parents. For the data of offspring of heterozygous parents, we show that it can be used in association study in the presence of linkage between a marker and a QTL. The theoretical basis is the difference between the conditional mean of trait value given an allele is transmitted and the conditional mean of trait value given an allele is not transmitted from a heterozygous parent. For the data of offspring of both homozygous and heterozygous parents we show that it can be used in general association study. In this case, the theoretical basis is the difference between the conditional mean of trait values given an allele is transmitted from a parent and the population mean. To build valid models, we calculate the variance covariance structure of trait values for offspring of a nuclear family. Besides, the reduction of the number of parameters is discussed under an assumption of tight linkage between the trait locus and the marker for the data of offspring of heterozygous parents. Based on mixed models, two test statistics are proposed for association study. By power calculation and comparison, we show that the proposed test statistics have higher power than that by collapsing alleles to be new ones. The proposed models are used to analyze chromosomes 4 and chromosome 16 data of the Oxford asthma data, Genetic Analysis Workshop 12.

IGES-49

Failure to narrow location of May-Hegglin anomaly (MHA) on chromosome 22 using single nucleotide polymorphism markers (SNPs)

Y.T. Fan[1], R. Gundry[2], D. Hayden[1], P. Boyce[1], E. Luong[3], A. Voltz[3], M. Kelley[4], J.E. Bailey-Wilson[3], K.F. Doheny[1], E.W. Pugh[1]

[1]Center for Inherited Disease Research, IGM, JHUSOM, Baltimore, MD, USA; [2]Dept of Pharmacology, JHUSOM, Baltimore, MD, USA; [3]IDRB, NHGRI, NIH, Bethesda, MD, USA; [4]Duke University, Durham, NC, USA

Previously we replicated linkage obtained using a 10 cM microsatellite genome scan for MHA to chromosome 22 in one large and two small families using 1,494 SNPs in the GeneChip HuSNP Mapping Assay and Microarray Suite 4.0 Software (Affymetrix). We attempted to localize the gene for MHA using SNPs by genotyping 61 additional SNPs on chromosome 22 using the Homogeneous MassEXTEND (hME) assay (Sequenom). We genotyped the large four-generation family ($n = 23$), and 24 diversity panel samples from the Coriell Institute for Medical Research. Each assay was done at least twice per DNA sample, there were 3 mismatches among 1,451 comparisons (0.21%). 21 (34%) of the 61 SNP assays failed. The missing rate for the 40 successful SNPs was 3.1% but only 25 of them were informative for linkage analysis in this family. These 25 SNPs are located near the MHA region (27–89 cM) with an average distance between SNPs of 1.82cM. The genetic locations of the SNPs are approximate based on physical location. Two point and multipoint linkage analyses were performed using FASTLINK and GENEHUNTER. Many informative markers had similar two point and multipoint lod scores. Five SNP markers had two point lod scores greater than 2.0. The highest multipoint lod score was 3.87 located at 73.63cM. Three regions had multipoint lod scores greater than 3.5 (54.43–74.71cM, 84.13–85.95cM, 95.41–100.8cM). The MHA gene is known to be located at approximately 62.32cM. Even in this simple case of a Mendelian dominant disease with complete penetrance, we were unable to narrow down the candidate region. In contrasts, by using microsatellite markers, the original investigators were able to identify a candidate gene region of 1MB ($\sim$6cM).

IGES-50

An examination of the properties of cumulative meta-analysis when applied to genetic association studies

B.A. Fijal
Department of Human Genetics, Roche Molecular Systems, Inc., Alameda, CA, USA

Assessing the association between potentially causal polymorphisms and complex disease outcomes has been hampered by the large sample sizes needed to achieve adequate power. An alternative to running a single, large study is to combine results from several smaller studies using cumulative meta-analysis. In cumulative meta-analysis, meta-analyses are repeatedly performed as new data become available until significance is reached or until a predetermined number of studies have been examined. Using simulation we examine the properties of cumulative meta-analysis, such as the number of studies of a given size that must be run, on average

before significance is reached; the effect of repeated analyses on the false positive rate; and the accuracy of the estimated odds ratio.

IGES-51

Efficient candidate gene study design for association studies in complex diseases

S.A. Fisher, A. Moody, C.G. Mathew, C.M. Lewis
Division of Medical and Molecular Genetics, Guy's, King's and St Thomas' School of Medicine, London, UK

Association studies are widely used to identify complex disease susceptibility loci, following the limited success of genome-wide linkage studies. However, a considerable amount of work is involved in first identifying the genetic variation within a candidate gene, and then genotyping a large study cohort to identify disease association. A short genomic region can contain many single nucleotide polymorphisms (SNPs) with complex patterns of linkage disequilibrium (LD). Technological or financial factors may limit the number of SNPs that can be genotyped within a candidate gene. We describe a method of identifying informative subsets of SNPs, based on partial genotype data from dHPLC which only distinguishes between heterozygoteous and homozygoteous genotypes. We show how this data can be used to estimate SNP allele frequencies and linkage disequilibrium (LD) measures. For each pair of SNPs, haplotype frequencies are estimated assuming Hardy Weinberg equilibrium and used to calculate Δ. SNPs are assigned to LD groups where, for each pair of SNPs within an LD group, values of Δ exceed a threshold K. K is chosen so that the association study has sufficient power to detect association with a SNP which is in LD ($\Delta > K$) with the true disease susceptibility allele. A single SNP is then selected from each LD group and genotyped in the full study cohort. This efficient candidate gene study design is illustrated with SNPs from the chemokine gene cluster *SCYA22, SCYD1, SCYA17* on chromosome 16.

IGES-52

Neuropsychological phenotypes for identifying children with fetal alcohol syndrome

L. Flury[1], S.N. Mattson[2], P.W. Kodituwakku[3], C. Adnams[4], G. Turner[5], T-K. Li[1], E.P. Riley[2]
[1]Indiana Univ SOM, USA; [2]San Diego State Univ, USA; [3]Univ of New Mexico, USA; [4]Cape Town Univ, South Africa; [5]Trinity College, Ireland

Fetal alcohol syndrome (FAS) can occur in the offspring of women who drink heavily during pregnancy. The diagnosis is heavily dependent upon craniofacial anomalies and behavioral problems that are often encountered in these children. This project is an initial attempt to combine independent samples of FAS and control children collected in San Diego, Moscow and South Africa, representing very different ethnic groups and living conditions and thereby enhancing the power to identify neuropsychological measures that can accurately classify these children. Subtests of IQ measure were used in logistic regression analysis to predict FAS. A derived composite verbal/performance IQ measure and the digit span WISC subtest significantly predicted FAS ($p < 0.0001$, $p < 0.09$ respectively). Maximizing sensitivity and specificity yielded %corr = 69.7, %false neg = 0.1. Further analyses will use bootstrap techniques to obtain average error rates. These results demonstrate that control and FAS children can be accurately classified based on specific measures of intellectual ability, omitting measures of dysmorphia. Future studies will focus on the development of novel quantitative phenotypes that could be used in prospective studies of environmental and genetic contributors to FAS susceptibility, in addition to distinguishing nondysmorphic children prenatally exposed to alcohol and controls. Funded by AA07611, AA10820, AA10417, R01AA9440 and AA010417.

IGES-53

Prognosis and molecular genetic research

A. Fröhlich[1], U. Mansmann[2], A. Ziegler[1]
[1]Inst. of Med. Biometry and Statistics, Med. Univ. Lübeck, Germany; [2]Inst. of Med. Biometry and Informatics, Ruprecht-Karls-Univ. Heidelberg, Germany

With the rapid advance of molecular genetical and biological technologies, it is hoped that the array of prognostic markers available to pathologists greatly increases [Dunlop, 1997, BMJ]. However, unless principles of clinical epidemiological sciences are analogously applied, the molecular science that has produced magnificent genetic advances may also lead to an epidemic of harm from spurious associations, unnecessary fears, and unfulfilled promises of benefit [Bogardus et al., 1999, JAMA]. In this presentation, we firstly discuss the requirements of prognostic factor studies to be accepted in clinical routine. By use of examples from the literature, we secondly illustrate that these prerequisites are often violated in today's molecular marker studies. Thirdly, we discuss the enhancements necessary for gene expression studies identifying prognostic markers to be accepted in clinical practice.

IGES-54

Disentangling compound phenotypes and pleiotropic effects in multifactorial diseases using covariate adjustments: FCHL as an example

F. Gagnon[1,2], G.P. Jarvik[2], A.G. Motulsky[2], S.S. Deeb[2], J.D. Brunzell[2], E.M. Wijsman[2]
[1]Dept. of Epid. and Com. Medicine, Univ. of Ottawa, Canada; [2]Dept. of Medicine, Univ. of Washington, USA

Multifactorial diseases demonstrate complex patterns of inheritance and are characterized by an association of related phenotypes. Familial combined hyperlipidemia (FCHL) is an example of such complex diseases, for which no genetic defects of major effect have yet been identified. Pleiotropy affecting both plasma high-density lipoprotein (HDL) and triglycerides (TG) has been suggested. We completed a genome scan for HDL in 4 ($N = 255$) FCHL families ascertained through hyper-TG probands, and analyzed the relationship between TG and the quantitative trait loci (QTLs) for HDL. Using covariate-adjusted (age and sex) joint linkage and segregation analysis based on Bayesian MCMC, we localized 2 QTLs with individual contributions of 27% and 11% of total HDL variance, respectively. One of these QTLs is located at the APOC3 region on chromosome (ch) 11. Three QTLs with a smaller effect on the variance were also identified. Adding TG as a covariate (cov) decreased 4 out of the 5 linkage signals, including the signal at APOC3, suggesting pleiotropy at 4 QTLs. Joint analysis based on 90 markers of the ch with positive linkage signals further supported linkage of the two large QTLs identified in the initial analyses, with intensity ratios (IR) for linkage (calculated for 2 cM intervals) of 21 for the signal at APOC3 and 45 for the other locus. The IR is the posterior acceptance rate of QTL positions in an interval to the prior such rate. Analyses of Ln TG using the same joint ch analysis also revealed QTLs for TG with IR of 70 and 19 but in different locations. Using cov adjustments will help identifying pleiotropic effect of related phenotypes while simultaneous analyses of ch will be useful in fine mapping of genes implicated in the expression of multifactorial diseases.

IGES-55

The case-only design to detect G×E interaction for a survival trait

J. Gauderman, J. Millstein
USC Department of Preventive Medicine, USA

Many complex diseases exhibit variability in age of onset. Such traits are best analyzed using survival analysis methods that consider both disease status (D), and age of onset or censoring age (T). For analysis of $G \times E$ interaction, the proportional hazards model has form $\lambda(T) = \lambda_0(T) \times \exp(\alpha G + \beta E + \delta G \times E)$, and interest focuses on $\exp(\delta)$, a measure of interaction on the hazard rate scale. In the Children's Health Study (CHS), we follow children for new-onset asthma with a goal of testing for interaction between GSTM1 and air pollution. The CHS consists of over 6,000 children, too many to obtain complete genotype data, and thus we consider a case-only (CO) design to estimate δ. The standard CO approach estimates the odds-ratio (OR) between G and E in a sample of cases, without regard for T. However, this OR estimates $G \times E$ interaction on a relative-risk scale, and will be a biased estimator of $\exp(\delta)$ unless D is rare or followup time is short. We propose a modified CO analysis in which we fit a logistic regression for the relationship between G and E in cases, with terms added to the model to account for T. In preliminary work, we simulated 100 replicates of a cohort of 3,000 members, assuming a 3-yr disease rate of 19%. Assuming no $G \times E$ interaction ($\delta = 0$), a standard CO analysis yielded an inflated Type I error rate (50% instead of 5%), while our modified CO yielded a correct rate (4%). Under an alternative hypothesis of $\delta = 2$, our modified CO yielded an unbiased estimate (average estimate = 2.03) while the standard CO estimates were severely biased (average = −0.12). We compare performance under a variety of models.

IGES-56

Testing deviations from Hardy-Weinberg equilibrium in the single-locus case

F. Geller[1], T. Görg[1], A. Ziegler[2]
[1]Institute of Medical Biometry and Epidemiology, Philipps-University, Marburg, Germany; [2]Institute of Medical Biometry and Statistics, University of Lübeck, Germany

A well-known problem in the analysis of genetic data are deviations from Hardy-Weinberg equilibrium (HWE). Several tests are suggested in the literature to investigate general or specific deviations from HWE. We discuss tests that are suitable in the single-locus case with multi-allelic markers. We investigate the size and the power of these tests considering two alternatives, null alleles and population stratification. We simulate data for a microsatellite marker with 20 alleles in 100 and 500 individuals, respectively. Additionally, we investigate the same marker with reduced heterozygosity, combining 16 alleles with low frequencies to one allele. Null alleles are investigated by deleting one allele from all observations. Population stratification is generated by simulating an equal amount of genotypes from two different sets of allele probabilities. We perform the Guo-Thompson test, which is based on Markov chain Monte Carlo theory, the binomial test based on the total number of heterozygous genotypes in the data and the Chi-square-goodness-of-fit test. The asymptotic Chi-square-goodness-of-fit test does not hold the level for sparse tables and has low power in the remaining situations. In the case of population stratification all tests have low power. Thus, other methods using multi-locus information should be applied. Null alleles at minimum frequencies of 10% ($N = 100$) or 5% ($N = 500$) are detected with quite good power using the binomial test.

IGES-57

A likelihood-based haplotypic method to estimate the age of rare mutations

E. Génin[1], A. Tullio-Pelet[2], F. Begeot[2], S. Lyonnet[2], and L. Abel[3]
[1]INSERM U535, Le Kremlin Bicêtre, France; [2]INSERM U393, Hôpital Necker, Paris, France;

[3]INSERM U550, Université René Descartes, Paris, France

We present a new method to infer the age of rare mutations using the information provided by the size of the disease haplotypes shared by affected individuals. The age of the most recent common ancestor (MRCA) of the mutation is estimated by maximum likelihood, and a Bayesian confidence interval is provided. Simulations performed under the hypothesis that complete Identity-By-Descent (IBD) information is available show that the method provides satisfactory results with as few as two affected individuals, and quite reliable estimates when the sample contains at least five individuals. In the more realistic situation of incomplete IBD information, we propose a correction to account for frequencies of shared marker alleles that works very well except in the configuration combining high allele frequencies (.30) and large number of generations (100). In this latter case, densifying the map corrects the bias, and also makes the method more robust to allele frequency misspecification. We reanalyse two published datasets concerning the mutation causing Congenital Chloride Diarrhea and the CCR5-Δ;32 AIDS resistance allele. We find that our method provides age estimates consistent with the published ones, and, interestingly, leads to smaller confidence intervals. Finally, we use our method to date the origin of a recently identified founder mutation involved in Triple-A syndrome from nine patients and find an estimated age of 1175 years (95% CI 700–2000).

IGES-58

Detecting disease genes via a new Markov chain Monte Carlo approach for multipoint linkage analysis

A.W. George[1], E.M. Wijsman[2] and E.A. Thompson[1]
[1]Department of Statistics, University of Washington, USA; [2]Division of Medical Genetics, University of Washington, USA

Our ability to correctly detect and position chromosomal regions containing disease genes (called trait loci) is increased through the multipoint linkage analysis of data observed on large pedigrees and multiple genetic markers. Exact methods of computing the required multipoint likelihoods quickly become intractable with increasing pedigree complexity and increasing numbers of genetic markers. In this talk, we present a Markov chain Monte Carlo (MCMC) method for performing multipoint linkage analysis of data far exceeding the computational boundaries of exact methods. We will focus on the application of our MCMC method to data originating from a genetic study into early-onset Alzheimer's disease. Due to several of the pedigrees containing substantial missing data, previous attempts at estimating multipoint lod scores via MCMC methods were problematic. We will show that it is now possible to obtain multipoint lod scores quickly and accurately.

IGES-59

Non-parametric alternatives for sib-based QTL mapping: some statistical comparisons

S. Ghosh[1], P.P. Majumder[2], T. Reich[1]
[1]Department of Psychiatry, Washington University School of Medicine, USA; [2]Anthropology and Human Genetics Unit, Indian Statistical Institute, India

Unlike qualitative or binary traits which can be characterized completely by allele frequencies and genotypic penetrances, quantitative traits require an additional level of modeling: the probability distribution of the underlying trait. Hence, likelihood based methods like variance components, which requires assumptions like multivariate normality of trait values within a family, may yield misleading linkage inferences when underlying model assumptions are violated. The Haseman-Elston regression method (1972) and its extensions do not assume any specific probability distribution for the trait values, but assume a linear relationship between the squared sib-pair trait differences and the estimated identity-by-descent scores at a marker locus. Since it is often difficult to test the validity of these assumptions, it is of interest to explore for non-parametric alternatives. We (Ghosh and Majumder 2000) have therefore proposed that it may be strategically more judicious to empirically estimate the nature of dependence of the two above-mentioned variables using non-parametric diagnostics like rank correlation or kernel-smoothing regression. In this study, we extend our earlier methodologies to multipoint mapping and compare their performances to the linear regression procedures of Elston et al. (2000). Ghosh and Reich (2002) developed a so-called "contrast function" which integrates trait values within a sibship into a linear combination with sum of the coefficients zero. They proposed a test for linkage based on a linear regression of the squared contrast function on a quadratic function of the estimated i.b.d. scores of all possible sib-pairs within a sibship. We propose a non-parametric regression method on the lines of Ghosh and Majumder (2000) using the contrast function. We compare the power of the proposed method to those of the linear regression methods of Elston et al. (2000) and Ghosh and Reich (2002). For both sib-pair and sibship data, we find that while the non-parametric method is marginally less powerful than the linear regression methods in the absence of dominance, it performs increasingly better as dominance increases.

IGES-60

Evidence of a genetic locus for very-late onset alzheimer disease on chromosome 20

K.A.B. Goddard[1], J.M. Olson[1], H. Payami[2], M. Van der Voet[3], H. Kuivaniemi[3], G. Tromp[3]
[1]Case Western Reserve University, Cleveland, OH, USA; [2]Oregon Health Sciences University, Portland, OR, USA; [3]Wayne State University, Detroit, MI, USA

We recently reported results of a genome scan for late-onset Alzheimer Disease (AD), in which we utilized a novel covariate-based linkage approach, and demonstrated strong evidence of linkage on chromosome 20 in a subset of very old affected sib-pairs (ASPs) [Olson et al., Am J Hum Genet 71:154–161, 2002]. Here we report on subsequent association studies to further evaluate the candidate region on chromosome 20 for evidence of a susceptibility locus. The study subjects include 49 ASPs (89 individuals) recruited as part of the NIMH AD genetics initiative with a high probability of linkage to the region, and 129 healthy control individuals with an age distribution similar to the cases. We evaluated multiple chromosome 20 markers for an association with AD. The most compelling evidence of an association was at D20S200, where 54% of the ASPs are likely to share at least one 265 allele at this locus identical by descent (IBD) based on multipoint IBD sharing probabilities. Estimates of the population frequency for the 265 allele are about 23% from several sources. D20S200 is located approximately 1 cM from cystatin C (CST3), a strong candidate gene for AD. Polymorphisms in CST3 have previously been associated with AD, and CST3 co-localizes with amyloid beta in the brains of AD patients. These results support the existence of a novel susceptibility locus on chromosome 20 for very-late onset AD.

