ABSTRACTS FROM THE

ANNUAL MEETING OF THE INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

1

General Class of Family-based Association Tests for Sequence Data, and Comparisons with Population-based Association Tests

Iuliana Ionita-Laza (1) Seunggeun Lee (2) Vladimir Makarov (3) Joseph D Buxbaum (3) Xihong Lin (2)
(1) Columbia University (2) Harvard University (3) Mount Sinai School of Medicine

Recent advances in high-throughput sequencing technologies make it increasingly more efficient to sequence large cohorts in many complex traits. Making sense of these massive data depends critically on the use of efficient study designs and appropriate statistical methods. We discuss a general class of sequence-based association tests for family-based designs, that corresponds naturally to previously proposed population-based tests, including the classical burden and variance-component tests. This framework allows for a direct comparison between the power of sequence-based association tests for family- and population-based designs. We show that for dichotomous traits, family-based designs result in similar power levels as the population-based designs, while for continuous traits (in random samples) the population-based designs can be substantially more powerful. A possible disadvantage of population-based designs is that they can lead to increased false-positive rates in the presence of population stratification, and although adjustment for population substructure using principal component analysis can properly adjust at a small loss in statistical power, in more subtle scenarios, it can fail to completely adjust for such confounding. This is unlike the family-based designs, which are, by definition, robust to population stratification. We will also present results from an application to an exome-sequencing family-based study on autism spectrum disorders.

2

Family-based Designs for Sequencing Studies

Duncan C Thomas (1)
(1) University of Southern California

The cost of next-generation sequencing is approaching that of early genome-wide SNP genotyping panels, but still out of reach for the large-scale studies with the tens of thousands of subjects that will be required for testing association with disease. The millions of rare variants also pose a serious challenge for distinguishing causal from non-causal variants. Family-based designs for sequencing substudies provide a means to identify novel variants and prioritize them for their likelihood of causality by exploiting their co-segregation with disease. Other advantages include the possibility of enriching for family history, improved imputation, and efficient two-step analyses that exploit between- and within-family comparisons. Using simulation, we compare simple rule-based, likelihood ratio, Bayes factor, and score statistics in terms of their ability to distinguish causal from non-causal variants. We find that the Bayes factor criterion provides the best discrimination, but the score test is a close contender, is model free, and computationally simpler. Two-phase designs consistently provide better power for subsequent association testing than single-phase case-control designs. The approach is illustrated using two substudies within the Colorectal Cancer Family Registries, one large pedigree with 50 members genotyped at GWAS SNPs, one with whole exome data on a subset of families from a previous linkage scan.

Published online 31 August 2012 in Wiley Online Library (www.wileyonlinelibrary.com)
DOI: 10.1002/gepi.21677

3

Linkage Disequilibrium-based Subset Selection Approach for Rare Variants Analysis

Rajesh Talluri (1) Sanjay Shete (1)
(1) The University of Texas, M. D. Anderson Cancer Center

Rare variants have increasingly been cited as major contributors to a variety of diseases. Recently, several approaches have been proposed to analyze rare variants association with disease. Some approaches sum or group rare variants near a particular locus to improve power while others select the best group of rare variants for testing association such as the step-up approach. We propose a new approach based on the difference in Linkage Disequilibrium (LD) pattern between cases and controls to select the best subset of variants to include in the model. The motivation for using LD information to select the variables comes from the fact that linkage disequilibrium patterns are different in cases and controls. The LD based approach is robust to deleterious and protective effects of rare variants unlike the step-up type methods. The LD based subset selection approach is compared to the step-up method; we also explore three other approaches similar to the step-up method, which are the step-down, step-updown and the step-downup approach. The simulations were performed based on HapMap3 dataset of DRD2 gene. 1000 replicates each with 1000 cases and 1000 controls were simulated based on linkage disequilibrium pattern in DRD2 Gene. Permutations were used to control the type 1 error. The power comparisons after controlling the type 1 error show that the LD based approach is an attractive alternate method that can be considered for subset selection of rare variants.