IGES-61

Using unconditional analysis in a case-sibling control design increases power to detect gene-environment (G×E) interaction

A.M. Goldstein[1], N. Andrieu[2], M-G. Dondon[3]
[1]Genetic Epidemiology Branch, DCEG, NCI, Bethesda, USA; [2]INSERM EMI00-06, Evry & Service de Biostatistiques, Institut Curie, Paris, France; [3]INSERM XU521, IGR, Villejuif, France

The interest in $G \times E$ interaction assessment continues to increase. Two types of control groups are generally used for examining $G \times E$ interactions: unrelated or related (e.g. sibling) controls. Although previous work has shown that relative controls have good power for detecting interaction when risk factors are rare, some of these studies may have assumed unrealistic numbers of available relative controls. Limited reviews of cancers have suggested that about 50% of cases may have one sibling control. In order to estimate both main effects and $G \times E$ interaction effects, the analysis should be a conditional analysis with each matched set comprised of a case and an unaffected sibling of the case (when a case has an available sibling control). The validity of this design requires no difference in the distribution of variables of interest between cases who have sibling-controls versus those cases without such sibling-controls and exchangeability of covariates of interest in cases and sibling controls. To further increase the power for detecting $G \times E$ interactions, we propose an unmatched analytic strategy. This strategy allows estimation of the $G \times E$ interaction effect and the E main effect under the assumption of no correlation in E between siblings. The G main effect cannot be correctly estimated from the unconditional analysis (however, it can be estimated from a separate conditional analysis). Since an unconditional analysis strategy generally uses more information from data than does a conditional analysis, this strategy may lead to improved efficiency for detecting interaction when compared to a conditional analytic approach. We illustrate the spectrum of improvements in power and discuss the limitations of this strategy.

IGES-62

Analysis of sulfotransferase 1A2 haplotypes in a colorectal adenoma case-control study

E.L. Goode[1], J.D. Potter[1], L. Fosdick[2], R.M. Bostick[3], and J. Bigler[1]
[1]Fred Hutchinson Cancer Research Center, Seattle WA, USA; [2]University of Minnesota, Minneapolis MN, USA; [3]University of South Carolina, Columbia SC, USA

Sulfotransferase 1A2 (SULT1A2) is involved with metabolism of estrogen and of non-steroidal anti-inflammatory drugs and is a candidate gene for colorectal adenoma and cancer susceptibility. We analyzed haplotypes based on these 2 SULT1A2 polymorphisms (P19L and N235T) in a study of 471 colorectal adenoma cases and 545 controls from Minneapolis, MN. The rare alleles of each polymorphism were found to be in negative linkage disequilibrium (LD) ($D' = -0.95$). Associations between haplotypes and adenoma risk were examined using a) logistic regression analysis assuming complete LD and b) haplotype estimation and "model-free" analysis of association (Zhao et al., 2000). When haplotypes were assigned assuming complete LD, individuals carrying 2 copies of the L-N haplotype ($n = 18$) appeared to be at decreased risk (OR 0.3; 95% CI 0.1–0.9) while individuals carrying a P-T and an L-N haplotype ($n = 73$) appeared to be at increased risk (OR 1.8; 95% CI 1.0–3.1) compared to individuals carrying 2 copies of the P-N haplotype ($n = 280$). However, comparison of likelihoods allowing versus not allowing association (maximized over mode of transmission) and tests for heterogeneity of haplotype frequencies did not support association ($p = 0.72$). Though the directions of the estimated ORs for certain SULT1A2 haplotypes suggest hypotheses about risk, these haplotypes overall may not be strong predictors of colorectal adenoma risk.

IGES-63

Genetic imprinting effects on body mass index in children

O. Gorlova[1], C. Amos[1], W. Wang[1], S. Shete[1], S. Turner[2], E. Boerwinkle[3]

[1]Dept Epidemiology, Univ Texas MD Anderson CA Ctr, Houston, TX, USA; [2]Division of Hypertension, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA; [3]Human Genetics Center and Institute of Molecular Medicine, The University of Texas Health Science Center, Houston, TX, USA

Obesity predisposes to hypertension, heart diseases and other chronic diseases and shows a high degree of familial aggregation. However, the genetic factors that predispose for the development of obesity are not yet well understood. Body mass index (BMI) is used as a measure of fatness, and obesity is defined by a threshold imposed on the distribution of BMI where individuals with values beyond the threshold are considered obese. Here we performed a genome-wide scan for genes related to BMI, while allowing for the possible effects of imprinting. We applied a sib pair linkage analysis to a sample of primarily children and young adults by using the Haseman-Elston method, which we modified to model the separate effects of paternally and maternally derived genetic factors. After stratification of sib pairs according to age, a number of regions showing linkage with BMI were identified. Most linkage and imprinting effects were found in children 5–11 years of age. Six regions showing evidence of imprinting were 3p23-p24 (locus-specific empirical p-value for the test for imprinting 0.036, paternal expression), 4q31.1-q32 ($p = 0.025$, maternal expression), 10p14-q11 ($p = 0.04$, paternal expression), 12p12-pter ($p = 0.02$, paternal expression) in children and 4q31-qter ($p = 0.0095$, paternal expression) and 8p ($p = 0.013$, paternal expression) in adults. Strongest evidences for linkage, all in children, were found on Chromosome 20 at 20p11.2-pter near the marker D20S851 (LODTotal = 4.22, $p = 0.000015$) and near the marker D20S482 (LODTotal = 3.55, $p = 0.000059$), and Chromosome 16 at 16p13 near the marker ATA41E04 (LODTotal = 3.12, $p = 0.00014$), and for those loci no evidence for imprinting was found.

IGES-64

Confirmation of linkage to and evidence of association at SLEB3

C. Gray-McGuire[1,2], J.A. Kelly[1], J.B. Harley[1,3,4]
[1]Arthritis and Immunology Program, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, USA; [2]Case Western Reserve University, Cleveland, Ohio, USA; [3]Department of Medicine, University of Oklahoma, Oklahoma City, Oklahoma; [4]US Department of Veteran Affairs Medical Center, Oklahoma City, Oklahoma, USA

Systemic lupus erythematosus (SLE) is a common, complex disease likely to involve several susceptibility genes. Linkages to several regions, including chromosome 4p16-15.2, support this position. A candidate locus, SLEB3, was identified in a genome scan of 77 European American (EA) pedigrees and supported by analysis of an additional, independently collected, 187 EA families. To localize and ultimately identify this gene, we genotyped 10 microsatellite markers over a 10cM region, first in the original sample (sample 1), and then an independent collection of 76 EA families (sample 2). Multipoint regression analysis of concordant and discordant full-sibling pairs resulted in maximum uncorrected p-values at D4s3007 of 2×10^{-4} in sample 1, 2×10^{-4} in sample 2, and 2×10^{-6} in the combined sample. Results also localized the linkage to approximately 5cM. To identify a genetic model, model-based linkage analysis was performed, resulting in a maximum LOD score (4.2) using a recessive model with 10% disease allele frequency and 100% penetrance. Association analysis of the 5cM support interval using haplotypes yielded uncorrected p-values of 0.003, 0.014, and 0.021. While linkage to SLEB3 is now localized to 4p16 association results do not further narrow the region indicating variability of linkage disequilibrium that must be addressed.

IGES-65

Haplotype Blocks, haplotype map, and haplotype mapping – the utility of haplotype structure analysis in search for complex disease genes

C.C. Gu[1], G.R. Abecasis[2], M.A. Province[1], D.C. Rao[1,3]
[1]Division of Biostatistics, Washington University School of Medicine, St. Louis, USA; [2]Center for Statistical Genetics, University of Michigan, Ann Arbor, USA; [3]Departments of Genetics and Psychiatry Washington University School of Medicine, St. Louis, USA

The establishment of large databases of single nucleotide polymorphism markers (SNP) across the whole genome has made massive genetic association analysis possible by scanning thousands of SNPs simultaneously. As efforts to unlock the information in these massive data sets continue recent studies of human population data indicate some structuring of linkage disequilibrium (LD). That is, there exist discrete blocks of haplotypes manifesting high LD, separated by hot spots of recombination where LD is suppressed. This can have implications both in designing studies of complex diseases, as well as in mapping the diseases genes. We perform simulation studies to analyze the effect of several design factors influencing the identification of block boundaries of the haplotype structure, especially the effect of admixture of a sample population. An empirical Monte Carlo procedure is applied to test for possible heterogeneity in the sample. Practical strategies are discussed to deal with the presence of heterogeneity, including the use of relevant covariates and possible gene-environmental interactions.

IGES-66

Genotype-by-age interactions influence normal quantitative variation in platelet-derived growth factor (PDGF) in the San Antonio Family Heart Study

L.M. Havill, M.C. Mahaney, L. Almasy, J. Blangero, J.W. MacCluer
Department of Genetics, Southwest Foundation for Biomedical Research, USA

We reported previously that genes and age contribute significantly to quantitative variation in platelet stored PDGF, a potent mitogen and chemotactant implicated in hemostasis and atherogenesis, in Mexican American families from San Antonio. We conducted the analyses reported here to better characterize the effects of genes and age, including their interactions, on variation in this substance. We assayed platelet stores of PDGF-AB in serum from 836 subjects, 16 to 92 years old, from 48 pedigrees. Quantitative genetic analyses, using a maximum likelihood based variance decomposition approach, estimated significant heritability ($h^2 = 0.31$, $p < 0.000001$) and a marginally significant mean decrease with age ($p = 0.07$). Neither sex nor any other age term exerted similar significant effects. To test for genotype-by-age ($G \times A$) interaction we modeled the additive genetic component of the variance in PDGF as a function of age. We also modeled the genetic correlation between relative pairs at different ages as an exponential decay function across age values. We estimated these two $G \times A$ interaction terms plus their environmental counterparts. Likelihood ratio tests revealed a significant decrease in genetic variance with increasing age ($p < 0.000001$), but the genetic correlation did not differ from 1.0. We conclude that, although the additive effects of genes decrease with age the same gene or suite of genes likely contributes to the variance in PDGF throughout adulthood.

IGES-67

Familial aggregation and prevalence in the restless legs syndrome

W.A. Hening[1], T. Washburn[2], R.P. Allen[2], C.J. Earley[2]
[1]UMDNJ-RW Johnson Med School, New Brunswick, NJ, USA; [2]Johns Hopkins Univ, Baltimore, MD, USA

In a preliminary study, we found an elevated frequency of restless legs syndrome (RLS) in relatives of patients (Allen et al., 2002). Another recent study found that RLS families whose proband had an age of onset of 30 or less showed a dominant pattern of inheritance in segregation analysis (Winkelmann et al., 2002). A linkage to chromosome 12 has been reported using a model of recessive inheritance (Desautels et al., 2001). We are currently engaged in a controlled family study of RLS. Consecutive RLS patients seen at the Johns Hopkins Bayview Medical Center were invited to participate in this study. After the proband signed an approved consent, relatives were contacted to provide background data and were interviewed by a blinded expert (WH) who determined RLS diagnosis using a validated telephone interview (Hening et al., 2001). First degree relatives of control probands free of RLS were similarly contacted and diagnosed. 50 families have now been analysed. Family aggregation was noted with 32 of 38 RLS families (84%) having at least one affected first degree relative member compared to 6 of 12 (50%) control families. The prevalence of RLS in first degree relatives of RLS proband families was 44% (57 of 130) compared to 24% (8 of 33) in control families. For probands with onset below 45 years, 47% (44 of 93) of first degree relatives were affected and for those with onset 30 or below, 54% (28 of 52). A very high percentage of relatives of RLS probands are affected. A still high, but distinctly lower, percentage of relatives of control probands also have RLS. Our work is consistent with previous studies (Allen et al., 2002; Winkelmann et al., 2002) which found more support for a genetic basis in families of those with earlier onset of RLS. RLS is most likely a complex disorder with important genetic and environment.

IGES-68

Octapeptide repeat insertions in the prion protein gene and transmissible spongiform encephalopathies

J.J. Houwing-Duistermaat[1], E.A. Croes[1], J. Theuns[2], C. Van Broeckhoven[2], C.M. Van Duijn[1,2]
[1]Dept. of Epi & Biostat, Erasmus MC, Rotterdam, The Netherlands; [2]Dept. of Mol Genet, Univ of Antwerp, Antwerpen, Belgium

To asses the relation of octapeptide repeat insertions in the prion protein gene (PRNP) with age at disease onset and disease duration in transmissible spongiform encephalopathies (TSE), a meta-analysis was conducted of all patients with PRNP octarepeat insertions (59 cases, 23 families). The relation of age at onset with the number of repeats was modelled using linear regression including a normal distributed random effect to estimate the residual correlation within families. The relation between disease duration and the number of repeats was studied using a Cox proportional hazards model adjusting for age at onset and for residual familial correlation by modelling a gamma distributed frailty. An increasing number of repeats was significantly associated with younger age at onset ($p = 0.001$). When adjusting for PRNP codon 129 (CDN129), a polymorphism determining susceptibility in Creutzfeldt-Jakob Disease the relation still hold ($p < 0.001$). Adding the CDN129 reduced familial correlation from 30% to 10%. Duration of disease increased significantly with length of the repeat ($p = 0.021$). This relation was borderline significant when also CDN129 was included ($p = 0.077$). In both Cox models residual familial correlation was negligible ($p > 0.900$). Our findings show a significant association of the number of

repeats with age at onset and duration in TSE independent of CDN129. Both genetic factors explain for a large part the observed familial correlation.

IGES-69

Familial myopia study

G. Ibay[1], B. Doan[2], L. Reider[3], M. Schlifka[3], M. Alexander[3], J. O'Neill[1], J. Bailey-Wilson[1], D. Stambolian[3]
[1]Inherited Disease Research Branch, NHGRI/NIH, USA; [2]Center for Inherited Disease Research, JHU; [3]Univ. of Pennsylvania, USA

Myopia, or nearsightedness, is highly prevalent in the United States and can lead to blindness in severe cases. Our goal is to identify regions of the human genome containing genes responsible for nonsyndromic myopia using pedigrees from four ethnic groups of the Myopia Family Study: Jewish, Amish, African- and Chinese-American families. Power studies were conducted on the first 44 Jewish families. Under tight linkage ($\rho \leq 0.1$), these families have good power of detecting linkage with a LOD score of 3.0 even when only 50% carry a disease gene linked to the same marker locus. With an expected number of about 200 families we anticipate adequate power to detect linkage even when only 24% of the families are linked to a single locus. Young et al. (1998a, 1998b) mapped major "high" myopia loci to chromosomes 12q21-q23 and 18p11.3. We genotyped 12 markers in these regions in 34 Jewish and 60 Amish pedigrees to determine whether these loci are also important in families with other, less severe, clinical forms of myopia. Although there was no strong evidence of linkage in either population 1 Amish family showed evidence of linkage (LOD >1.0) to the region previously reported on chromosome 12q; another 3 Amish families gave LOD >1.0 to 2 markers on chromosome 18p; and 3 Jewish families gave LOD >1.0 to 5 markers on chromosome 12q. Additional analyses of these candidate genes are ongoing as more families are collected. A genome wide scan of 46 Jewish and 66 Amish families is ongoing.

IGES-70

The "combination SNP test": a powerful strategy to detect the role of a candidate gene

A-S Jannot, L. Essioux, M.G. Reese, F. Clerget-Darpoux, INSERM U535, France

A topical question is the optimal use of the information provided by the genotyped SNPs in order to detect the role of a candidate gene in a multifactorial disease. We propose a strategy called "combination test" that tests the association between the phenotype and all possible phased combinations of a various number of SNPs. By simulating different models of correspondence between the genotype and the phenotype we show that the "combination SNP test" is able to detect the role of a candidate gene even when several functional SNPs interact with weak marginal effects for each. For this type of model, testing the association with each SNP separately or with the multilocus genotype does not enable to detect the role of the candidate gene. Interestingly, when there is no interaction between SNPs, the "combination SNP test" keeps a good power. So the "combination SNP test" is a suitable strategy for a wide type of situations. Its drawback is that the number of possible phased combinations increases drastically with the number of genotyped SNPs. To prevent from this increasing number of tests to perform we propose to test the combinations sequentially with an increasing number of SNPs included in the combination. We show the interest of this approach using GAW12 data.

IGES-71

Using word frequencies for testing the equivalence between two DNA sequences

I. Jansen[1] K. Van Steen[1], G. Molenberghs[1], M. De Wit[2], M. Peeters[2]
[1]Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium; [2]Tibotec-Virco, Mechelen, Belgium

Knowing the "closeness" between two DNA sequences is very important in the field of accrediting new laboratories. It can also be used as a measure of consistency if multiple sequences are generated by a single lab. The paper of Wu, Burke and Davison (1997) introduced a family of word-based dissimilarity measures defining a distance between two sequences by simultaneously comparing the frequencies of all n-words (sequence of n adjacent nucleotides) in both sequences. The Mahalanobis distance, accounting for both the variances and covariances between frequencies of n-words, turned out to give the best performance. We make some crucial modifications to this Mahalanobis distance and apply it to real life data about accrediting new labs, i.e. on testing whether DNA sequences composed in different labs are sufficiently close. Afterwards we develop an equivalence test, using simulation studies. Evaluating this test results in a sufficiently high power (approximately 98%). Before calculating the distance between two DNA sequences, several parameters need to be chosen, such as the word size n, the window length in which the word frequencies are calculated, etc. The performed simulations also gather knowledge about the impact of these assumptions. We can conclude that there is no single best choice for those parameters. Nevertheless, conclusions about the equivalence between both DNA sequences are consistent over different parameter choices. The introduced methodology can also be extended to amino acid sequences instead of sequences of adjacent nucleotides.

IGES-72

Modelling the familial clustering of breast cancer for individual risk assessment

M.A. Jonker[1], W.E. Hoogendoorn[2], G.H. De Bock[2], J.C.Van Houwelingen[1]
[1]Department of Medical Statistics; [2]Department of Medical Decision Making, Leiden University Medical Centre, Leiden, The Netherlands

At family cancer clinics an individual's risk to develop breast cancer is estimated based on the family history of breast and ovarian cancer. To provide accurate risk assessment it is important to use a genetic model that explains the real life breast cancer risks, including the population incidence and relative risks for females with an affected closely related family member. BRCA1 and BRCA2 explain only a small proportion of these increased risks. Therefore we considered two genetic models with in addition to BRCA1 and BRCA2 either a dominant or a recessive third hypothetical gene, called BRCAu. The model parameters concerning BRCAu (allele-frequency and penetrance function) were estimated so that the three genes jointly explain the empirical (relative) risks. BRCAu then fills up the gap between the relative risks observed in real life and the risks due to BRCA1 and BRCA2. For the estimation only published estimates of the population incidence and relative risks were used [1]. To evaluate the two estimated models as they are used for individual risk assessment based on complete families, a simulation study was performed. The study showed that for most counselees the advice for further diagnostics is similar for both models. [1] Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies. Lancet 2001; 358:1389–99.

IGES-73

Polygenic features of a family history of CHD

M.G. Brandt, A.M. Levin, S.L.R. Kardia
Department of Epidemiology, University of Michigan, USA

Family history of CHD is well established as a predictor of an individual's future risk of CHD. The explanation for the observed familial disease aggregation is not well understood except for the general knowledge that genetic and environmental factors predisposing to CHD also aggregate in families. In the Rochester Family Heart Study, we divided the initial 256 three-generation pedigrees into those with a strong and weak family history based on the CHD status of their 4 grandparents. Thirty-four single nucleotide polymorphisms (SNPs) in 14 genes (*ACE, ADD, ADRB2, AGT, APO E, APO B, APO AI-CIII-AIV, DRD2, LDLR, LIPC, LPL, PON 1*) were measured in these families. With few exceptions (i.e. *PON 1*) variations in these loci do not significantly discriminate between children with and without a strong family history ($P < .05$). However, when we defined potential susceptibility genotype classes liberally (using $P < 0.5$ to capture small effects) and use replication in male and female children as an additional constraint, we found that 10 SNPs were potentially informative. Counting the overall number of susceptibility genotypes per person we found that male children ($N = 75$) with a strong family history carried an average of 26% of these susceptibility genotypes versus 21% in male children ($N = 113$) with a weak family history ($P = 0.07$). Female children ($N = 79$) with a strong family history had an average of 28% of these susceptibility genotypes compared to 22% in female children with a weak family history of CHD ($P = 0.02$). This study demonstrates that devising ways to measure the cumulative effect of many genes may help researchers understand the predictive nature of family history information.

IGES-74

Choice of sib-pair design when sibling relative risk is small

K.J. Keen
Department of Epidemiology and Biostatistics, Case Western Reserve University, USA

Sib-pair analysis for linkage was introduced by L. S. Penrose in 1935. His approach was to analyze sibships ascertained with one or more affected sibs, not two or more as in the affected sib-pair (ASP) design. Penrose's method thus provides controls against environmental similarities among sibs, whereas, the modern ASP design does not. Nevertheless, the ASP design has been demonstrated to be remarkably successful in detecting linkage for highly penetrant recessive Mendelian disorders. In this study, we consider a disorder, of an unknown genetic model, of late onset with an estimated sibling relative risk of approximately 2.0. As noted by Risch, small values of sibling relative risk argue against use of the ASP method. For 80% power at a p-value of 0.0001 (corresponding approximately to a LOD score of 3.0), the estimate of the relative efficiency of the discordant sib-pair (DSP) design with respect to the ASP design is 168% assuming a dominant model and 435% assuming a recessive model. Modification of the ASP design to incorporate unaffected pairs is demonstrated not to improve efficiency with respect to the DSP design in this scenario. Therefore, blind adoption of the ASP design cannot be recommended in all situations. Indeed the DSP design provides some measure of control with respect to similarities among sibs analogous to the original design introduced by Penrose.

IGES-75

Non-parametric linkage analysis of the thiazide-sensitive Na-Cl cotransporter gene in families with a history of hypertension

A. Keszei, A. Mente, S.B. Bull, A. Tisler, A.G. Logan
Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada

The thiazide-sensitive Na-Cl cotransporter (TSC) plays a major role in sodium and chloride

reabsorption in the distal convoluted tubule of the kidney. Mutations in the TSC gene at 16q13 influence blood pressure regulation and urinary calcium excretion. Calcium abnormalities are associated with sodium sensitive forms of hypertension, suggesting the possibility of the TSC gene as a candidate for hypertension susceptibility. To test for linkage between hypertension or hypercalciuria and the TSC gene, we collected 55 families in whom the probands were under the age of 50 years had a positive family history of hypertension and excreted increased urinary calcium ($Ca^{2+} \geqslant 0.5$ mmol per mmol creatinine). Consenting probands and first-degree relatives provided a venous blood sample for genotyping and a first morning urine sample for calcium determination. Five polymorphic microsatellite markers over a region of 10 cM spanning the TSC gene locus were used for genotyping. Affection status was defined as the presence of hypercalciuria or drug treated hypertension. Data were analysed using non-parametric allele-sharing method (Genehunter). Analysis of all the families resulted in a peak NPL score of 2.45 ($P = 0.002$) at marker D16S3253 flanking the TSC gene locus. A subset of 48 Caucasian families revealed a maximum NPL score of 2.95 ($P = 0.0003$) at the same marker. These data suggest a role of the genetic region containing the TSC gene in essential hypertension associated with increased urinary calcium excretion.

IGES-76

Familial aggregation of ocular refraction

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

Ocular refraction refers to the power of the external lens to bring images into focus on the retina. Refractive errors, myopia (nearsightedness) and hyperopia (farsightedness) are common conditions that require corrective lenses. It has been well established that myopia clusters in families and linkage to regions on chromosome 12q and 18q has been shown through studying highly myopic individuals. The potential for genetic effects through the entire range of refraction has been less well studied. Twin studies have indicated a high heritability for overall refraction and there is evidence showing that refraction is highly correlated between siblings. This study examines the familial aggregation and pattern of inheritance of ocular refraction in an adult population using data collected as part of the Beaver Dam Eye study. Analyses were based upon 2,138 individuals with complete data in 620 extended pedigrees. After adjustment for age and education correlation between pairs of relatives was assessed using FCOR. Substantial positive correlation was found between siblings (0.35) and parents and offspring (0.24) with lower correlation among cousins (0.10) and avuncular pairs (0.09). Measurements between spouses were negatively correlated (−0.16). Commingling analysis indicated that multiple distributions provided a better fit to the data than does a single distribution. Formal segregation analysis results will be presented.

IGES-78

Combined linkage analysis of data sets typed for different marker maps

D.L. Koller, D. Lai, T. Foroud
Dept. of Medical and Molecular Genetics
Indiana Univ. School of Medicine, Indianapolis, Indiana, USA

Marker screening sets employed for genome screens have evolved rapidly in recent years. In consequence, ongoing studies of complex traits often include sample subsets typed for different markers. Etzel and Guerra (AJHG 71:56, 2002) have proposed a meta-analysis technique allowing for heterogeneity in marker maps and other study parameters. We have performed a simulation study to examine the effect on QTL LOD scores in the worst-case scenario of differing marker maps: two subsets of the data each typed for 10 cM screens, but with marker sets staggered by 5 cM. Quantitative trait and marker data were simulated using the GASP package (AJHG 59:A193, 1996). Data sets of size 200-1000 sib pairs were generated, with varying values of QTL effect size, parents available per pair, and position of the QTL within the marker map. Trait heritability was assumed to be 80%, and marker data ($PIC = 0.8$) was generated for an 80 cM chromosome. Mapmaker/SIBS (AJHG 57:439, 1995) was used to compute LOD scores for 1000 replicates of each set of parameter combinations. LOD scores were compared for two situations: 1) each half of the sample typed for staggered 10 cM maps, and 2) the full sample typed for all markers. LOD scores decreased by approximately 15% for the staggered case as compared to the fully typed case, with the decrease ranging from 12% to 18% across all combinations of parameters. We conclude that in some situations differing marker sets could act with other between-study heterogeneity to cause loss of power in combined and meta-analytic studies of complex disease.

IGES-79

Group sequential study designs for the analysis of candidate genes in case control studies

I.R. König, A. Ziegler
Medical University of Lübeck, Institute of Medical Biometry and Statistics, Germany

In the past years, the focus of genetic-epidemiological studies has shifted to the analysis of complex diseases. Here, a single gene often contributes only little to the manifestation of a trait; hence, many patients have to be included in a study to reliably detect small effects. To reduce the number of required phenotypings and genotypings in a study

and thus facilitate the analysis of complex traits, sequential study designs can be applied. For the sequential analysis of candidate genes in association studies, we describe at first the procedure by Sobell et al. [1]. This includes the successive testing of many candidate genes with an adjustment of the significance level. Thus detected associations are validated in independent samples. Based on results from Monte-Carlo simulations, we discuss the efficiency of this procedure. Secondly, we present the adaptation of optimized group sequential study designs by Müller and Schäfer [2] to the analysis of candidate genes. In this procedure, the sample of cases and controls is enlarged sequentially; after the genotyping of each subsample, association is analyzed in the cumulative data. Error rates and the efficiency of this proceeding are similarly investigated by Monte-Carlo simulations. Finally, we compare both procedures regarding error rates, efficiency, and practical applicability. Literature: [1] Sobell J. L.Heston L. L.Sommer S. S. (1993) Am J Med Genet 48:28–35. [2] Müller H.-H.Schäfer, H. (1999) Stat Med 18:1769–1788.

IGES-80

Estimation of haplotype frequencies and measures of linkage disequilibrium using SAS/Genetics

I.R. König, A. Ziegler
University of Lübeck, Institute of Medical Biometry and Statistics, Germany

With the introduction of the module Genetics, SAS Institute Inc. has promised new statistical procedures specifically tailored for genetic data. The newly available procedures are specifically designed for marker data and aim at supplying tools needed for the association mapping of a complex trait or disease (http://www.sas.com/rnd/papers/sugi27/genetics.pdf). The features include the examination of marker properties (Hardy-Weinberg equilibrium, marker informativeness), estimation and tests of haplotype frequencies (expectation-maximization algorithm), and measures of linkage disequilibrium. The new module Genetics might offer some advantages over the use of available software packages. Firstly, data are entered in a commonly used format. Secondly, many calculations might be possible with the same data format without having to switch programs. The key disadvantage is the costs of purchasing the SAS module. We therefore compared the module SAS/Genetics with freely available software. Using data from a case-control study, we focused on the ease of data entry, comparability of results, and further features.

IGES-81

The Contribution of genes and smoking to the transmission of pancreatic cancer susceptibility

J.F. Korczak[1], O.E. Johnson[1], M. Schenk[1], M. Kinnard[2], A.G. Schwartz[1], D.H. Garabrant[3]
[1]Karmanos Cancer Inst. and Wayne State University, USA; [2]Case Western Reserve University, USA; [3]University of Michigan, USA

We recently reported familial aggregation of pancreatic cancer (PC), based on cancer histories of first-degree relatives of 247 newly diagnosed PC patients and 420 population-based controls in southeastern Michigan (Schenk et al., 2001). To determine whether transmission of PC within the families of the 247 PC patients (probands) was consistent with a genetic etiology, we performed segregation analysis using Class A regressive logistic Models 1 (variable age of onset) and 2 (variable susceptibility) as implemented in the S.A.G.E. program REGTL. For each model, hypotheses of genetic, environmental, or no transmission were tested against a general model with arbitrary transmission probabilities. Because Schenk et al. (2001) also found the relatives' smoking history to be a significant risk factor for PC, analyses were performed both without and with smoking status as a covariate. According to the Akaike Information Criterion (AIC), Model 1 fit the data as well or better than Model 2 for all transmission hypotheses. Chi square tests were predominantly significant for the non-genetic transmission hypotheses but not for the genetic hypotheses, indicating a better fit of the genetic hypotheses to the data. Including smoking as a covariate significantly improved the fit of the models. The most parsimonious model was that of recessive inheritance of an allele for early onset of pancreatic cancer, including smoking. Since Schenk et al. (2001) also found a significant interaction between the relatives' smoking status and family history of PC in the subset of families of young onset probands we are pursuing further segregation analysis of this subset and plan to model a gene-smoking interaction in the analysis.

IGES-82

p53 codon 72 polymorphism and persistence of cervical infection by human papillomavirus

A. Koushik[1], T. Rabachini[3], M-C. Rousseau[1], A. Ghosh[2], G. Matlashewski[2], L. Villa[2], E. Franco[1]
[1]Epidemiology and Oncology; [2]Microbiology & Immunology, McGill University, Canada; [3]Virology Unit, Ludwig Institute for Cancer Research, Brazil

The Arg/Arg genotype vs. Arg/Pro or Pro/Pro at codon 72 of the p53 gene has been implicated in increasing susceptibility of the cervix to human papillomavirus (HPV) infection and thus altering cancer risk. Of a number of studies, however, only a few have been supportive. One study demonstrated that genotype results using allele-specific PCR varied substantially across different laboratories, suggesting that misclassification due to random genotyping errors may have led to many previous null findings. We examined this polymorphism in relation to persistence of cervical infection by HPV in a cohort

of sexually active women in Brazil, using three genotyping methods: Taqman, dot blot, and denaturing high-performance liquid chromatography. Genotype agreement between the three methods was high, with Kappa statistics ranging from 0.86–0.94. Among women with 3 agreed genotype results, a non-significant increased risk for overall persistence of HPV types 16 or 18 was observed for Arg/Arg. In particular, the association was high for long-term persistence: odds ratio (OR) = 2.7, 95% confidence interval 0.5–16.4. Despite high agreement among genotyping methods, when the same analyses were done separately by method, ORs for long-term persistence varied from 1.5–2.6. Thus, Arg/Arg may increase the risk of persistence of HPV-16/18. However, our results show that an elevated association can be obscured when only one method is used.

IGES-83

Family-based expression-association studies

P. Kraft[1], S. Horvath[1,2]
[1]Department of Biostatistics; [2]Department of Human Genetics, University of California, Los Angeles, CA, USA

Messenger RNA transcript abundances a.k.a. gene expression levels potentially provide more detailed and immediate information on the genetics of complex traits than simple genotypes. We discuss the prospects for detecting associations between expression levels and disease using case-control studies and present a statistic for family-based expression-association studies. This Pedigree Expression Test (PET) is similar to the Pedigree Disequilibrium Test (PDT) proposed by Martin et al. [Am J Hum Genet 67 (2000) 146–154]. Where the PDT exploits Mendelian laws for the transmission of alleles, the PET is based on analogous rules for expression levels. We review relevant data on the heritability and transmission of expression levels. Family-based studies protect against population-stratification bias, which we argue is as reasonable a concern for expression-association studies as for gene-association studies. However, as in the gene-association case, family-based controls can be less powerful then population-based controls. Simulation studies show that sibling controls are 20%–45% less powerful than population controls in common situations. We apply PET to both real and simulated data sets.

IGES-84

Negative effect of smoking on bone density is modulated by estrogen receptor beta gene in postmenopausal French Canadian women

N. Laflamme[1,5], S. Giroux[2], G. Cardinal[2], S. Dodin[3], C. Blanchette[3], K. Morgan[4], F. Rousseau[2]
[1]INSPQ; [2]URGHM; [3]U. Rech. Endocrinol. Reprod.Dept. Biol. Méd. & Ob.-Gyn.U. Laval, Québec, Canada; [4]Dept. of Hum. Gen. & Med.McGill U. & Res. Inst. McGill U. Health Centre; [5]SignalGene Inc. Montreal, Canada.

Osteoporosis is a multifactorial disease that involves environmental and genetic factors. Bone density is highly inherited but little is known on the interaction between environmental and genetic factors. Our aim was to study the influence of the estrogen-receptor beta (ESR2) gene alone, and in interaction with environmental risk factors, on bone density in a large sample of women. In a cross-sectional study, bone density of 1189 ambulatory postmenopausal French Canadian women was evaluated by right calcaneal quantitative ultrasound. Results were expressed as broadband ultrasound attenuation (BUA), speed of sound and stiffness index. Women were genotyped by allele-specific PCR for a coding (G/A) polymorphism in the 5th exon of the ESR2 gene. This polymorphism was in HW equilibrium. An exploratory analysis (ANCOVA) was performed in a first randomized subgroup of 572 women. A negative association was observed between smoking and bone density mostly in ESR2 heterozygotes. The same association was replicated in the remaining independent subgroup of 617 women. In the final analysis comprising all women, those bearing the GA genotype (8%) who were smoking had a 13 db/MHz (1.3 SD) lower heel-BUA ($p < 0.0001$) than non-smoking women bearing the same genotype (interaction factor ESR2*smoking: $p = 0.001$). This may translate into a 2 to 3 fold difference in the risk of fracture. The potential confounding effect of standard risk factors for low bone density were evaluated and did not affect results. These findings suggest that the negative effect of smoking on heel bone density in postmenopausal women is modulated by variation at the ESR2 locus.

IGES-85

Familial aggregation of components of the multiple metabolic syndrome

K.E. Lee, B.E.K. Klein, R. Klein
Department of Ophthalmology and Visual Science, University of Wisconsin Medical School, Madison, WI, USA

The multiple metabolic syndrome is defined by a cluster of risk factors including hyperglycemia, hypertension, central obesity and dyslipidemia. Persons with the syndrome have elevated risks for cardiovascular disease as well as diabetes. We wish to quantify the familial aggregation of the components of the syndrome in the Beaver Dam Eye Study population. The study population consists of 4,926 adults between the ages of 43–86 years at a baseline exam. Measurements of glucose, glycosylated hemoglobin, cholesterol (total and HDL), and uric acid were obtained from casual blood samples. Mean arterial blood pressure (MABP) and body mass index (BMI) were calculated from blood pressure and height and weight measurements obtained during a standardized examination. We identified familial relationships among all participants. Analyses were based on the 2,104 people in 602 extended pedigrees with complete data who were free of diabetes at the baseline exam. After adjustment for age and gender effects, correlations

between pairs of relatives were assessed using FCOR in SAGE. Sibling correlations were high for serum cholesterol (0.14 (total), 0.26 (HDL)), BMI (0.23), MABP (0.12) and uric acid (0.19) but not for glycosylated hemoglobin (0.06). Parent-child correlations were also high for these same measures, except MABP. Cousin and avuncular correlations were considerably lower. Spousal correlations were generally low, except for uric acid (0.27). Within an individual, uric acid and BMI were correlated with all other measures. Between siblings, no measure was strongly correlated with any other measurement. This data is supportive of a familial relationship for serum cholesterol, BMI and blood pressure, which are risk factors for the development of cardiovascular disease.

IGES-86

Genomewide Scan for Genes Influencing Blood Pressure in Mexican Americans

D.M. Lehman[1], R. Arya[1], L. Almasy[2], R.J. Leach[1], P. O'Connell[3], J. Blangero[2], M.P. Stern[1], R. Duggirala[2]
[1]Depts. of Medicine and Cellular and Structural Biology, University of Texas Health Science Center at San Antonio, USA; [2]Southwest Foundation for Biomedical Research, USA; [3]Virginia Commonwealth University, USA

The intraindividual variation in blood pressure levels has been shown to be heritable, yet the gene(s) influencing blood pressure are largely unknown. We have performed genomewide scans for susceptibility genes influencing blood pressure in a Mexican American population. The data relate to 351 individuals distributed across 27 low-income extended Mexican American pedigrees of the San Antonio Family Diabetes Study (average age 48.2 ± 15.4, 39.3% hypertensive, 39% male). We used a variance components technique to conduct multipoint linkage analyses for systolic (SBP) and diastolic blood pressure (DBP). After accounting for the effects of blood pressure medication, sex and age terms, we found significant evidence for linkage (LOD of 3.49, $p = 0.000062$) of SBP measures to a genetic location between markers D10S568 and D10S676 on chromosome 10q and for DBP measures (LOD of 3.12, $p = 0.00015$) to a location near marker D8S284 on chromosome 8q. These regions continued to exhibit significant evidence for linkage even after accounting for the effects of diabetes status or body mass index. In addition, we also observed weak evidence for linkage ($p < 0.01$) of SBP to chromosomes 1p (LOD of 1.30, D1S220), 11p (LOD 1.44, D11S988), and 19p (LOD of 1.2, D19S247). Regions exhibiting weak evidence of linkage to DBP were chromosomes 7p (LOD 1.26, D7S510), 10q (LOD 1.67, D10S568-D10S676), 15q (LOD 1.32, D15S11), and 19q (LOD 1.34, D19S714). In conclusion, these results implicate a novel QTL at chromosome 10q that influences variability in SBP and replicates a number of the previously published loci for blood pressure.

IGES-87

Estimation of the inbreeding coefficient from multipoint marker data

A.L. Leutenegger[1,2], B. Prum[3], F. Clerget-Darpoux[2], E.A. Thompson[1]
[1]University of Washington, USA; [2]INSERM, France; [3]CNRS, France

Many linkage studies are done in small isolated populations and populations where marriages between relatives are encouraged. In such populations, accurate genealogies might not be available. It has been shown that underestimation of relationships may lead to false linkage conclusions. We propose a maximum likelihood method for estimating the inbreeding coefficient (f) of individuals from such populations using a hidden Markov model with the EM-algorithm. This methodology also allows us to infer the full probability distribution of the identity-by-descent (IBD) status at each point along the genome and provides a variance for the estimates. We simulate a full genome scan of 672 markers (every 5cM) for (1) first cousin pedigree (2) quadruple-second cousin pedigree. We present 95% confidence intervals for the values of f and examine the accuracy of the imputation of the IBD segments. We show how these results are sensitive to marker allele frequencies and marker spacing.

IGES-88

Powerful and Robust Exact Family Based Tests of Linkage and Association

J.P. Lewinger[1,2], S.B. Bull[2,3]
[1]Dept. of Stat; [2]Samuel Lunenfeld Research Institute; [3]Dept. of Public Health Sciences, Univ of Toronto, Canada

The Transmission Disequilibrium Test (TDT) is a simple family based test of linkage. It is also a valid test of allelic association when only one affected child per nuclear family is included. Conditional on parental genotypes and children's phenotypes, the null distribution of the TDT statistic is independent of the unknown mode of inheritance and population structure, making the TDT an exact test. However, the TDT requires complete parental genotypic information, and its power is reduced by ignoring the unaffected children and the parental phenotypes. We introduce exact tests of linkage and association that make use of all available information in a nuclear family including the ascertainment when this is known. The tests are based on likelihood ratio statistics obtained from a standard two point linkage model extended to include allelic association parameters and ascertainment corrections. All parameters are estimated subject to appropriate constraints. The likelihood method allows families with any pattern of missing data and arbitrary number of affected and unaffected children. Exact and robust tests are then obtained by conditioning on sufficient statistics for a nonparametric family of distributions for the appropriate null hypothesis (D. Rabinowitz, N. Laird, Hum Hered 2000; 50: 211–223). Additionally, p-values can be efficiently

estimated by Monte Carlo importance sampling under the fitted model. We show by simulation that these tests have greater power than the TDT for a range of relevant scenarios.