4

NGS Transcriptome Genomic Analysis in Three Tissues of a Twin Cohort

Alfonso Buil (1) Andrew Brown (2) Matthew Davies (3) Ana Vinuela (3) Mercedes Gallardo (4) Dan Glass (3) Maria Blasco (4) Richard Durbin (2) Timothy D Spector (3) Emmanouil T Dermitzakis (1)
(1) University of Geneva, Geneva, Switzerland (2) Wellcome Trust Sanger Institute, Hinxton, United Kingdom (3) King's College London, London, UK (4) Centro Nacional de Investigaciones Oncologicas, Madrid, Spain

In this work we use RNA-seq data in a sample of twins to address the question of tissue specificity of the genetic regulation and dissect attributes of heritability of gene expression. We use the sample of the EUROBATS project that consists of ∼400 twin pairs with RNA from fat, LCLs and skin (2330 RNA-seq samples in total). At the moment, we have analyzed a pilot set of samples with 200 individuals in the three tissues. We quantified exonic reads and used imputed genotypes to determine cis eQTL. We observed 1901, 1776 an 1816 genes with at least a significant cis eQTL for fat, LCL and skin (FDR=10%). We also found that between 35% and 50% of the eQTL are shared among two tissues and that a 25% are shared by the three tissues. RNA-seq data allows the quantification of allele specific expression (ASE). At a 10% FDR we observed that 9.5%, 9.3% and 9.1% of the heterozygous sites had a significant ASE effect in fat, LCL and skin. We also found that 8.8% of the significant ASE sites were present in the same individual in the three tissues and that in all of them the direction of the ASE ratio was the consistent. Finally, using the ASE ratios of each sample in every heterozygous site, we defined an ASE distance among samples. We observed that the ASE distance is smaller for MZ twins than for DZ twins and for DZ twins than for unrelated individuals. With the analysis of the whole dataset we will study the tissue specific genetic determinants of alternative splicing.

5

Kernel Canonical Correlation Analysis for Assessing Gene-Gene Interactions

Nicholas B. Larson (1) Gregory D. Jenkins (1) Melissa C. Larson (1) Robert A. Vierkant (1) Thomas A. Sellers (2) Ellen L. Goode (1) Brooke L. Fridley (1)
(1) Mayo Clinic (2) Moffitt Cancer Center

While single-locus approaches have been widely applied to identify individual disease-associated polymorphisms, complex diseases are thought to be the product of multiple interactions between loci. This has led to the recent development of statistical methods for detecting statistical interactions between loci. Canonical correlation analysis (CCA) has been proposed to detect gene-gene co-association (Peng et al., 2010). However, this approach is limited to detecting linear relations and cases where the number of observations exceeds the number of SNPs in a gene. This point is particularly important for next-generation sequencing, which could yield a large number of novel variants on a limited number of subjects. To overcome these limitations, we propose an approach to detect gene-gene interactions based upon a kernelized version of CCA (KCCA). Simulation studies showed that KCCA controls the type I error. For genes with a moderate number of SNPs, we found that our method is comparably or more powerful than SNP-level logistic regression, PCA-based approaches, and the CCA procedure. We also applied KCCA to assess all interactions between 200 genes in the NFKB pathway for risk in a set of 3,869 ovarian cancer cases and 3,276 controls genotyped on the Illumina 610 array, followed by genotype imputation using HapMap 2. We identified 91 significant gene-pairs relevant to ovarian cancer risk (local FDR < 0.05).