IGES-89

Definition of family history of stroke in predicting stroke among female young adults

R. Li[1], B. Mitchell[2], W. Giles[3], Q. Song[1], G. Gibbons[1], S. Kittner[2]
[1]Morehouse School of Medicine, USA; [2]University of Maryland School of Medicine, USA; [3]CDC, USA

Family aggregation of stroke depends on family size and age of family members. We compared different definitions of family history of stroke in predicting young adult (age 15–44 years) stroke in the Baltimore-Washington Cooperative Young Stroke Study. The study included 252 (133 white, 119 African American, AA) female stroke patients, and 369 (236 white, 133 AA) age-matched female controls. Family history of stroke and the age at disease onset was collected from probands' parents and full siblings. Family size, not including proband, ranged from 1 to 15 for cases and 1–14 for controls. The number of stroke in family members was 0–3 in case family and 0–2 in control family. The family history of stroke was defined as mother/father without stroke or with stroke at age $\leqslant 55$, >55, number of stroke patients in a family, family risk score using Hunt's method (FRS1) and FRS using Silberberg's method (FRS2). The association between family history of stroke and stroke was assessed by logistic regression with and without adjusted for probands' age, smoking status, hypertension, diabetes and oral contraceptive use.

Mother stroke at age $\leqslant 55$ strongly predicts daughter's stroke, suggesting significant genetic effects between mother and daughter and emphasizing the increased genetic component of early onset stroke. Other definitions of family history of stroke less strongly predicted stroke in the proband.

IGES-90

PPL based re-analysis of a genome screen for schizophrenia

M.W. Logue[1,2], V.J. Vieland[1,2], R.J. Goedken[1], E.W.C. Chow[3], A.S. Bassett[3], L.M. Brzustowicz[4,5]
[1]Division of Statistical Genetics; [2]Dept. of Psychiatry, University of Iowa, USA; [3]Dept. of Psychiatry, University of Toronto, Canada; [4]Dept. of Genetics, Rutgers University, USA; [5]Dept. of Psychiatry, NJMS, UMDNJ

The PPL has been developed as a Bayesian alternative to traditional linkage analysis. It differs from both lod scores and "non-parametric" methods in several respects: it directly measures the probability of linkage, given the data and it incorporates prior genomic information including the prior probability of linkage. Recently we have implemented a version of the PPL which treats the trait model as a vector of nuisance parameters which are integrated out of the likelihood. This differs from other methods that it is based on the likelihood as a function of the trait model, but without fixing the model at any specific value. In the current application we have re-analyzed the Canadian schizophrenia data set reported by Brzustowicz et al. [Science, 2000] using the model-integrated PPL. As expected, salient results are highly correlated with the original analyses, with the initially-reported maximum heterogeneity lod under a recessive model of 5.8 corresponding, at the same marker, to a PPL of 98.7%. However, the PPL graphs have some advantages over the lod graphs: a single graph suffices, rather than separate graphs for different genetic models; and the graphs show extremely clean results over the rest of the genome. Furthermore, localization based on additional markers in the region of interest appears to be very precise.

IGES-91

Likelihood for quantitative and threshold traits under a mixed model of inheritance revisited

J.C. Loredo-Osti[1], B.R. Smith[2], K. Morgan[1]
[1]Dept. of Genetics and Medicine, McGill University and the Research Institute of the McGill University Health Centre, Montreal, Canada; [2]Mathematics and Statistics Dept., Dalhousie University, Halifax, Canada

Computing the likelihood for a mixed model of inheritance is an issue regardless of the complexity of the pedigree. In 1974, Morton and MacLean (Am J Hum Genet 26:489-503) developed in detail a method for evaluating the likelihood for a vector Y of observations of a quantitative trait in nuclear families with unrelated parents under a mixed model of inheritance where every element Y_i comes from a mixture of normals with the same variance, and the means are determined by the oligogenic part of the model. Also, they sketched a way to evaluate the likelihood of $Z = I_{(Y>t)}$, the threshold trait with underlying liability Y and threshold value t. The key feature of the Morton and MacLean method is that given the parental phenotypes and the vector of genotypes, x, amongst the offspring, the distribution of the vector of differences $Y_i - x_I$ is invariant under permutation. Therefore, there exists an independent random variable such that conditioning on it, the distribution of these differences are independent. This holds even if the parents are related. Here, the method is extended and a numerical procedure for quantitative threshold and bivariate traits in general nuclear families is proposed. The extension can be used in a Monte Carlo sampling scheme in blocks of sibships with phenotypic and genotypic data at the bottom of an extended pedigree.

IGES-92

Power estimation for quantitative trait linkage for the CEPH reference families

A. Malhotra, M. F. Leppert, S. J. Hasstedt
University of Utah, Salt Lake City, UT, USA

The Centre d'Etude du Polymorphisme Humain (CEPH) project began as a collaborative effort to map genetic markers using pedigree data. Each pedigree collected consisted of 4 grandparents, 2 parents, and 8–10 offspring. These families were collected without regard to any disease phenotype. Recently, members of 30 Utah CEPH pedigrees have undergone a battery of tests and have been measured for a large number of quantitative phenotypes. A simulation study is being performed to estimate the power to detect linkage of a quantitative trait locus to a linked marker under both dominant and recessive models for members of these 30 families. We have simulated 500 replicates for this purpose. We assumed a heterozygosity of 75% and a recombination fraction of 0.04 based on the marker information already available for the CEPH families. We are estimating the power for a two-point linkage analysis using parametric and variance components methods. Using phenotypes simulated under a dominant model, power estimates ranged from 50%–93% depending on the displacement and prevalence assumed. Power estimates for the variance components method are comparable to parametric methods in these pedigrees. A power of less than 10% was estimated when the dominance is misspecified. These results strongly suggest that these data have sufficient information to detect linkage using both parametric and nonparametric methods. In addition, the power to detect linkage under a misspecified model is significantly low. We are currently analyzing phenotypes simulated under a recessive model. In addition, 2-locus models will be studied.

IGES-93

Using allele transmission data from both parents and multiple affecteds to detect linkage and association

J.D. Malley[1], J.E. Bailey-Wilson[2], R.A. Redner[3], T.A. Severini[4], S. Pajevic[5]
[1]NIH, CIT; [2]NIH, NHGRI; [3]University of Tulsa, USA; [4]Northwestern University, USA; [5]NIH, CIT

The standard TDT statistic uses allele transmission data from single (statistically independent) parents as a simple hypothesis test of the null model of no association and/or no linkage. When the null model is rejected it does not provide any estimates or confidence intervals for association (linkage disequilibrium) or linkage. As a more comprehensive and potentially more powerful alternative we derive the likelihood for allele transmission data to multiple affecteds from both parents (in a family trio, with parents not assumed independent), and obtain separate confidence intervals for association and linkage. We find that power is improved for detecting linkage or association: in our example with strong association present, and a single affected, we detect theta = 0.3 with power .975, and detect association with power 1.00, while the TDT has power .948 (for association and/or linkage) (all with the nominal alpha level at or below 0.05). More importantly, we provide confidence intervals for linkage and association with high coverage probabilities, and hence disentangle the effects of the linkage and association parameters. We show how to extend the method to multiple affecteds, again with increases in power over the conventional TDTwhich only allows data from a single affected.

IGES-94

A genome-wide scan of a myotonic dystrophy pedigree and statistical inference of allele-sharing statistics in the context of a single multiplex pedigree

M. Martinez[1], I. Le Ber[2], D. Campion[3], D Hannequin[2]
[1]INSERM00-06, Evry, France; [2]Dept. of Neurology, CHU Rouen, France; [3]INSERM99-06, Rouen, France

Myotonic Dystophy (DM) is an inherited disease characterized by progressive muscle weakness, myotonia, ocular cataracts, and cardiac arrhythmias. We have identified a five-generation French family in which, as in a few number of other families with similar phenotype, linkage to known DM1 and DM2 loci was excluded, suggesting that at least one other susceptibility gene may exist. We have conducted a genome-wide scan of our DM pedigree with 392 markers (CNG-panel), and using an "affected-only" parametric LOD score approach, assuming dominant inheritance and no phenocopies. The power and the maximum information in the pedigree that could be obtained with fine mapping was evaluated through simulations using SLINK. The expected maximum LOD is 2.38, and 74%, 41% and 0% of replications have a LOD > 1, 2 or 3, respectively. Two-point LOD > 1 were found on 8 chromosomes, and the highest peak was on chromosome 15q22-q24 (LOD = 2.38). Typing further 24 markers yielded similar results (peak multipoint LOD = 2.38). Non-parametric analyses of allele sharing among affected relatives (MERLIN v0.9.2) confirmed these results. However, to our surprise, significance was much higher with NPL = 12.82 (nominal-$P < 10^{-7}$) than with Lod-Zlr = 1.53 (nominal-$P = 4 \times 10^{-3}$). Evaluation of the empirical significance rates showed that, in our data, NPL is liberal and Lod-Zlr is conservative and the adjusted significance levels of our observed scores are 7.8×10^{-4} (Lod-Zlr) and 7.7×10^{-4} (NPL).

IGES-95

Evaluating heterogeneity of linkage to the thiazide-sensitive Na-Cl cotransporter gene by positive and negative family history of hypertension

A. Mente, S.B. Bull, L. Mirea, A. Keszei, A.G. Logan
Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada

We demonstrated suggestive linkage between the thiazide-sensitive Na-Cl cotransporter (TSC) gene and hypertension or increased urinary calcium excretion in families with a positive history of hypertension. To ascertain the role of family history in determining linkage we collected 32 additional families in whom there was no family history of hypertension. All other selection criteria, genotyping procedures and determination of affectation status were identical to those with a positive family history of hypertension. The results showed no linkage at the TSC gene locus for the families with a negative family history of hypertension (NPL score −0.37 compared to that of 2.45 for positive family history). To compare whether the NPL scores demonstrated significant heterogeneity we developed a new permutation test that is applicable to families since previous similar procedures were limited to affected-sib pairs. The null hypothesis is that there is equal identity-by-descent allele sharing among pairs of affected relatives in families with a positive and negative family history of hypertension. A rejection of the null hypothesis would indicate that linkage is specific to only families with a positive family history of hypertension.

IGES-96

Reducing dimensionality in the search for multi-gene interactions

J. Millstein, W.J. Gauderman
Division of Biostatistics, Department of Preventive Medicine, University of Southern California, USA

We propose a method, Dimension Reduction for Multi-gene Interactions (DRMI), for identifying multi-way interactions among a set of candidate genes. The goal is to solve the multiple testing problem of assessing all possible combinations, e.g. with 20 candidate genes, there are 1,140 possible 3-way interactions. Assuming a case-control sample, clusters of subjects are formed based on genotype similarity for all combinations of K genes, where K is the highest order interaction under consideration (e.g. K = 3 for 3-gene interactions). Genotype similarity is defined using a score that relates the observed sharing of alleles to that expected based on population allele frequencies. Disease status is not used during the clustering process. The subsets of K genes yielding the densest clusters are then analyzed for main and interactive effects using standard logistic regression. In preliminary work into the feasibility of this approach, we simulated 30 replicate data sets of 20 candidate genes from 200 cases and 200 controls. Under the null hypothesis of no interactions among any genes, our empirical estimate of the Type I error rate matched the desired level (0.05). Under an alternative hypothesis that a subset of 3 of the 20 genes interacted to increase disease risk by a factor of 4.0, we were able to reduce the 1, 140 possible 3-way interactions to 20 or less in 23 (77%) out of the 30 replicates. In 20 (87%) of these 23 replicates, the true interaction was significant at the 0.05 level. The multiple comparison issue regarding the final 20 subsets will be discussed.

IGES-97

Multifactor dimensionality reduction is an ideal discriminator of discrete clinical endpoints using multilocus SNP genotypes

J.H. Moore, L.W. Hahn
Program in Human Genetics, Vanderbilt Univ., USA

The identification and characterization of genes whose effects are primarily through interactions with other genes and environmental factors remains a statistical and computational challenge in the genetic epidemiology of common, complex multifactorial diseases. We have previously developed a multifactor dimensionality reduction (MDR) approach to identifying gene-gene and gene-environment interactions in case-control and discordant sib-pair study designs. MDR is nonparametric in that no parameters are estimated and is genetic-model free in that no particular genetic model is assumed. Using both simulated and real data, we have demonstrated that MDR has excellent power for identifying high-order gene-gene interactions in the absence of any detectable independent main effects of each gene. Further, we have demonstrated that the power of MDR to identify gene-gene interactions is robust in the presence of genotyping error and phenocopy. In the present study, we outline a mathematical proof that MDR ideally discriminates between discrete clinical endpoints using multilocus SNP genotypes. In this proof, we first define the context of the problem, the MDR decision rule, and the error evaluation on which the decision rule is optimized. We then prove that the MDR decision rule is optimal for the given error evaluation. This proof suggests that no analytical approach will classify discrete clinical endpoints using multilocus genotypes better than MDR. Based on this proof we propose that MDR should be the ''gold standard'' with which other methods for identifying gene-gene and gene-environment interactions are compared.

IGES-98

A novel strategy for selecting optimal subsets of SNPs for the analysis of gene-gene interactions

J.H. Moore, L.W. Hahn, B.C. White
Program in Human Genetics, Vanderbilt Univ., USA

A dense map of single nucleotide polymorphisms (SNPs) is expected to facilitate the identification of

disease susceptibility genes. A complicating factor is the reality that the effects of some genes will only be detected when considered in the context of other genes (i.e. epistasis). The implication of this reality for using SNPs to identify disease susceptibility genes is that each SNP must be considered in the context of other SNPs. The evaluation of subsets of SNPs presents an enormous combinatorial problem when the total number of SNPs is large. Even with an accurate haplotype map, powerful computer systems and efficient search algorithms will be needed to facilitate the identification of optimal SNP subsets from an effectively infinite number of possibilities. We have implemented and evaluated a novel parallel search algorithm that is inspired by the problem solving abilities of ant colonies for food foraging. With this ant colony optimization (ACO) approach, each artificial ant initially carries out a random walk through multilocus SNP space. Each set of SNPs selected by an artificial ant is statistically evaluated and the probability of selecting those SNPs is increased or decreased based on how well the set of SNPs distinguishes affected from unaffected individuals (i.e. the statistical fitness). This is analogous to updating a pheromone trail. Many ants are evaluated and the process of updating the virtual pheromone trail is iterated until the artificial ant colony converges on a best SNP subset (i.e. a single pheromone trail to a food source). Using simulated data with 10,000 SNPs we demonstrate the ACO approach is competitive with other search methods such as parallel genetic algorithms. The simplicity and efficiency of the ACO approach make this an attractive algorithm for the development of SNP subset selection software.

IGES-99

A genetic locus for systemic lupus erythematosus susceptibility linked to chromosome 11p in African-American families stratified by the presence of discoid lesions

S.K. Nath, B. Namjou, J. Kilpatrick, R.H. Schofield, J.B. Harley
Arthritis and Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, USA

Systemic Lupus Erythematosus (SLE), a chronic, complex disease, is the prototype of human autoimmune diseases. Although there are many environmental factors that play a crucial role in triggering SLE, its genetic predisposition has been established by family as well as twin studies. During the past few years, there has been considerable interest in identifying of genomic segments linked to SLE through either whole genome scan or candidate gene approach. One of the major cutaneous manifestations in SLE, which is more prevalent in African American patients, is discoid lesions (DLE). In the present study, we have identified 58 multiplex families, 27 African American (AA), 26 European American (EA) and 5 others, where at least one SLE patient also afflicted with DLE, from the collection of families as a part of our ongoing linkage study for SLE. A genome-wide parametric and non-parametric linkage analyses was conducted with 320 markers. Significant evidence of linkage was identified in one chromosomal location (11p13) for 27 AA families. The maximum two-point and multipoint lod scores were 5.5 and 4.5, respectively, obtained at 47 cM. The segregation behavior of this gene indicates a dominant mode of inheritance. These results suggest that grouping families based on their clinical features and racial origin not only homogenize families but also increase power to detect linkage for complex diseases.

IGES-100

A comparison of measures to determine the important inputs to an artificial neural network with application to a late onset Alzheimer's disease dataset

R.J. Neuman, N.L. Saccone, J.P. Rice, A. Goate, L. Sun.
Dept. of Psychiatry, Washington Univ.St. Louis, MO, USA

Our main focus in this study was to compare various measures which are used to determine the inputs which are most important in artificial neural network (ANN) statistical analysis. To that end, we implemented alternative measures of importance and evaluated them by examining the consistency of the results over alternative network architectures. Previous authors have demonstrated a lack of repeatability of results in certain situations (Marinov and Weeks, Hum Hered 2001;51:169–1760). We applied ANNs to an Alzheimer's Disease (AD) dataset containing 429 affected sibpairs and 247 sibpairs discordant for AD (DSP). In both cases, inputs to the ANNs were the sibpairs' mean IBD data at multiple loci. Target outputs for both datasets were the ASP and DSP phenotypes. A complete genome scan had been performed previously on the AD data using traditional IBD sharing methods (Myers et al. Am J Med Genet 2002;114:235–44). Cross-validation was used to identify networks demonstrating good discrimination between the two phenotypic classes. As a first step in ANN analysis of the AD data each chromosome was considered separately. One of the purported strengths of ANN techniques is the ability to analyze markers residing on different chromosomes simultaneously. Accordingly, additional analyses used multiple chromosomal regions to detect epistatic interactions among non-syntenic loci. We were able to demonstrate that many measures of importance were extremely consistent in the order of importance in which inputs were ranked. However, this consistency deteriorated somewhat over alternative network architectures. Nevertheless, for some ANN architectures, we obtained results that were concordant with chromosomal regions identified in the Myers et al. report.

IGES-101

Interaction between MTHFR haplotypes, dietary methionine intake, and aggressive prostate cancer

N.L. Nock[1], L. Li[1], F.R. Schumacher[1], M. Cicek[2], G. Casey[2], and J.S. Witte[1]
[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio; [2]Department of Cancer Biology, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH, USA

The etiology of prostate cancer is complex and appears to be driven by both genetic and environmental factors. Methylenetetrahydrofolate reductase (MTHFR) and other products and cofactors in the methione and folate pathways affect DNA methylation and synthesis. Previous reports indicate that polymorphisms in the MTHFR gene and deficiencies in dietary intake of folate and methione play a role in cancer development, but their potential impact on prostate cancer has not been well studied. Here, we evaluate the effect on prostate cancer of the interaction between haplotypes defined by two MTHFR polymorphisms (C677T and A1298C) and dietary intake of methionine in a moderately large family-based case-control study (416 cases, 449 sibling controls). We found an inverse association between the 677T – 1298A haplotype and prostate cancer (OR = 0.67, 95% CI = 0.46–0.97, p = 0.03). Furthermore, we observed an interaction between the T-A haplotype and increased methionine intake (p = 0.02) among men with more advanced disease. Our findings suggest that the 677T and 1298A haplotype, in methionine sufficiency, is associated with a reduced risk of prostate cancer among men with more advanced disease. These results provide considerable support for the role of DNA methylation in prostate cancer aggressiveness.

IGES-102

Associations between selected candidate genes and type 2 diabetes after stratification by physical activity and dietary fat

J.M. Norris[1], C.D. Engelman[1], M.M. Barmada[2], R.E. Ferrell[2]
[1]Dept. of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, CO, USA; [2]Dept. of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA

Type 2 diabetes mellitus (T2DM) may result from an interaction between environmental exposures and genetic susceptibility. Both high dietary fat and low physical activity have been associated with T2DM. Investigation of candidate genes for T2DM have produced inconsistent results, possibly because potential gene-environment interactions were not addressed. We used the family based association test (FBAT) (Hum Hered 2000;50:227–33) to look for linkage and association to T2DM in 421 affected Hispanic individuals from 92 extended families living in the San Luis Valley, Colorado. The FBAT, a generalization of the transmission/disequilibrium test, allows for missing parental genotypes, and accounts for the relatedness of affecteds when using large pedigrees. We conducted an analysis overall, and in the upper and lower tertiles of average occupational and leisure physical activity prior to diabetes diagnosis and current percent dietary fat intake of the affected siblings. No linkage/association was observed between T2DM and the beta-3 adrenergic receptor (B3AR) (p = 0.14), the fatty acid binding protein (FABP) (p = 0.49), and the lamin gene (LMNA) (p = 0.76) when all families were analyzed. However, both B3AR and LMNA were associated with T2DM in families in which the affecteds had low physical activity (lower tertile) prior to diagnosis (p = 0.017 and 0.048, respectively). FABP was associated with T2DM in families in which the affecteds had a high fat diet (upper tertile) (p = 0.003). In conclusion, the association between candidate genes and T2DM may be more readily detected when environmental exposures are taken into account.

IGES-103

A family study of alcoholism: combining sources of diagnostic information

J.I. Nurnberger Jr., S. O'Connor, E.T. Meyer, T. Reich, J. Rice, M. Schuckit, L. King, T. Petti, L. Bierut, K. Bucholz
Institute of Psychiatric Research, Department of Psychiatry, Indiana University School of Medicine, and the Collaborative Study of the Genetics of Alcoholism (COGA), USA

The COGA study includes diagnostic information on 8412 first-degree relatives of alcoholics as well as 1365 first degree relatives of controls. The majority have algorithmic diagnoses based on SSAGA interview. A subset of COGA subjects (2267, or 26% of the interviewed sample) have best estimate diagnoses. More than a third of first-degree relatives, however, have not been directly interviewed. A previous analysis of family history data (Rice et al, 1995) concluded that three or more implications of alcohol dependence by relatives predicted a positive interview diagnosis 98% of the time. In the present analysis we compare family history to best estimate diagnosis (N = 1044). Among persons with 1–3 implications of alcohol dependence by a relative, 69.4% received a best estimate diagnosis of alcohol dependence by DSM-III-R. With four or more implications the figure was 92.2%. With no or uncertain implications, 26.2% of first-degree relatives were diagnosed with alcohol dependence. Among first-degree relatives without a direct interview (N = 3641), 2117 had 1–3 implications and 707 four or more implications. We have derived estimates of percent affected among non-interviewed relatives using such analyses. Data are now being cumulated and age-corrected to estimate lifetime morbid risk and relative risk for alcohol dependence and other Axis I diagnoses among relatives and controls.

IGES-104

Genetic dissection of linkage to complex human diseases: Application to late-onset Alzheimer disease

J.M. Olson, K.A.B. Goddard, D.M. Dudek, Y. Song
Department of Epidemiology and Biostatistics, Case Western Reserve University, USA

Locus heterogeneity presents a challenge in the analysis and interpretation of linkage results for common, complex human diseases. We propose a systematic approach for partitioning a set of affected sib pairs (ASPs) into genetically homogeneous subsets. First, genome scan data is analyzed using an ASP linkage method that allows inclusion of covariates that serve as surrogate measures of locus heterogeneity. The largest covariate signal(s) is selected and the estimated model parameters are used to estimate the proportion of ASPs linked to the selected genomic location(s) and to assign to each ASP, conditional on its covariate value and estimated allele-sharing probabilities, a probability of linkage to the location. The linked subset is then isolated by choosing the cutpoint on the linkage probability measure that maximizes the difference in lod score between the linked and unlinked subsets. The unlinked subset is then reanalyzed by repeating the steps above, until all subsets are identified and isolated. Finally, all subsets are reanalyzed using the baseline (without covariates) model. We use the method of false discovery rate to control for Type I error due to multiple testing. An application of this strategy to ASPs available through the NIMH Alzheimer Disease Genetics Initiative reveals three genetically homogeneous subsets of ASPs, each characterized by linkage to two to three distinct loci and varying average ages of onset. Multilocus models with covariates describe the interactions among loci within each subset.

IGES-105

Evidence of Linkage of Parkinson Disease to Chromosome 2q comes mainly from families with additional affected individuals

N. Pankratz[1], W.C. Nichols[2], S.K. Uniacke[2], C. Halter[1], A. Rudolph[3], C. Shults[4, 5], P.M. Conneally[1], T. Foroud[1], Parkinson Study Group
[1]Indiana Univ SOM, USA; [2]Cincinnati Children's Hospital Research Foundation, USA; [3]Univ Rochester, USA; [4]Univ California, San Diego, USA; [5]VA San Diego Healthcare System, USA

To identify susceptibility genes contributing to the more common, later onset Parkinson disease (PD), families consisting of at least one pair of living affected siblings were recruited. A Diagnostic Checklist with inclusion criteria consisting of clinical features highly associated with autopsy-confirmed PD and exclusion criteria highly associated with other non-PD pathological diagnoses was used to classify study subjects as having verified PD (VPD) or nonverified PD. Families with a positive lod score in the parkin gene were screened for mutations, and mutation-positive families were removed from the subsequent genome-wide linkage analyses. Affected sibpair linkage analysis in a sample of 126 sibling pairs with VPD yields a lod score of 3.0 on chromosome 2q. When only the 37 (of 118) families with an affected parent are analyzed under an autosomal dominant model of PD inheritance, the lod score is 2.1. Further limiting the sample to the 31 families having at least four affected individuals among the 1st or 2nd degree relatives of the affected sibling pair yields a lod score of 2.5. Thus stratifying the sample based on a more extensive family history of PD has identified a subset of families with a potentially Mendelian form of disease in which to focus ongoing molecular studies of chromosome 2q.
Study supported by: NINDS NS37167.

IGES-106

Screening for genetic and environmental effects in association studies through a cross-classified data paradigm

V.S. Pankratz
Division of Biostatistics, Mayo Clinic, USA

Genetic association studies are often used to examine multiple genetic and environmental factors in relationship to the presence of disease. In addition to examining the associations between single factors and disease, these studies make it possible to examine the interplay among the factors of interest, including interactions. To address these issues, appropriate statistical methods are needed. Data analysis methods that were developed for cross-classified data can effectively make use of the multiple experimental factors to address research questions of interest. These analysis methods are a direct extension of methods commonly used for 2×2 contingency tables. As such, they are based on modeling the structure of multi-way contingency tables. Goodness-of-fit tests, and tests of specific model parameters, are then used to address hypotheses of interest. These methods have a number of strengths. They include the ability to simultaneously test alleles of multiple loci and multiple environmental factors, individually and in interactions. Also, the goodness-of-fit statistic asymptotically follows a chi-square distribution. However there are limitations. To completely assess the structural relationships among all factors of interest, it is necessary to observe at least one observation for each of the possible combinations of the various factors. Even more are needed for asymptotic reasons. I use simulated data sets to illustrate the properties of this analysis method when used as a screening tool to identify main and interactive effects of multiple genetic and environmental factors.

IGES-107

Confidence set of markers tightly linked to a disease gene: the effect of marker polymorphism

C. Papachristou, S. Lin
Department of Statistics, Ohio State University, USA

The effect of incomplete marker polymorphism on a confidence set of markers tightly linked to a disease locus is examined here. A non-parametric test statistic, termed modified mean test, is defined for data from affected sib pairs, and their parents if available. Based upon this statistic, a confidence set is constructed. The construction of the confidence requires specification of the disease incidence data, which can be usually estimated from the population. Simulations are performed to assess the properties of the new approach under several factors, including the disease model, amount of data and heterozygosity of markers. The results show that for a wide range of disease models, when the number of alleles of the marker is not too small, say five, an increase of about 20–70% in the amount of data is sufficient to achieve the same power as when a 100% polymorphic marker is used. Furthermore, parental data, if available, usually help to reduce the number of families needed to achieve comparable power. Finally, the approach can be extended to include families for which information is available for only one parent, or additional siblings.

IGES-108

Methods for testing familial aggregation of diseases in population-based samples: Application to lymphoproliferative cancers in Swedish registry data.

R. Pfeiffer[1], M. Gail[1], K. Hemminki[2], L. Goldin[1]
[1]DCEG, NCI, USA; [2]Karolinska Inst. Stockholm Sweden

Familial aggregation of diseases can be tested in case-control or cohort studies. A unique cohort for conducting such tests is the Swedish Family Cancer Database which includes 10.2 million individuals with defined familial relationships, over 1 million of whom have a confirmed cancer diagnosis. From this database, we selected relatives of all cases diagnosed with a lympho-proliferative cancer along with relatives of matched controls. We propose a survival analysis method for testing familial aggregation comparing disease occurrence in families ascertained through a case proband with disease occurrence in families ascertained through a control proband. We extend a method by Liang (1991), letting t_{ij} denote the age or age at onset of disease for member j in family i and modeling it by a marginal Cox model $\lambda(t_{ij}|\ X_{ij},\ Z_{ij}) = \lambda_0(t_{ij})\ \exp(\beta X_{ij} + \alpha Z_{ij})$. λ_0 is the baseline hazard function, X_{ij} denotes measured covariates for that individual, e.g. gender, and Z_{ij} is an indicator of the proband's disease status ($Z_{ij} = 1$ if the proband of family i is a case and 0 otherwise). Testing for familial aggregation corresponds to testing $\alpha = 0$. The parameters are estimated under a working independence assumption accounting for truncation arising from the database design. The robust sandwich covariance matrix accounts for the dependence of the family members. Hodgkin's disease (HD) is used to illustrate the method. Here we find an increased risk of HD in first degree relatives of cases compared to controls.

IGES-109

Long-term survival model for genetic association studies of complex disease with variable age at onset: small simulation study

J. Pitkäniemi[1,2], L. Haapala[2], E. Moltchanova[2], V. Hyttinen[2]
[1]University of Helsinki, School of Medicine, Finland; [2]National Public Health Institute, Finland

We propose a long-term survival (LTS) model for genetic association studies when the population under study consists of susceptible subjects with variable age at onset and non-susceptible subjects. Using the EM-algorithm we maximize complete the data likelihood function of a simple statistical model with two parts: logistic regression model for susceptibility (alpha and beta) and Poisson regression model (gamma and delta) for the age at onset. In our simulation model disease gene affects both susceptibility and age at onset and 38% of the population is expected to be susceptible. At the age of 25 subjects were censored if still disease free. We simulated 200 datasets with 500 subjects and results of the simulation are shown below.

For the proportion of susceptibles the mean value in the simulation was 38.33 % (s.d. 0.023). Our small simulation example shows that we obtain unbiased estimates of both susceptibility and age at onset parameters of interest. In the LTS model information of susceptibility proportion is based on the long enough follow-up and not on the unlinked marker information that is hard to obtain. Clearly, our simulation model is too simple for modeling of complex traits like CHD or diabetes, and therefore we plan to use a more flexible survival model for the age at onset and Bayesian methods. To evaluate properties of the proposed model and compare this model with other traditional association models a more comprehensive simulation is needed.

IGES-110

A recessive major gene predisposes to human herpesvirus-8 infection in an endemic population of African origin.

S. Plancoulaine[1,2], A. Gessain[2], M. Van Beveren[2], P. Tortevoye[2], L. Abel[1].
[1]Human Genetics of Infectious Diseases, INSERM U550, Paris, France; [2]EPVO, Pasteur Institute, Paris, France.

In a large endemic population of African origin from French Guiana, infection by human herpesvirus-8 (HHV-8), the aetiological agent of all forms of Kaposi's sarcoma, was recently found to exhibit familial aggregation with mother-child and sib-sib correlations (Plancoulaine S et al, Lancet, 2000; 356:1062–65). To investigate whether this familial

pattern could be explained in part by genetic factors, we conducted a segregation analysis in all pedigrees of the same villages from French Guiana. The study included 83 pedigrees of African origin with 1623 subjects (871 females, 752 males) of whom 193 (11.9%) were HHV-8 positive. The analysis was performed using regressive logistic models which test for the presence of a major gene taking into account covariates influencing HHV-8 infection (i.e. age) and other sources of familial dependences (i.e. due to virus transmission routes). Results show the presence of a recessive major gene predisposing to HHV-8 infection with a significant residual mother–child correlation. The frequency of the allele D predisposing to HHV-8 infection is estimated at 0.22 indicating that $\sim 5\%$ of the population is DD homozygous. The probability to be HHV-8 positive with HHV-8 negative mother is >0.7 by age 12 for DD subjects, while it is <0.1 by age 35 for other genotypes. Furthermore, most of HHV-8 positive children <15 years with a negative mother are predicted to be genetic cases while most of those with a positive mother are sporadic cases. In addition to familial routes of viral transmission (likely through saliva), familial aggregation of HHV-8 positivity in endemic populations could be explained by a genetic predisposition to infection. Linkage studies with genetic markers are ongoing to confirm and identify this major gene.

IGES-111

Adiponectin and Cardiovascular Risk in the Old Order Amish

T.I. Pollin, B.D. Mitchell, W-C. Hsueh, K. Tanner, A.R. Shuldiner, J.C. McLenithan
University of Maryland, Baltimore, MD, USA and University of California, San Francisco, CA, USA

Adiponectin is a fat-derived hormone involved in insulin sensitivity and free fatty acid clearance. We measured adiponectin levels via radioimmunoassay in 126 subjects from the Amish Family Diabetes Study. Adiponectin levels increased with age independent of sex and body mass index (BMI) (Spearman's $r=0.50$, $p<0.0001$), were significantly higher in females compared to males (13.5 ± 1.1 vs. 9.9 ± 1.1 μg/ml, $p=0.0001$, corrected for age and BMI), and were not significantly associated with diabetes in this sample. Adiponectin levels were negatively correlated with obesity independent of age and sex (BMI, $r=-0.37$, $p<0.0001$; fat mass, $r=-0.26$, $p=0.01$) and with several insulin resistance-related traits, including fasting insulin ($r=-0.28$, $p=0.005$), fasting glucose ($r=-0.23$, $p=0.01$), the HOMA insulin resistance model ($r=-0.31$, $p=0.0006$), and triglycerides ($r=-0.28$, $p=0.001$), independent of age, sex and BMI. Similarly, adiponectin and HDL cholesterol were positively correlated ($r=0.24$, $p=0.02$). Adiponectin and leptin were inversely correlated, controlling for age and sex ($r=-0.34$, $p=0.0001$). The relationship between decreased adiponectin level and insulin resistance and cardiovascular risk factors appears strong in the Amish and similar to other populations. These results, coupled with prior studies showing this trait to be heritable provide a rationale for undertaking further studies in this founder population to determine if genes influencing this trait also influence susceptibility to diabetes and heart disease.

IGES-112

A sequential multiple decision procedure (SMDP) for microarray gene expression analysis

M.A. Province
Division of Biostatistics, Washington Univ. School of Medicine, St. Louis, MO, USA

The SMDP theory of Bechhofer, Kiefer and Sobel (1968) is used to develop a global, sequential, statistical test for the gene discovery/classification problem of microarray analysis. The sequential test adds gene chips one at a time, re-analyzing until a pre-specified stopping criterion is reached, wherein the process automatically terminates identifying the signal genes. Traditional fixed-sample statistics must trade power for the inflated experiment-wise type I error arising from the multiple testing paradigm. The problem is very severe in microarray analysis due to the large number of genes tested on each chip. However, sequential tests allow tight control of both type I and type II errors without trading one for the other, by letting sample size vary. Sequential tests are known to be amongst the most efficient, requiring smaller N on average than fixed sample ones. This efficiency is especially important for expensive microarray experiments. Further, the SMDP avoids multiple testing entirely. At each stage a single test is done for all genes on the chip at once instead of testing each gene individually. It therefore automatically defines the optimal way to split a fixed sample of chips between a hypothesis generating set and the confirmation/refutation set. The efficiency of the method is demonstrated via simulation, as well as in application to the publicly available data of Golub and Lander (1999) [Science, 286:531–537] on 72 gene chips each measuring the expression levels of 6,817 genes to distinguish acute myeloid leukemia from acute lymphoblastic leukemia tumors.

IGES-113

A genome scan for autoimmunity loci using lupus-related antibody profiles

P.S. Ramos[1], J.A. Kelly[2], C.M. Meyer[1], A.N. Leiran[1], W.A. Ortmann[1], K.J. Espe[1], J.B. Harley[2,3], K.L. Moser[1]
[1]Dept. of Medicine, Univ. of Minnesota, USA; [2]Oklahoma Medical Research Foundation, USA; [3]Dept. of Medicine, Univ. of Oklahoma Health Sciences Center, USA

Autoantibodies are found in all patients with systemic lupus erythematosus (SLE) and many of their unaffected relatives. We sought to identify loci that predispose to humoral autoimmunity in human SLE. We have tested for intrafamilial association of SLE autoantibodies between siblings from 229 multiplex SLE pedigrees and found significant concordance for antinuclear antibodies, anti-La and anti-nRNP ($p<0.005$). Based on the presence of SLE autoantibodies, we have defined a novel phenotype for lupus-related autoimmunity (LRA) and performed linkage analysis. We reclassified 1668 total subjects (525 SLE affecteds) for the affectation status and genotyped 279 microsatellites across the genome. A total of 850 individuals were reclassified as affected for the LRA phenotype. We used the revised Haseman-Elston algorithm (SIBPAL) to identify regions of increased allele sharing within these pedigrees. We found strong evidence for linkage ($p<0.005$) in 131 European-American pedigrees at 7p21, 7q11 and 15q24, and in 73 African-American pedigrees at 12q24. Linkage to 7p21, 12q24 and 15q24 was also observed when all 229 pedigrees were analyzed. Comparison of these loci to previously reported results indicates overlap with linkages previously identified in rheumatoid arthritis (7p15), SLE (7q21) and inflammatory bowel disease (12q24), while the effect at 15q25 appears to be novel. These results provide evidence of the presence and locations of genes that are involved in the susceptibility to development of an intermediate autoimmune phenotype and are likely to help further unravel the complexity of SLE as well as other autoimmune diseases.

IGES-114

Heritability and linkage analysis of premature coronary artery disease and related traits based on a study of 428 ascertained multiplex families

S–Q. Rao[1,2], G–Q. Shen[1,2], R. Cannata[1], E. Zirzow[1], R.C. Elston[3], X. Li[4], L. Li[1,2], E.F. Plow[2], E.J. Topol[1,2], Q. Wang[1,2]
[1]Center for Cardiovascular Genetics, Dept. of Cardiovascular Medicine, Cleveland Clinic Foundation, USA; [2]Dept. of Molecular Cardiology, Cleveland Clinic Foundation, USA; [3]Dept. of Epid. & Biostat, Case Western Reserve Univ, USA; [4]Dept. of Math, Harbin Medical Univ., China

Coronary artery disease is a complex disease with contributions from both genetic and environmental factors. Sample heritability was assessed and linkage analysis was performed to elucidate its genetic determinants. In the heritability analysis, we estimated genetic and environmental variance components of premature coronary artery disease (PCAD) and related traits in 2030 participants from 428 multiplex families in the Cleveland Clinic Foundation Gene Quest Study. The two types of analysis, mixed linear models and family aggregation analysis, agreed with each other in implicating the existence of genetic cases in the manifestation of PCAD and related phenotypes. The sample heritability was 0.45 for PCAD, and varied considerably for the related phenotypes. In the linkage analysis, we genotyped markers D2S129 and D14S1426 with an ABI3100 Genetic Analyzer. The two markers were found to be significantly linked to coronary heart disease in the Finnish population (Pajukanta P. et al., 2000, Am. J. Human Genet. 67:1481–1493) and to myocardial infarction in the German population (Broeckel U. et al., 2002, Nature Genetics 30:210–214), respectively. The linkage analysis was carried out using the new Haseman-Elston regression. No significant linkage evidence was observed for either marker in our Gene Quest population. We conclude that PCAD has an important heritable component. However, the susceptibility genes in the Gene Quest population may be different from those in the Finnish and German populations.