6

Adjusted Sequence Kernel Association Test for Rare Variants Controlling for Cryptic and Family Relatedness

Karim Oualkacha (1) Zari Dastani (2) Rui Li (3) Tim Spector (4) Chris Hammond (4) Brent Richards (5) Antonio Ciampi (6) Celia Greenwood (7)
(1) UQAM University, Department of mathematics; McGill University, Epidemiology, Biostatistics and Occupational Healt (2) Lady Davis Institute; McGill University, Epidemiology, Biostatistics and Occupational Health (3) Lady Davis Institute (4) St. Thomas Hospital, Twin Research Unit (5) Lady Davis Institute; McGill University, Medicine; McGill University, Epidemiology, Biostatistics and Occupational Healt (6) McGill University, Department of Epidemiology, Biostatistics and Occupational Healt (7) Lady Davis Institute; McGill University, Oncology; McGill University, Epidemiology, Biostatistics and Occupational Healt

The advent of high-throughput sequencing technologies is providing an unprecedented opportunity to examine rare and unique variants possibly associated with complex traits. However, the results of such efforts depend essentially on the use of efficient statistical methods and study designs. Although family-based designs might enrich a data set for familial rare disease variants, most existing rare variant association approaches assume independent sampling. We present here a framework for association testing of rare variants in family-based designs. This framework is an adaptation of the sequence kernel association test (SKAT) which allows us to control for family structure. Our adjusted SKAT (ASKAT) combines the SKAT approach and the factored spectrally transformed linear mixed models (FaST-LMM) algorithm to capture family effects based on a linear mixed model incorporating the realized proportion of the genome that is identical by descent between pairs of individuals. We evaluated type I error and power of this proposed method based on simulation studies and we showed that, regardless the level of the trait heritability, our approach provides good control of type I error and good power. Since our approach uses FaST-LMM to calculate variance components for the proposed mixed model, ASKAT is fast and can analyze hundreds of thousands of markers. To illustrate the ASKAT methodology, data from the UK twins consortium are analyzed.

7

Combining p-values from Linear and Quadratic Tests for Rare Variants Provides Robust Statistics Across Genetic Models

Andriy Derkach (1) Jerald F Lawless (2) Lei Sun (3)
(1) Department of Statistics, University of Toronto (2) Department of Statistics and Actuarial Science, University

of Waterloo (3) Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto

Many association tests have been proposed for rare variants, however, the practical choice of a powerful test is often uncertain because of limited information available on the true genetic model. Proposed methods use either linear statistics, which are powerful when most variants are causal with same direction of effect, or quadratic statistics, which are more powerful in other scenarios. The differences in power, unfortunately, can be significant in many settings, e.g. 20% vs. 60% or vice versa. To achieve robustness, it is natural to combine the evidence of association from two or more complementary tests. To this end, we consider the minimum-p and Fisher's methods of combining p-values from linear and quadratic statistics. Extensive simulation studies of over 10,000 different models show that both methods are robust across models with varying proportions of causal, deleterious and protective rare variants, variant frequencies, effect sizes and the relationships between variant frequencies and effect sizes. When a linear or quadratic test has moderate power of 20% or more, Fisher's test has considerable better power than each test for 90% of the models considered, while the maximum absolute power loss is 8% for the remaining 10% of the models. An application to the GAW17 quantitative trait Q2 data based on sequence data from the 1000 Genomes Project shows that Fisher's test has comparable power for all 13 functional genes and has the best power for more than half of them.

8

Re-ranking Next Generation Sequencing Variants for Accurate

Laura L Faye (1) Shelley B Bull (1) Peter Kraft (2) Lei Sun (3)
(1) Dalla Lana School of Public Health, University of Toronto; Samuel Lunenfeld Research Institute, Mount Sinai Hospital (2) Program in Molecular and Genetic Epidemiology, Dept of Epidemiology, Dept of Biostatistics, Harvard School of Public Health (3) Dalla Lana School of Public Health, University of Toronto; Department of Statistics, University of Toronto