IGES-115

Detection and Characterization of Gene-Gene Interactions in the Presence of Missing Genotypes

M.D. Ritchie, L.W. Hahn, J.H. Moore
Program in Human Genetics, Vanderbilt Univ., USA

We have previously developed the multifactor dimensionality reduction (MDR) method to identify gene-gene interactions. In brief, MDR is a method that reduces the dimensionality of multilocus genotypes to identify combinations of SNPs associated with increased risk of disease. This approach develops a multilocus model for defining disease risk by pooling high-risk genotype combinations into one group and low-risk combinations into another. Ten-fold cross validation and permutation testing are used to identify optimal models. A limitation of MDR is its inability to consider observations with missing genotypes. The goal of this study was to develop a modified MDR method that can model interactions in the presence of missing data without significant loss of power. Here, we modified MDR by allowing a fourth level to each variable that is used when that particular genotype is missing. Thus, in addition to the genotype encoding of 0, 1, and 2, we introduce a fourth encoding of 3 for the missing genotype. This allows MDR to make use of the information contained in genotypes at other loci for that observation. For example, assume a three-locus interaction is being modeled and a single genotype at one locus for one observation is missing. MDR will now be able to use the genotype information at the remaining two loci whereas before it would remove that observation from the dataset. In this example, it is anticipated that the power to identify the three-locus interaction will be improved if the other two-loci have some level of effect that is independent of the third locus. Using simulated interactions among two to five SNPs case-control data we validate the MDR missing data approach and demonstrate improved power to identify functional SNPs. We

anticipate this new addition to the MDR method will increase its usefulness for identifying gene-gene interactions when genotypes are missing from multilocus SNP data.

IGES-116

Genome scan for obesity genes in type 2 diabetes patients from West Africa

C. Rotimi[1], G. Chen[1], J. Zhou[1], A. Amoah[2], K. Agyenim-Boateng[3], B. Eghan Jr.[3], J. Oli[4], F. Abbiyesuku[5], T. Johnson[6], O. Fasimade[6], Y. Chen[1], H. Daniel[1], G. Dunston[1], F. Collins[7]
[1]National Human Genome Center, Howard University, Washington, DC, USA; [2]University of Ghana, Accra; [3]University of Science and Technology, Ghana; [4]University of Nigeria, Enugu; [5]UCH, Ibadan, Nigeria; [6]University of Lagos, Nigeria; [7]NHGRI, NIH, Bethesda

We conducted a genome search, using GENEHUNTER, for genes to multiple obesity traits (BMI, fat mass -FM, percent fat mass PFM) in a cohort of diabetes patients from West Africa. A total of 341 families including 691 diabetes cases were studied. The strongest evidence of linkage was obtained for FM (LOD = 4.0, P = 0.000009) between D2S2739 and D2S441 followed by a LOD score of 3.3 (P = 0.000049) for PFM between markers D2S2739 and D2S441 on chromosome 2. Strong linkage signals were also obtained for BMI on chromosome 4 (LOD = 3.4; P = 0.000038) between markers D4S1647 and D4S2623; and for FM on chromosome 5 (LOD = 3.5; P = 0.00003) near marker D5S1725. All analyses were adjusted for gender and age. The Quantitative transmission/disequilibrium test (QTDT) demonstrated evidence of association between markers and obesity traits in these linkage regions. Interestingly, our linkage signals lie in regions that have been implicated as harboring obesity susceptibility genes in several previous studies conducted in multiple populations and ethnicity. In conclusion, we found and replicated strong evidence of linkage and association with BMI on chromosome 4 with PFM on chromosome 5 and with FM on chromosomes 2 and 5 in our cohort of diabetes patients.

IGES-117

Application of statistical methods for analyzing collapsibility in regression models to the Regression of Offspring on Mid-Parent (ROMP) approach

M.-H. Roy-Gagnon[1,2], A.J.M. Sorant[1], A.F. Wilson[1]
[1]Inherited Disease Research Branch, NHGRI, NIH, Baltimore, MD, USA; [2]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

The Regression of Offspring on Mid-Parent (ROMP) method is an extension of the traditional linear regression of offspring on mid-parent used to estimate trait heritability. ROMP also provides an estimate of the heritability attributable to a candidate locus by including a locus effect in the model. The change in the regression coefficient of the mid-parent value when the locus effect is added to the model is the locus-specific heritability estimate. ROMP can be set in the general problem of collapsibility in regression analysis. The goal in analyzing collapsibility is to compare the coefficient β_1 from the reduced model $Y = \beta_1 X + e_1$ to the coefficient β_2 from the full model $Y = \beta_2 X + \beta_3 Z + e_2$ using the difference $\delta = \beta_1 - \beta_2$. The covariate Z is said to be collapsible if $\delta = 0$. Applying results from the collapsibility theory to ROMP provides a theoretical standard error for the locus-specific heritability estimate and shows that testing for $\delta = 0$ is equivalent to testing for a significant locus effect, i.e. $\beta_3 = 0$. Standard errors calculated in simulated samples of parent-offspring trios with locus-specific heritabilities ranging from 10% to 50% were similar to, but smaller than their corresponding Monte Carlo estimates. Average lengths of 95% confidence intervals ranged from 0.01 for a true heritability of 0 to 0.1 as heritability increases. Nonparametric bootstrap and permutation methods are also considered.

IGES-118

Power and sample size calculations for studies of gene-gene and gene-environment interactions

C.L. Saunders, D.T. Bishop, J.H. Barrett
Genetic Epidemiology Division, Cancer Research UK Clinical Centre, University of Leeds, UK

We apply a simple power calculation method (1) to case-control studies of gene-gene and gene-environment interactions. For any proposed design, power or required sample size to detect interaction can be estimated if the expected frequencies among cases and controls of the risk factors under study under the alternative hypothesis can be determined. It is assumed that the likelihood ratio test statistic for the interaction term is distributed as a non central chi squared distribution under the alternative hypothesis and the test statistic from the analysis of a data set with expected frequencies is an approximation to the non-centrality parameter of this distribution. We have also carried out simulations for power using the same expected frequencies and show that the results of the two methods are very similar, indicating the reliability of this large sample approximation. We have applied this method to family, population-based and flexible- and counter-matching study designs for interactions, and show that power varies with population and sampling parameter values in a non-uniform manner between designs. The application of these calculations at the planning stage of a study is therefore useful, since there is no one sampling scheme that can be systematically identified as more powerful. Calculations using the non-central chi-squared distribution can generally be carried out using standard statistical software such as SAS or Stata, which makes the wide application of these

under-used methods straightforward. (1) Longmate JA (2001) Am J Hum Genet 68:1229–37.

IGES-119

Analysis of two candidate genes for obesity and their implications for the design of genome-wide searches for linkage

H. Schäfer[1], F. Geller[1], A. Dempfle[1], A. Wermter[2], A. Hinney[2], J. Hebebrand[2]
[1]Institute for Medical Biometry & Epidemiology, Philipps-Univ. Marburg, Germany; [2]Clinic for Child & Adolescent Psychiatry, Philipps-Univ. Marburg, Germany

We present asymptotic variance formulae for the maximum likelihood estimators of genotype relative risks, allele frequencies and attributable risks in the case parent trio design (Schaid and Sommer, 1993; Knapp, Wassmer and Baur, 1995). With these formulae, confidence intervals can be calculated with a simple pocket calculator or spreadsheet program without need for a statistical package. We apply these estimators to the analysis of a frequent SNP in a candidate gene for extreme obesity. We will also investigate the following question: What is the potential of a whole genome scan for linkage to detect the regions of candidate genes like those which we have analysed. We discuss the implications for the design of whole genome scan studies in affected sib pairs for obesity and other complex diseases.

IGES-120

No evidence of linkage for age-related maculopathy on chromosome 1 markers near the ABCR gene

J.H. Schick[1], S.K. Iyengar[1], K. Reading[1], R. Liptak[1], C. Millard[1], K. Lee[2], S. Tomany[2], R. Klein[2], R.C. Elston[1], B.E. Klein[2]
[1]Dept. of Epi. and Biostat. Case Western Res. Univ., Cleveland, OH, USA; [2]Dept. of Ophthal. & Vis. Sci.Univ. of Wis. Med. School, Madison, WI, USA

Age-related maculopathy (ARM) is a multifactorial disorder involving the retinal pigment epithelium, choriocapillaris and retina which primarily but not exclusively affects the macular region of the eye leading to the loss of peripheral central vision. ARM is the leading cause of irreversible loss of vision in the elderly population of western countries. The ABCR gene, whose location has been refined to a 2cM interval between polymorphic markers D1S406 and D1S236 on chromosome 1p22-p21, has been associated with an increased risk of this disease. We tested the hypothesis that this candidate gene region also segregated in ARM families from a community sample from Beaver Dam, Wisconsin. We genotyped 98 families ($N = 325$), which included 240 informative sib pairs. Relationships were reclassified using 376 autosomal markers in a genome-wide scan for ARM. Our quantitative ARM trait was adjusted for age and age[2] and analyzed using S.A.G.E. version 4.2. Neither the results from our multipoint nor our single-point analyses verified any significant major gene effect for ARM on chromosome 1 markers, including the ABCR gene. We conclude that the ABCR gene does not have a major effect on ARM pathophysiology in this population. This study is supported in part by U. S. P. H. S. research grants GM28356 and EY10605, resource grant RR03655 and training grant HL07567.

IGES-121

A case-control study of ovarian cancer susceptibility to CYP17 and androgen receptor gene polymorphisms

J. Schildkraut, R. Wenham, J. Lancaster, B. Calingaert, S. Halabi, J. Marks, K. McLean, A. Berchuck
Duke University Medical Center, USA

Several lines of evidence suggest that androgen may play a role in the etiology of ovarian cancer. The aim of this study is to determine whether polymorphisms in the CYP17 or the androgen receptor (AR) gene, two genes in the androgen pathway, affect susceptibility to ovarian cancer (OC). The identification of polymorphisms that increase OC risk could facilitate identification of high-risk women who would be candidates for screening and/or prevention interventions designed to decrease mortality. In the North Carolina Ovarian Cancer study, an ongoing population-based study, we have accrued 400 epithelial OC cases and 450 controls frequency matched on race and age. Odds ratios (ORs) were computed to assess associations between genotypes and case-control status using logistic regression adjusting for age, race, parity, body mass index, and duration of oral contraceptive use. For the CYP17 polymorphism, ORs were 0.7 (95% confidence interval (CI) = 0.4–1.0) for heterozygotes and 0.8 (95% CI = 0.5–1.5) for homozygotes. For the CAG polymorphism in the AR gene, there was no difference in the mean allele repeats between cases and controls. Additionally the frequency of either very long (>27) or a very short CAG alleles (<16) did not differ between cases and controls. There was no evidence of an interaction between polymorphisms in CYP17 and AR. This study is not supportive of the hypothesis that polymorphisms in AR or CYP17 affect OC risk. We will examine these genes to determine if they modify associations between known OC risk factors.

IGES-122

Association of vitamin D receptor variants and prostate cancer

F.R. Schumacher[1], M. Cicek[2], G. Casey[2], and J.S. Witte[1]
[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA; [2]Department of Cancer Biology,

Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, Ohio, USA

Despite the high incidence and mortality of prostate cancer, few well-established risk factors exist for this disease. Genetic and environmental factors, as well as their interactions with each other, drive the complex etiology of prostate cancer. Previous studies suggest that vitamin D may play a role in cancer by regulating cellular growth and differentiation. Prostatic cells express the vitamin D receptor (VDR), and the VDR mediates the function of 1, 25-dihydroxyvitamin D3, the bioactive form of vitamin D. Therefore, we evaluate the association between four VDR gene variants and prostate cancer risk and aggressiveness in a moderately large family-based case-control study (611 cases, 456 sibling controls). Using an age-adjusted matched analysis, we observed an inverse association between the BsmI variant and low aggressive prostate cancer among all study subjects (OR = 0.58, 95% CI = 0.34, 0.99). In addition, we will present results from evaluating the interaction between the VDR variants and dietary sources of vitamin D. Our findings suggest the BsmI variant protects against latent forms of prostate cancer.

IGES-123

Bayesian TDT analysis allowing for genotyping errors

S.R. Seaman, C. Berzuini, L. Bernardinelli, P. Holmans
MRC Biostatistics Unit, Cambridge, UK

Genotyping errors are known to reduce power to detect associations between disease and marker loci, and also to bias the estimates of genotypic relative risks. In many situations, genotyping errors will be difficult to detect. Therefore, methods to perform TDT analyses allowing for undetected genotyping errors are of considerable interest. Bayesian analysis provides a natural framework for taking uncertainty in genotypic error probabilities into account. We present two methods for performing a TDT analysis with a biallelic marker on parent-offspring trios. The first method assumes a multiplicative disease model and Hardy-Weinberg equilibrium, giving a likelihood equivalent to an unmatched case-control study, with the two transmitted alleles as ''cases'', and the two untransmitted alleles as ''controls'' (Clayton 1999, AJHG 65:1170–1177). The second method makes no assumptions about the parental genotype frequencies, or the disease model, and uses a similar likelihood formulation to that of Weinber et al. (1998, AJHG 62:969–978). The performances of the two methods were compared to each other, and also to that of a standard TDT test assuming no genotyping errors, under a variety of disease and error models. It was found that the Bayesian analyses gave credible intervals for the genotypic relative risks with approximately correct coverage probabilities, and, in the absence of errors similar power to a standard TDT analysis.

IGES-124

Is the choice of the MLS or NPL statistic in an affected sib-pair linkage analysis, equivalent?

H. Selinger-Leneman[1], M.C. Babron[1], F. Clerget-Darpoux[1]
[1]INSERM U535, Le Kremlin Bicêtre, France

When searching for susceptibility genes in complex diseases, by a systematic search of linkage in the whole genome, linkage analysis is often performed on affected sib pair samples. MLS [Risch, 1990] and NPL [Kruglyak et al, 1996] are widely used for testing linkage. In this paper, we address the question: can the conclusion of the search depend on the choice between the two statistics. We first assessed for both statistics, their respective thresholds, for the marker map considered and a 5% type I error, by simulating under the null hypothesis of no susceptibility gene. Then we compared their power by simulations under several H1 hypotheses (i.e. assuming several genetic models for the susceptibility gene). Relative power of MLS and NPL are shown to be model dependant. Besides, there is no simple correspondence between the MLS and the NPL. Different samples with a same MLS value may have different NPL values and conversely. In particular, a large proportion of samples for which the NPL values are over the threshold, does not reach significance with MLS and conversely. The choice of one or the other of these two statistics could thus change the linkage analysis conclusion. The choice of the statistic may also have an impact in the conclusions of meta-analyses, such as GSMA [Wise et al, 1999].

IGES-125

APOE gene polymorphism and type 1 diabetic complication: genotype-phenotype associations study

N.S. Shcherbak
Laboratory of Molecular Cardiology, St. Petersburg State Medical University, Russia

Vascular complications of diabetes are a major cause of morbidity and mortality, but the mechanisms for their development remain elusive. Neuropathy is a common complication of type 1 diabetes mellitus. These conditions usually result from diabetic microvascular injury involving small blood vessels that supply nerves (vasa nervorum). A number of facts suggest that diabetic neuropathy may involve genetic susceptibility. Recently, it is reported that APOE polymorphism may influence to development of diabetic neuropathy. The gene encoding apolipoprotein E (APOE) has been proposed as a candidate gene for vascular complication in type 1 diabetes. Apolipoprotein E was discovered as a plasma protein involved in the metabolism of lipoproteins. There are three common alleles, E2, E3, and E4, which code for three major isoforms, resulting in six common genotypes. The aim of this study was to investigate the influence of APOE gene polymorphism in the development of diabetic neuropathy (genotype-phenotype associations

study) in type 1 diabetes patients. The study consists of 51 patients with diabetic neuropathy and 150 without diabetic neuropathy matched to the patients by age gender and diabetes duration. APOE polymorphism was detected by the restriction fragment length polymorphism method after a polymerase chain reaction. The distribution of APOE genotypes and alleles frequency showed no difference between the patients with diabetic neuropathy and without this complication ($p > 0.05$). The difference between the groups was tested by Fisher's exact test. The present study no found the strong genotype-phenotype associations.

IGES-126

Genetic mapping in the presence of gene-environment interactions

D. O. Siegmund[1], H.K. Tang[2]
[1]Dept. of Statistics, Stanford University, USA;
[2]Hewlett Packard, Palo Alto, USA

We discuss a model of gene-environment interactions for both quantitative and qualitative traits. For quantitative traits and a population based sample of sib pairs, we compare the "naïve" score statistic, which neglects the gene-environment interaction and uses a single degree of freedom to detect an additive genetic effect with the true, three degree of freedom, score statistic. Numerical examples show that when the between sibs correlation in the environmental variable is small, (a) the power to detect linkage can be very low and (b) the correct score statistic can be much more powerful than the naive statistic. Since determination of the true score statistic requires detailed specification of a model for the gene-environment interaction, we also investigate a more easily implemented strategy for mapping a qualitative trait which uses a subsample of a sample of affected sibpairs, selected on the basis of their environmental variables.

IGES-127

A comparison of cluster analysis methods on continuous data with an excess of zeros

K.D. Siegmund[1], P.W. Laird[2]
[1]Department of Preventive Medicine;
[2]Departments of Surgery and Biochemistry & Molecular Biology, Keck School of Medicine, University of Southern California, Los Angeles, USA

We compare cluster analysis methods for the discovery of novel disease subtypes based on DNA methylation profiles. The distribution of DNA methylation measured using the MethyLight technology is a mixture of discrete and continuous observations; MethyLight measures the frequency of fully methylated alleles for any region of DNA, finding none for many samples and variable levels for others. In a simulation study, we compare standard model-based clustering methods for continuous data and for discrete data when the outcome is from a mixture distribution. Data are simulated for two groups under a variety of outcome distributions. In general, the misclassification rates for all models increase as the proportion of zeros in the dataset increases. The discrete and continuous data methods perform similarly when the frequency of zeros does not differ between the two subgroups. When the proportion of zeros is higher in the group with the higher positive methylation values, the error rate is lower for the discrete data approach. When the proportion of zeros is higher in the group with the lower positive methylation values, the error rate is lower for the continuous data approach. We find the approaches perform equally well when applied to data from a study of small-cell and non-small cell lung cancer cell lines (Virmani et al., CEBP 11:291–297, 2002), having misclassification error rates around 20% and cross-validation error rates around 25%.

IGES-128

Complex simulation using G.A.S.P.

A. J. M. Sorant, A. F. Wilson
Inherited Disease Research Branch, NHGRI, NIH, Baltimore, MD, USA

The Genometric Analysis Simulation Program (G.A.S.P.) is a software tool for generating samples of family data based on a user-specified generating model including major locus, polygenic, common sibship environment, and/or covariate components for quantitative traits. Discrete traits or markers can also be simulated. Realistic complications such as missing data, genotyping errors, reduced penetrance, genetic heterogeneity, epistasis, and non-random ascertainment of samples can be added with modifications to the G.A.S.P. driver. Driver-based trait modifications include: functional transformation classification of a quantitative trait according to a threshold, and randomly determined expression of a discrete trait (reduced penetrance). More complicated trait models allow for the contribution of one locus to depend on the genotype at another locus. Non-random ascertainment can be modeled by discarding families which do not meet the desired criterion. The driver can allow for mixtures of populations, modeling partial association or genetic heterogeneity, by maintaining separate model specifications and random number streams for each population and including families from different populations in the desired proportions. Most simulation studies require the use of thousands of samples from each population model considered. It is necessary to generate data invoke analysis programs, extract relevant output, and accumulate results for each replicate. A unix shell script is used in combination with G.A.S.P. and appropriate driver programs to accomplish these tasks.

IGES-129

A comparison of efficient genotype samplers for complex pedigrees and multiple linked loci

C. Stricker[1], M. Schelling[2], F. Du[3], I. Hoeschele[3], S.A. Fernández[4], and R.L. Fernando[4]
[1]Applied Genetics Network, Schweigrutistrasse 20, 8852 Altendorf, Switzerland; [2]Insitute of Animal Sciences, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland; [3]Virginia Tech, Department of Dairy Science, Blacksburg, VA, USA; [4]Iowa State University, Department of Animal Science, Ames, IA, USA

First, the limitations of maximum likelihood and the sampling of genotypes or allelic origins (segregation indicators) by a Gibbs sampler are discussed. Then, different 'non-Gibbs' Markov chain Monte Carlo strategies are presented and compared: sampling genotypes by a blocking strategy where the block size is equal to the whole pedigree (the ESIP sampler of Fernández et al., 2002a, 2002b) or sampling segregation indicators instead. For the latter the approaches of Du and Hoeschele (2002) updating blocks of segregation indicators jointly and those of Schelling (2002) extending the sampling of allelic origins over multiple linked loci are presented and compared.

IGES-130

A likelihood ratio approach to determining sample size in genetic studies

L.J. Strug[1,2], C.A. Rohde[3], M. Corey[1,2]
[1]Dept of PHS; [2]Hospital For Sick Children, University of Toronto, Canada; [3]Dept of Biostat, Johns Hopkins University, USA

Statistical Genetics is one of the few fields to embrace the Law of Likelihood; evaluating statistical evidence using lod scores is the gold standard in linkage studies. But why stop there? Like mainstream medical research, the genetic literature is rife with studies that prove difficult to replicate. Cardon and Bell (Nature Rev Gen, 2, 2001, 91–99) discuss contributing factors in genetic association studies, including poor design and insufficient sample sizes. A sample size method is proposed that uses the concepts of weak (w) and misleading evidence (m) (Royall, 1997, Chapman and Hall), defined using likelihood ratios, as an alternative to power and significance based methods. It produces 4 plots displaying the probability of w and m for increasing sample size, under the two competing hypotheses for an effect. Sample size decisions are generally driven by w probabilities which allow for control of ambiguous results. An illustration of this method for a genetic association study is presented. It provides sample size decisions to detect a cross-product ratio (CPR), allowing for different specifications of Mendelian transmission. A review of a study by Dunning et al. (Human Molec Gen, 6, 1997, 285–289) exhibited w probabilities of 0.609 and 0.613 (CPR = 1.251) for its size under a dominance hypothesis. The original study concluded insufficient evidence of an association. The 4-plot method leads to larger sample size decisions than standard methodology. 4-plot methods for linkage studies and genome scans will be considered in future work.

IGES-131

Approximating linkage disequilibrium patterns in the HLA region of chromosome 6 as a covariance function

M. Swartz[1,3], M. Kimmel[1], P. Mueller[2], C.I. Amos[3]
[1]Dept. of Statistics, Rice University, USA; [2]Dept. of Biostatistics, University of Texas, M. D. Anderson Cancer Center, USA; [3]Dept. of Epidemiology, University of Texas, M. D. Anderson Cancer Center, USA

Using linkage disequilibrium (LD) to induce a covariance structure in a hierarchical model may help smooth out noise encountered in analysis of a complex disease. However, the common measures for LD can be noisy and may not produce a positive definite matrix necessary for a covariance measure. LD in the human leukocyte antigen (HLA) region of Chromosome 6 decays as a function of genetic distance, similar to covariance structures commonly encountered in spatial statistics. Therefore, we chose a rational quadratic semi-variogram, commonly used in spatial statistics, to model LD. Fitting the semi-variogram to the LD measures between markers from the HLA region of Chromosome 6 gives us approximate LD measures between markers that also generate a positive definite matrix. Visual inspection of the predicted LD values shows that the rational quadratic semi-variogram captures many of the features of the real data, however, the semi-variogram also has an upward bias for LD between more distant markers. We use a Chi-square test to evaluate the fit of the semi-variogram.

IGES-132

Regional inference procedure using smoothing techniques for genome-wide association studies

N. Tanaka[1], T. Yamaguchi[1]
[1]Dept of Biostat, the Univ. Tokyo, Japan

In recent years, genome-wide association studies are feasible through use of PCR methodologies with pooled DNA samples and microsatellite variation, and SNP variation. Ordinarily, we assess that there is *significant* evidence for association between marker and disease from *p-values* based on the statistical testing. However, those studies are being conducted on many complex diseases where the susceptibility genes involved are likely to have moderate to small effects, requiring investigation of large numbers of samples. In such situations, we would like to know not only whether the marker is associated with disease but also the strength of association. In case-control studies, odds ratio is typically used as a measure of the strength of association. Instead of plotting the odds ratios estimated for each marker from the case-control samples, using smoothing technique, we propose to compute and plot the weighted average of odds ratio across the number of markers at the informative markers. The estimated correlation coefficient was

used as weight. Moreover, 95% confidence intervals for each marker and band for smoothed curves are calculated using resampling method to adjust multiplicity. We consider several situations and present some simulation examples. Our method provided that association between marker and disease can be assessed adjusting multiplicity and correlation between markers and visualization the most susceptible genomic region as well as the power to detect the difference.

IGES-133

Do HLA DRB1 and TNFα play both a role in Rheumatoid Arthritis

S. Tezenas du Montcel[1,2], E. Teixeira[3], A. Mallet[2,4], J. Osorio[3], F. Cornelis[3], F. Clerget-Darpoux[1]
[1]INSERM U535, Kremlin-Bicetre, France; [2]Service d'Informatique Medicale, CHU Pitie Salpetriere, AP-HP, France; [3]GenHotel, Evry, France; [4]INSERM U436, Paris, France

Rheumatoid arthritis (RA) is a frequent systemic autoimmune disease of unknown cause. However, RA is a multifactorial disease with contributions from both genetic and environmental factors. To date, HLA DRB1 is considered as a susceptibility factor for RA. Although controversial, susceptibility is based on a common epitope shared by some DRB1 alleles that increases the risk to develop RA. Other HLA genes are also suspected to be involved in RA development, in particular the tumor necrosis factor α gene (TNFα) located in the class III region of HLA. TNFα is one of the central cytokines involved in the inflammatory process and anti-TNFα have proved to be efficient in RA. The aim of the study is to refine the HLA modelling of RA using DRB1 and TNFa, a microsatellite of TNFα, in two kinds of French samples: 95 trios (one affected and his two parents) and 132 affected sibs pairs. The MASC method is used to revisit the shared epitope hypothesis and a log linear analysis is performed to test if TNFα contributes to the risk of developing RA additionally to the risk due to DRB1.

IGES-134

Validity of two-stage QTDT for small samples

H.K. Tiwari[1], V. George[1], R.C. Elston[2]
[1]University of Alabama, Birmingham, AL, USA; [2]Case Western Reserve University, Cleveland, OH, USA

George, et al. (Am J Hum Genet 65:236, 1999) proposed a regression-based TDT for linkage between a marker locus and a quantitative trait locus by modeling the trait as the dependent variable and the transmission status as one of the independent variables along with other predictors and confounders in a linear regression model. We 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. If a significant association is found at the first stage, then a regression based QTDT is performed (George & Tiwari, Genet Epidemiol 17:192, 1999). We obtained a good large sample regression model to estimate the overall p-value associated with this two-stage test procedure. In this presentation, we investigate to what extent this large sample regression model can be used to estimate appropriate p-values for the two-stage procedure in small samples.

IGES-135

Variance component models for the analysis of longitudinal family data

M.D. Tobin, K.J. Scurrah, N.A. Sheehan, P.R. Burton
Genetic Epidemiology and Integrated Social Science, Institute of Genetics, University of Leicester, UK

Many longitudinal studies of nuclear and extended families are underway to investigate the etiological architecture of complex traits. These studies are powerful and informative because they reduce measurement error and allow the study of the determinants of trends in a trait over time. This is important if we are to properly understand the genetic and non-genetic influences underpinning the natural history of complex diseases. However, the analysis of such data is more complex than that of data from cross-sectional family studies. We describe extensions of our standard Gibbs sampling-based variance component models (eg Scurrah et al, 2000) that correctly model the repeated measurements and permit the investigation of the etiological determinants (both observed and latent, genetic and non-genetic) of trends in a complex trait. The models can also be used for both linkage and association analyses. Although the extensions we describe involve many additional parameters, the ease with which Gibbs sampling models can be extended to new problems renders them relatively straightforward. We will present the relevant theory and illustrate its application to data based on both nuclear and extended families. Like our standard models, these methods are applicable to phenotypes with any distribution from the exponential family including: continuous normally distributed traits, binary phenotypes and censored traits. These models have wide applicability in biomedical science. Scurrah KJ, Palmer LJ, Burton PR. Genetic Epidemiology 2000;19:127–141.

IGES-136

Haplotype association analysis by use of the stochastic-EM algorithm

D.A. Trégonët, L.Tiret, S. Escolano, J.L. Golmard
INSERM U525, France

It is now widely admitted that haplotypic information can be of potential interest for investigating

the role of a candidate gene in the etiology of complex diseases. In absence of family data, haplotypes are not readily deducible from genotypes, except for individuals who are homozygous or heterozygous at only one site. New statistical inference methods need then to be developed for estimating from genotypic data haplotype frequencies and their corresponding haplotypes effects on the phenotype of interest. Up to very recently, Expected-Maximisation (EM) algorithm was the standard technique for estimating haplotype frequencies but was mainly applied for testing difference in haplotypes frequencies distribution between cases and controls. We here describe how a stochastic version of the EM algorithm (SEM) can be used for testing association between haplotypes and any quantitative or qualitative phenotype. The technique is general enough for incorporating information on additional covariates and investigating haplotypes × covariates interaction. Statistical properties of the SEM algorithm are investigated through a simulation study in a large range of practical situations including small/large samples and rare haplotype frequencies. Results were compared to those obtained by use of the standard maximum likelihood approach based on the Newton-Raphson algorithm.

IGES-137

Comparisons of power for three model-based tests for linkage

A. Ulgen[1], N.R. Mendell[2], S.J. Finch[2], D. Gordon[3]
[1]G.H. Sergievsky Center, Columbia Univ, USA; [2]Dept. of Applied Math and Stat, SUNY at Stony Brook, USA; [3]Lab. of Statistical Genetics, Rockefeller Univ, USA

We conducted a power study comparing three statistics: the lod score statistic assuming the operative genetic model (LOD-C); the MMLS procedure of Hodge et al. (MMLS), which maximizes lod score over twenty fixed models for twenty-six fixed recombination fractions (ρ values); and the LOD-M procedure (Ulgen et al., 2002), which maximizes lod score over a five dimensional model space. We used simulation to study three genetic models with $\rho = 0.02$: one dominant and two intermediate. We used 1200 replicates of a study with 100 nuclear families. Both MMLS and LOD-M had relatively small loss of power compared to LOD-C. The MMLS and LOD-M procedures had essentially equal power for the alternatives studied. Multipoint analysis lowered power of all three procedures. The effect on LOD-C and LOD-M was minor; the effect on MMLS was larger. Consequently, we recommend LOD-M because one searches the complete parameter space to find the model that best describes the data and overall level of significance is controlled with little apparent loss of power.

IGES-138

Are mutations in the HFE gene associated to disease?

C.M. Van Duijn[1], B.Z. Alizadeh[1], F. Elsheikh[1], O.T. Njajou[1], A. Hofman[1], H.A.P. Pols[2], J.W.C. Witteman[1]
[1]Department of Epidemiology & Biostatistics; [2]Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands

Two mutations in the HFE gene (C282Y and H63D) explain up to 85% of all patients with hereditary hemochromatosis. Although this is a lethal disease if not treated, recent studies suggested carriers of these mutations are not at increased risk of disease (The Lancet 2002; 359:211–18). Both HFE mutations and smoking may be involved in cardio and cerebrovascular diseases through the same mechanism, i.e. oxidative stress. We studied the association of these mutations with the determinants of atherosclerosis, myocardial infarction and stroke in over 3000 subjects derived from a population-based follow up study, The Rotterdam Study. The effect of the mutations was studied overall and in interaction with smoking. HFE mutations significantly modified the relation between smoking and determinants of cardio and cerebrovascular disorders i.e. hypertension ($P = 0.01$), atherosclerosis ($P = 0.001$), and angina pectoris ($P = 0.01$). For each factor, smoking associated risk in carriers was more than twice increased than the risk in those without mutations. The association between smoking and myocardial infarction and stroke was also modified by HFE mutation ($P = 0.01$). All the modifications were found in both heterozygous and homozygous carriers. However the effect of the mutations in the absence of smoking was not significant. Heterozygosity and homozygosity for HFE C282Y and H63D mutations are determinants of atherosclerosis, cerebro and cardiovascular disorders through the modification of smoking associated risks.

IGES-139

Linkage disequilibrium haplotype diversity and association with asthma in the ADAM33 gene

P. Van Eerdewegh[1], J. Dupuis[1], K. Falls[1], R. Little[1], B. Hayward[1], A. Bureau[1], R. Del Mastro[1], F. M. Cuss[2], S. T. Holgate[3], T. Keith[1]
[1]Genome Therapeutics Corp.USA; [2]Schering-Plough Research Inst.USA; [3]U. of Southampton General Hospital, UK

We performed a genome wide scan on 460 Caucasian families and identified a locus on chromosome 20p13 linked to asthma (MLS = 2.94) and asthma with bronchial hyperresponsiveness (MLS = 3.93). We tested 135 polymorphisms in 23 genes spanning a 1-LOD support interval and identified ADAM33 as being associated with both phenotypes by case-control and transmission disequilibrium (TDT) tests ($p = 0.04$ to $p = 0.000003$). Statistical haplotyping of 80 SNPs in 7 genes

spanning 247 kb around ADAM33 was performed to elucidate the haplotype diversity and block structure of LD in that region of chromosome 20. The low diversity illustrated by the small number of haplotypes with an estimated frequency of 1% or larger is in contrast to the more complex picture found with the measure of LD (Δ) that is relevant to the identification of a causal variant within a gene in a case-control association study. The pattern of Δ differed between the US and UK populations, and SNPs in strong LD by this measure formed groups of non-adjacent SNPs with similar level of association with the asthma phenotype. Detection of an association is critically dependent on both the allele frequency and the non-monotonic nature of LD as a function of distance. This is illustrated by measuring the pairwise LD among the 37 SNPs typed in ADAM33 and the association of specific SNPs and haplotypes with the asthma phenotype in this study.

IGES-140

Merits of the multivariate dale model in genetic association studies

K. Van Steen[1], G. Molenberghs[1], N. Tahri[2,3]
[1]Center for Statistics, LUC, Diepenbeek, Belgium; [2]INSERM U.525, C.H.U. Pitié Salpétrière, Paris, France; [3]Genset Genomics Research Center, Evry, France

Until now, the most common parametric approaches to study the combined effects of several alleles at different positions are logistic regressions at the diploid level and haplotype-based methods at the haploid level. However, in particular settings (e.g. if multicollinearity between the covariates of interest is observed, or if the effects of several alleles include dominance phenomena), other methods may be more attractive. Alternatively, truly multivariate models can be considered. One such model is the multivariate Dale model (Molenberghs and Lesaffre 1994 JASA; 89: 633–644), which encompasses a whole family of parametric models, by the choice of different link functions for the margins and/or associations. Interpretations remain straightforward and similar to those drawn in a classical logistic modeling framework. Perhaps the most interesting merit of the model is that associations (expressed by means of odds ratios) between markers can be linked to relevant covariate information apart from disease aspects. Row-, column-, and cell-specific terms can be included in a flexible way. Apart from pointing out the benefits and drawbacks of the multivariate Dale model we will apply the model to a real-life data set of case-control individuals genotyped on SNPs within a 250kb region containing the APOE gene. In addition, we will address extensions of the model that are similar in spirit.

IGES-141

Further evaluation of an autism candidate gene on 7q31-33 via linkage disequilibrium

D. Wang[1], T.H. Wassink[2], J. Huang[1,3], J. Pietila[4], V.C. Sheffield[4], R. J. Goedken[1], V.J. Vieland[1,2], J. Piven[5]
[1]Dept. of Biostatistics Div. Of Statistical Genetics, Univ of Iowa, USA; [2]Dept. of Psychiatry, Univ. of Iowa, USA; [3]Dept. of Statistics and Actuarial Science, Univ. of Iowa, USA; [4]Dept. of Pediatrics and the Howard Hughes Medical Institute, Univ. of Iowa, USA; [5]Neurodevelopmental Disorders Research Center and Dept. of Psychiatry, Univ. of North Carolina, USA

Following up on an initial report of linkage disequilibrium (LD) to WNT2 [Wassink et al., 2002], we genotyped two additional SNPs, EX4A and IN3b5b in 75 affected sib-pair (ASP) families. The dataset was also split into a ''language abnormal'' subset with 50 families in which both autistic children had substantial language delay, and a ''language normal'' subset of the remaining 25 families. Two novel data analytic methods (LD-HET, allowing for locus heterogeneity; and a simple multinomial test) were used, as well as the TDT. Because our earlier analyses had suggested that both linkage and LD effects occurred only in the language-abnormal group, and also because there was evidence of locus heterogeneity within that group, we were interested in comparing the performance of these statistics in our data set. The LD-Het model detected significant LD at IN3b5b locus ($p = 0.0199$) for the pooled dataset ($N = 75$) but not in the subsets; the multinomial and TDT tests found significant LD at IN3b5b locus in the language-abnormal subset ($p = 0.002$; $p = 0.01$ respectively) but not in the whole dataset. No significant LD was found at EX4A locus. These results are consistent with earlier findings.

IGES-142

Efficient score statistics for mapping quantitative trait loci using multiple phenotypes

K. Wang
Div. of Statistical Genetics, Dept. of Biostatistics, Univ. of Iowa, USA

The use of correlated phenotypes may increase the power of detecting the underlying quantitative trait loci (QTLs). Current approaches that utilize multiple phenotypes include the regression method which generalizes the Haseman-Elston method, the variance of components method and the structural equations method. All these methods fail to recognize one important phenomenon when using multiple phenotypes, i.e. some model parameters are confounded under the null hypothesis that the putative locus is not linked to any QTL. Such confounding makes the null asymptotic distribution of the likelihood ratio statistic in the latter two methods more subtle and more complicated than usually expected. I propose a likelihood function

which is similar to, but different from, the likelihood function of the method of variance components. The problem of parameter confounding is overcome by making various plausible model assumptions. Most importantly, efficient score statistic, which is equivalent to the likelihood ratio statistic, and its asymptotic distribution are derived for each of the situations considered. The finite-sample property of these score statistics are studied using simulations. These score statistics are for use with general pedigrees. They are easier to compute than any of the three existing methods.

IGES-143

On the use of DNA pooling to estimate haplotype frequencies

S. Wang[1], K.K. Kidd[1,2], H. Zhao[1,2]
[1]Dept. of Epi and Public Health, Yale Univ., USA; [2]Dept. of Genetics, Yale Univ., USA

It may be necessary to conduct genome wide screens to identify genes underlying certain complex diseases. Because such studies can be extremely expensive, DNA pooling has been introduced as it may greatly reduce the genotyping burden. Parallel to DNA pooling developments, the importance of haplotypes in genetic studies has been amply demonstrated in the literature. However, DNA pooling of a large number of samples may lose haplotype information among tightly linked genetic markers. As a result, the usefulness of the DNA pooling strategy in haplotype studies has never been explored in the literature. In this study, we examine the cost effectiveness of DNA pooling in the estimation of haplotype frequencies from population data. When the maximum likelihood estimates of haplotype frequencies are obtained from pooled samples, we compare the overall cost of the study, including both DNA collection and marker genotyping, between the individual genotyping strategy and the DNA pooling strategy. We find that DNA pooling of two individuals can be more cost effective than individual genotypings, especially when a large number of haplotype systems are studied.