Identification of causal variants is critical to understanding the genetic underpinnings of complex diseases. Next generation sequencing (NGS) technology has dramatically increased our ability to interrogate disease-associations at the base-pair level. Precise identification of causal variants, however, remains a challenge. Differential and/or high error rates common to NGS due to low coverage, short read length, imperfect variant alignment and other sequence-specific errors, as well as the re-use of genome-wide association study (GWAS) data can induce bias in the evidence for association and obscure the location of the true causal variant. We quantify analytically the effects of these factors and show that they can reduce considerably the probability of correctly identifying the causal SNP, often by more than half. Further, we demonstrate that in certain common scenarios, failure to account for these factors causes evidence to accumulate at the wrong SNP as sample size grows. We develop a procedure to re-rank SNPs that accounts for the effects of differential genotyping error and linkage disequilibrium pattern among SNPs, and/or the use of GWAS signals to inform analysis. The re-ranking procedure substantially increases the localization success rate. We empirically evaluate our re-ranking method in sequencing data at several different read-depths generated by thinning reads from targeted high-coverage sequencing in the Cancer Genetic Markers of Susceptibility Study.

9

The Role of Aging in the Heritability of Epigenomic Markers

Alicia L. Lazarus (1) Jennifer A. Smith (1) Thomas H. Mosley Jr. (2) Stephen T. Turner (3) Sharon L.R. Kardia (1)
(1) Department of Epidemiology, University of Michigan (2) Department of Medicine (Geriatrics), University of Mississippi Medical Center, Jackson, MS (3) Division of Nephrology and Hypertension, Mayo Clinic

DNA methylation (DNAm) is a potential molecular mechanism to link the aging process to chronic disease outcomes that are a major cause of morbidity/mortality in the elderly. Since aging processes are known to be under genetic control, we evaluated 1,008 African-Americans (ages 39–94) from sibships within the Genetic Epidemiology Network of Arteriopathy study using the Illumina HumanMethylation27 array to estimate the heritabilities of four array-specific DNAm measures: 1) methylated signal, 2) unmethylated signal, 3) Beta Value and 4) M-Value. The four DNAm measures had high average heritability (0.30 to 0.41), and ~80% of the DNAm sites had significant heritabilities ($p<0.05$) ranging 0.11 to 1.0. In addition, we found that many of the heritabilities decreased after age adjustment, indicating that a portion of the genetic variance was due to age-related genetic factors. To explore this phenomenon, we then partitioned the heritabilities into age-related and residual heritability components. For the DNAm measures with significant heritability, the average percentage of total genetic variability due to age-related genetic factors ranged from 14.8% (M-Value) to 28.4% (unmethylated signal). In the M-Value, ~500 sites had $\geq$50% of their genetic variance due to age-related genetic factors. These results indicate that DNAm is heritable, and is potentially affected by genes that have differential effects across the age spectrum that may be related to chronic disease risk.

10

Variant Association Tools (VAT): A Pipeline to Perform Quality Control and Association Analysis of Sequence and Exome Chip Data

Gao Wang (1) Bo Peng (2) Suzanne M. Leal (1)
(1) Baylor College of Medicine (2) The University of Texas MD Anderson Cancer Center

Currently there is great interest in detecting associations of complex human traits with rare variants. Methods testing for rare SNV association aggregate variants across genes. We developed *variant association tools* (VAT), a tool-set that implements best practice for rare variant association studies. Highlights include variant site/call level quality control, summary statistics (HWE, Ti/Tv, etc), phenotype/genotype based sample selections, variant annotation and selection of loci for analysis. Within this pipeline a large number of rare SNV association tests have been implemented to analyze both qualitative and quantitative traits. Our association testing framework is regression based which readily allows for flexible construction of association models with

multiple covariates, weighting (based on frequency or predicted functionality), interactions terms and models for pathway analysis. VAT is capable of rapidly scanning through the exome using multi-processes computation, adaptive permutation and conducting multiple association tests simultaneously. Results can be view as text or graphs. It also provides a programming interface to allow for user implementation of novel association methods. The VAT pipeline can be applied to sequence data, imputed sequence data and exome chip array data. This pipeline should be beneficial in performing association analysis on the large amounts of sequence and exome chip data which is currently being generated for studies of complex traits.