IGES-144

Routine discovery of high-order epistasis models for simulation studies in human genetics

B.C. White, L.W. Hahn, M.D. Ritchie, T.A. Thornton, J.H. Moore
Program in Human Genetics, Vanderbilt Univ., USA

Simulation studies are useful in various disciplines for a number of reasons including the development and evaluation of new computational and statistical methods. This is particularly true in human genetics and genetic epidemiology where new analytical methods are needed for the detection and characterization of disease susceptibility genes whose effects are complex, nonlinear, and partially or solely dependent on the effects of other genes (i.e. epistasis or gene-gene interaction). Despite this need, the development of complex genetic models that can be used to simulate data is not always intuitive. In fact, only a few such models have been published. We have previously developed a genetic algorithm approach to discovering complex genetic models in which two single nucleotide polymorphisms (SNPs) influence disease risk solely through nonadditive interactions. In this paper, we extend this approach for the discovery of high-order epistasis models involving three to five SNPs. We demonstrate that the genetic algorithm is capable of routinely discovering interesting high-order epistasis models in which each SNP influences disease risk only through interactions with the other SNPs in the model. This study opens the door for routine simulation of complex gene-gene interactions among SNPs for the development and evaluation of new statistical and computational approaches for identifying common complex multifactorial disease susceptibility genes.

IGES-145

Family- and marker-specific empirically derived type I critical values for genomic screens

A.F. Wilson, G.J. Papanicolaou
Inherited Disease Research Branch, NHGRI, NIH, Baltimore, MD, USA

The use of theoretically derived critical values in testing for significance in genomic screening data are constrained by the relatively simple pedigree structures, markers and simplifying assumptions necessary to make the computations tractable. Conversely, simulation studies can provide critical values based on large numbers of repetitions of specific family structures and markers, but they are computationally intensive. An alternative approach for determining critical values based on actual family structures and markers is presented. Because a genomic screen includes a large number of markers that are not linked to putative trait loci, the empirical distribution of the null hypothesis, either no linkage or no association, can be obtained by using those markers not linked or associated with the specific trait. A marker with a known chromosomal location can be genotyped and used as a "pseudo-trait" in the genomic screen. The resulting lod-score or p-value is used to exclude linked or associated markers, and the null distribution is obtained using the results from the non-linked (or associated) markers. Like simulation methods, the resulting distribution is based on the actual family structures and markers, but requires no additional computation. In addition, by using the actual screening data, experimental errors specific to the analyses are incorporated into the determination of the critical value. Determination of the critical values with this approach and a theoretical approach are quite similar in an analysis of BMI in the Old Order Amish.

IGES-146

Statistical properties of affected sib-pair linkage tests

C.C. Wu, C.I. Amos
UT MD Anderson Cancer Center, Department of Epidemiology

Blackwelder and Elston (1985) investigate analytically the appropriateness and power of several identity by descent (IBD) test statistics by affected sib pairs including the mean test and the two-allele test. The authors refine their investigation on probabilistic behavior for normal approximations to the discrete test statistics, the mean test and the two-allele test, through numerical analyses. Both of them are consistently liberal (the actual significance level greater than nominal significance level) for various sample sizes, particularly for the two-allele test. Useful statistical properties are presented in which the actual null tail probability distributions for them can be approached in excellent agreement. A nonparametric power figure is presented which compares the power dominance between the mean test or the two-allele test.

IGES-147

Combined analysis of genomewide scans for adult height: results from the NHLBI family blood pressure program

X. Wu[1], R.S. Cooper[1], E. Boerwinkle[2], S.T. Turner[3], S. Hunt[4], R. Myers[5], R. Olshen[6], D. Curb[7], X. Zhu[1], D. Kan[1], A. Luke[1]
[1]Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, Maywood, IL, USA; [2]Institute for Molecular Medicine and Human Genetics Center, University of Texas Houston Health Science Center, Houston, TX, USA; [3]Division of Hypertension, Mayo Clinic, Rochester, MN, USA; [4]Cardiovascular Genetics Research Clinic, University of Utah School of Medicine, Salt Lake City, UT, USA; [5]Section of Neurogenetics, Boston University Medical Center, Boston, MA, USA; [6]Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA; [7]Pacific Health Research Institute, Honolulu, HI, USA

We performed a combined analysis of genome scans for adult height on 6,684 individuals from the NHLBI Family Blood Pressure Program. Linkage analysis was first performed separately in each of the eight ethnic groups in the four networks using the variance component method. To increase the power to detect the common genetic components affecting height for all the individuals, a linkage analysis was subsequently perform in the combined data set using Haseman-Elston regression by pooling the average allele sharing IBD for all groups. With this approach we found strong linkage evidence for a QTL at 6q16 (marker D6S1056, LOD = 3.40), which has been previously reported as linked with adult height. Less significant results were found on chromosomes 1, 2, 7, 8, 9, 12, 20 and 21, while evidence for linkage on chromosome 2, 7, 8, 20 has been reported before.

IGES-148

Genetic analysis of longitudinal data with DFREML: Framingham Heart Study

Q. Yang, L.A. Cupples
Department of Biostatistics, Boston University, USA

A unique feature of the Framingham Heart Study is that subjects were repeatedly measured usually every 2 or 4 years for the same cardiovascular related phenotypes and risk factors over a long period of time. In searching for the genetic determination of cardiovascular traits, such data provide an excellent opportunity to study the change of heritability over time as well as that of the influence of quantitative loci over time. We applied the DFREML programs (Meyer K. 1998. Genet. Sel. Evol.) to study the variation of the heritability of total cholesterol with age in Framingham offspring subjects. DFREML is based on a methodology that uses covariance functions to describe the covariance between measurements measured at any two ages. By using orthogonal polynomials as the covariance function, it enables a reduced fit that leaves fewer parameters to estimate than multivariate analysis estimating covariance components. We found that the heritability of total cholesterol in Framingham subjects increases from 0.30 to 0.43 with age before 35 and decreases gradually to 0.19 afterwards by age 75. Simulation studies showed that DFREML can provide reliable estimates of heritability and is superior to the strategy that breaks the data by age into several sets and estimates heritability for each dataset. Further exploration of the usefulness of DFREML in linkage analysis is underway.

IGES-149

CYP1A1 and other single nucleotide polymorphisms in relation to lung cancer risk: a case-control study of women in northeast china

X. Yang[1], S. Wacholder[1], Z-Y. Xu[3], M.Dean[2], V. Clark[2], B. Gold[2], L.M. Brown[1], B. Stone[1], N. Caporaso[1]
[1]Division of Cancer Epidemiology and Genetics, NCI, Bethesda, MD, USA; [2]Laboratory of Genetic Diversity, NCI, USA; [3]Liaoning Public Health and Anti-Epidemic Station, Shenyang, China

To investigate the relationship between polymorphisms of CYP1A1 I462V, GSTM1, and other cancer-related genes and lung cancer risk, we utilized data from a large population-based case-control study carried out in northeast China, where the rate of lung cancer among nonsmoking females is among the highest in the world. 204 female consecutive

cases and 204 controls with genomic DNA available for analysis were selected from the large case-control study. Subjects were genotyped for CYP1A1 I462V, GSTM1, and 46 other cancer-related SNPs. The CYP1A1 I462V genotype (combined ile/val and val/val) was significantly associated with increased lung cancer risk (OR = 2.5, 95% C.I.1.55–4.03) after adjustment for known risk factors including age, ever smoking, family history of cancer, and eye irritation during cooking. Furthermore, there was some evidence that the association was stronger in non-smoking group (OR = 3.67; 95% C.I.1.85–7.28). In contrast, in this study, we did not observe a significant association between GSTM1 null genotype and lung cancer risk. Among 47 SNPs screened in this study, 4 (two involving BRCA1, ardb3 and TGFB1) exhibited some association with lung cancer risk; however, only CYP1A1 met the strict Bonferroni significance criterion after accounting for multiple comparisons.

IGES-150

Effect of non-random selection on statistical power of variance components method for quantitative traits

Y. Yao[1], A. Sorant[2], H.H. Li[3], L. Spotila[4], M. Devoto[3], A. Wilson[2]
[1]Bloomberg SPH, Johns Hopkins Univ., Baltimore, MD, USA; [2]NHGRI, NIH, Baltimore, MD, USA; [3]duPont Hospital for Children, Wilmington, DE, USA; [4]Drexel Univ, Philadelphia, PA, USA

Although the variance components (VC) method is powerful for detecting linkage when pedigrees are randomly selected, the power can be reduced when ascertainment of families is non-random. In this study, the statistical properties of the VC method were evaluated under random and non-random ascertainment for various genetic models. In addition to random sampling, we considered three ascertainment schemes: selecting families with at least one offspring with a high trait value (H1); at least two with high values (H2); and extremely discordant offspring, one with a high trait value and one with a low value (ED). For models considered, the power of the VC method was consistently reduced under schemes H2 and ED, but increased under scheme H1 when the heritability was at least 50%. We applied this finding to a set of osteoporosis data consisting of 71 nuclear families collected under an H1 scheme, analyzing femoral neck bone mineral density, which has heritability estimated at 40%–70%. Analyzing this data using the VC approach, we detected a linkage signal on chromosome 1p36with a maximum lod score of 3.53 for linkage to a locus near marker D1S2694.

IGES-151

The genetic epidemiology of DNA concentration

D. Zabaneh[1], L.J. Collins[1], S.T. Bennett[1], I.J. Mackay[1] and the FAMOS study collaborators
[1]Oxagen Limited, Oxon, UK

The quantity of DNA extracted from small volumes of blood donated by volunteers is of fundamental importance in most genetic studies. In the limit it determines whether an individual can be included in the study. In general it will determine how much genotyping can be carried out before an individual is dropped from a study. Moreover, if DNA content itself is heritable, then recruitment into both family and population based studies may be biased towards alleles or haplotypes associated with QTL for increasing DNA content. We have studied the genetic epidemiology of DNA concentrations using a set of 561 pedigrees collected by the Family Osteoporosis (FAMOS) study as part of a study into the genetic architecture of osteoporosis. DNA concentration from 9 ml of blood was determined by the standard PicoGreen assay. Data were available from 2813 individuals in total. We show that log DNA concentration (measured in mcg/ml) is normally distributed and is of high heritability (0.55). We examine the effects of age, sex, height and weight of the donors, and report results from a whole genome linkage scan. The implications of these results for study design are discussed.

IGES-152

Linkage disequilibrium mapping of the MHC for juvenile oligoarthritis susceptibility loci

E. Zeggini[1], W. Thomson[1], W.E.R. Ollier[2], R.M. Lamb[1], P. Sevon[3], the BPRG Study Group, H. Toivonen[3], R.P. Donn[1]
[1]ARC EU, Univ. of Manchester, UK; [2]CIGMR, Univ. of Manchester, UK; [3]Dept. of Computer Science, Univ. of Helsinki, Finland

We have designed and carried out a sequential fine-mapping strategy aiming to examine whether the HLA loci associated with juvenile oligoarthritis are true susceptibility genes, and to investigate the presence of additional disease-predisposing loci within the MHC. One hundred and forty four simplex families consisting of an offspring affected with juvenile oligoarthritis and healthy parent(s) were initially typed for HLA-A and HLA-DRB1. Independent linkage and association of the two loci was established by using the following methods; ETDT, CETDT, PETDT and EHplus. Ten microsatellite markers spanning the region of interest were then typed and disease gene-containing intervals were refined through further complementary analyses. Four candidate genes were subsequently selected and extended haplotypes containing all 16 loci were constructed through GENEHUNTER and analysed through TDTphase, HPM and TreeDT. In addition, the independence of the effects observed was established by carrying out LD analyses and the CETDT. The MHC was found to harbour at least 5 independent loci that play a role in the

aetiopathogenesis of juvenile oligoarthritis; these are HLA-A and HLA-DRB1, together with 3 additional disease loci that we have localised. Fine mapping of the implicated intervals is now ongoing, in order to identify true juvenile oligoarthritis susceptibility genes within the MHC.

IGES-153

Latent variable models for genetic analysis of complex ordinal traits

H. Zhang, R. Feng, H.T. Zhu
Yale University School of Medicine, USA

Many health conditions including cancer and psychiatric disorders are believed to have a complex genetic basis and genes and environmental factors are likely to interact one another in the presence and severity of these conditions. Assessing familial aggregation and inheritability of disease is a classic topic of genetic epidemiology, which is commonly referred to as segregation analysis. While it is routine now to conduct such analyses for quantitative and dichotomous traits, there do not exist methods and software that accommodate ordinal traits. To this end, we propose and explore a latent variable model. The advantage of this latent variable model lies in its flexibility to include environmental factors (usually represented by covariates) and its potential to allow gene-environment interactions. The model building employs the EM algorithm for maximization and a peeling algorithm for computational efficiency. Asymptotic theory is provided for statistical inference and simulation studies are conducted to confirm that the asymptotic theory is adequate in practical applications. We also apply our model to examine the familial transmission of alcoholism, which is categorized into three ordinal levels: normal control, alcohol abuse, and alcohol dependence. Not only does our analysis confirm that alcoholism is familial, but also it suggests that the transmission may have a major gene component which was not revealed by previous analyses using dichotomous traits.

IGES-154

Transmission/disequilibrium tests using multiple tightly linked markers and haplotype sharing

S. Zhang, H.S. Cheng, J. Dong, R. Jiang
Department of Mathematical Science, Michigan Technological University, Houghton, MI, USA

Studies using haplotypes of 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 the identification of genes underlying complex traits. One problem for using multi-marker haplotype is that the larger number of haplotypes results in a large number of the degree of freedom for the test statistic. In this article, we propose novel statistical methods to analyze multiple tightly linked markers of family data. Our method both for quantitative trait and qualitative trait use the idea of TDT test and haplotype sharing. Simulation studies suggest that our methods are more powerful than the existing methods which include the existing single marker methods and the method based on multiple marker haplotypes.

IGES-155

On defining haplotype blocks

X. Zhu[1], R.S. Cooper[1], G. Cao[1], D. Kan[1], S. Zhang[2]
[1]Dept. of Preventive Medicine and Epidemiology, Loyola University Chicago, USA; [2]Dept. of Mathematical Science, Michigan Technological University, USA

Haplotype based association is a powerful approach to detect genetic variants of complex diseases. Recently, the global patterns of human genome are shown to be much simpler after defining haplotype blocks, in which there is little evidence for historical recombination and limited haplotype diversity. Here we propose a method to define the haplotype blocks based on haplotypic heterozygosity. A test statistic and confidence band to test a polymorphism belonging to a block is constructed by using the bootstrapping resampling procedure. We apply the method to the data of Gabriel et al. (Science, 2002). We compared our results with the published and the statistic properties are discussed.

IGES-156

Analysis of pooled DNA – A decision theoretic approach

A. Ziegler, I.R. König
Institute of Medical Biometry and Statistics, Medical University of Lübeck, Germany

Systematic analysis of the genetic background of complex diseases using single nucleotide polymorphisms (SNPs) affords a tremendous amount of genotypings. This is mainly due to two factors. Firstly, complex diseases are usually determined by interactions of multiple genetic and environmental factors. Under the conventional assumption that the influence of each single gene decreases with the overall number of involved genes, large sample sizes are required in a single study to obtain sufficient power. Secondly, the significance levels in systematic analyses have to be adjusted for multiple genome-wide testing. To reduce the amount of genotypings necessary and hence the overall cost of a case-control study with SNPs, some laboratories perform the genotyping in two stages: In the first, the DNA of all cases and all controls is mixed and genotyped

for each SNP. The frequency of both alleles is determined in both pooled DNA samples. If different frequencies are observed in the samples of the cases and the controls, genotyping is performed individually in the second stage and analysed as usual. This procedure is discussed under a methodological view. The aim is to find an optimal strategy regarding possible decision criteria for the individual genotyping. This can be viewed as a decision theoretic problem which is depicted in a decision tree. Various decision rules are discussed with respect to their applicability.

IGES-157

Promoter polymorphisms of the genes encoding TNF-α and IL-1β are associated with different subtypes of psoriasis characterized by early and late disease onset

A. Ziegler[1], R. Mössner[2], I.R. König[1], G. Westphal[3], C. Neumann[2], K. Reich[2]
[1]Inst. of Med. Biometry & Stat.Med. Univ. Lübeck, FRG; [2]Dept. of Dermatol.Georg-August Univ. of Göttingen, FRG; [3]Occupational Health, Georg-August Univ. of Göttingen, FRG

Gene polymorphisms that affect cytokine production may contribute to the disease-associated cytokine imbalance and influence susceptibility to psoriasis. Here, we investigated the interaction between polymorphisms in the genes encoding for TNF-α (G-238A, G-308A), IL-1β (C-511T, T + 3953C), and IL-1Ra (intron 2), and cytokine production in peripheral blood mononuclear cells of healthy donors, and analyzed the distribution of these polymorphisms in patients with psoriasis vulgaris (n = 231) and healthy controls (n = 345). Carriage of TNF A-238 allele 2 (−238*A) was associated with increased production of TNF-α in response to lipopolysaccharide in vitro, and with early onset disease (<40 y), especially in male patients with psoriasis [32% vs 7% in male controls; OR = 6.78, 95%CI = (3.18–15.15), p(adjusted) = 2×10^{-7}]. Carriage of the IL-1B-511*1 (−511*C) homozygous genotype was associated with increased production of IL-1Ra in response to lipopolysaccharide and IL-10, and with late onset psoriasis [⩾40y; 61% vs 44% in controls; OR = 2.04, 95%CI = (1.19–3.53)p(adjusted) = 0.04]. These findings indicate that gene polymorphisms associated with altered cytokine responses in vitro may modify age of onset of psoriasis. They also provide further evidence that patients with early and late onset psorias differ in their genetic background.

IGES-158

Analyzing Microarray Data in the Context of Diseases: Discovery and Interpretation

L.C. Lazzeroni
Stanford University

Microarrays and other new technologies have led to an enormous increase in the speed of data collection. Paradoxically, the rising quantity of available data makes it more difficult to interpret and validate the results of most statistical analyses in a usable way. This is true whether an analysis consists of a very large number of simple statistical procedures, such as t-tests, or is based on a more complicated pattern-discovery method, such as the plaid model or other types of cluster analysis. I will discuss this problem and describe some modifications of existing methods of analysis aimed at improving the interpretability of the statistical results.

IGES-159

Sample size estimates for candidate gene studies in pharmacogenetics

L.J. Sheffield
Murdoch Children's Research Institute, Royal Children's Hospital, University of Melbourne, Parkville, Victoria, Australia

The sample size required for a candidate gene case-control genetic association study is not established. There are frequent criticisms that case-control studies in pharmacogenetics have an inadequate sample size. I compare the sample sizes required for different genotypic relative risks and different gene frequency of marker alleles. Recently Risch and Teng have proposed a method for sample size calculations for pooled samples and McGinnis et al. have proposed a method for non pooled samples. Their methods are for whole genome scans rather than for candidate gene studies. For the latter studies a different correction needs to be made for multiple testing. I suggest a Bonferroni correction for 50 tests to allow for multiple alleles and a number of polymorphisms. It should be noted that until recently the only method available for estimating such sample sizes was the classical statistical approach which does not assume any genetic model. A comparison is made with this approach which both over estimates and underestimates the sample size for the dominant model. For example for an odds ratio of 3 (Type 1 error of .001, power of 80% and marker gene frequency of .2), the classical method suggests 148 cases and controls need studying whilst Risch and Teng's method estimate 135 for the dominant model and 219 for the additive model. As the known genotypic relative risk (for each genotype) and the genetic model may be able to be estimated from previous work, the genetic methods should be used. In pharmacogenetic studies, the genotypic relative risk is typically 20–40 and may be even higher. This means that quite modest sample sizes may be statistically adequate and is in contrast with the published examples in the sample size papers that use a maximum genotypic relative risk of 4. For example only 33 cases and controls are required with a genotypic relative risk of 20 (using the method of Risch and Teng).