11

A Framework for the Evaluation of Rare Variant Tests of Association

Nathan Tintle (1) Keli Liu (2) Shannon Fast (3) Matthew Zawistowski (4)
(1) Dordt College (2) Harvard University (3) Massachusetts Institute of Technology (4) University of Michigan

The tidal wave of next-generation sequencing (NGS) data has arrived, but more questions than answers exist about how to best analyze NGS data to investigate the potential contribution of rare genetic variants to human disease. Numerous rare variant association testing methods have been proposed which attempt to aggregate association signals across multiple variant sites in an effort to increase statistical power. While emerging simulation results suggest that some rare variant testing methods work better than others for particular genetic architectures, little concrete understanding of the tests is available. We propose a geometric framework which quickly classifies existing rare variant tests of association into two broad categories: length and joint tests. We then demonstrate how genetic architecture (relative risk distribution, allele frequency distribution and number of variants) directly relates to the behavior of length and joint tests. We go on to illustrate further implications of the geometric framework including the differential impact of variant weighting strategies, population stratification and genotype uncertainty on length and joint tests. We also describe how the geometric framework suggests numerous potential alternative rare variant association tests and how they will behave. The geometric framework articulates the connection between disease architecture and test behavior, providing a clear set of next steps for applied and theoretical researchers.

12

Efficient Meta Analysis of Rare Variant Associations via Summary Statistics

Dajiang J. Liu (1) Goncalo R. Abecasis (1)
(1) University of Michigan

There is great interest in understanding complex trait etiologies due to rare variants. In order to increase power, multiple cohorts can be jointly analyzed. MEGA analysis using individual participant data for large scale genetic studies can be challenging. Instead, performing meta-analysis using summary level statistics is desirable. For single variant association analysis, it is known that meta-analysis has comparable efficiency as MEGA analysis. However, for the detection of rare variant associations, multiple rare variants in a gene/region are usually jointly analyzed in order to aggregate information and increase power. Fundamentally different statistical methodologies, data sharing protocols as well as computational tools are needed for meta-analysis of rare variant associations. To address these challenges, we proposed a novel statistical framework for performing meta-analysis of rare variant association studies. We showed theoretically that through summary level data of minor allele frequencies, single variant association test statistics, as well as linkage disequilibrium information for the gene region (e.g.), most rare variant association tests (e.g. burden tests, variable threshold tests and sequence kernel association tests) can be exactly implemented as in MEGA analysis without losing any information. Our methods and tools were comprehensively evaluated via simulated data and being applied to a large scale meta-analysis of lipids levels using exome-chip data.

13

Testing for Rare Variant Associations in the Presence of Missing Data

Paul L Auer (1) Gao Wang (2) Suzanne M Leal (2)
(1) Fred Hutchinson Cancer Research Center (2) Baylor College of Medicine

For studies of complex diseases, many association methods have been developed to specifically analyze rare variants. When variant calls are missing, naive implementation of rare variant association (RVA) methods can lead to inflated false positive rates as well as a reduction in power. For case-control data, only power will be reduced when the rate of missingness is equivalent between cases and controls; however, if the rates are differential between cases and controls there can be an increase in the false positive rate. For next generation sequencing data, it is inevitable that variant calls will be unavailable in specific genomic regions and sets of samples. Differential missingness can be caused when cases and controls are sequenced in batches that are subject to different experimental conditions or if convenience controls are used. Using data from the NHLBI-Exome Sequencing Project we demonstrate that differential missingness can cause substantial increases in type I error for commonly used RVA methods. We developed extensions for four commonly used RVA tests and show that they control false positive rates without a reduction in power. Additionally, power to detect an association can be substantially improved by using these extended methods compared to removing variant sites with missing calls. In order to maintain proper control of type I error without sacrificing power the extended RVA methods should be implemented when analyzing sequence data.

14

A Method to Detect Differentially Methylated Loci With Next Generation Sequencing

Hongyan Xu (1) George Varghese (1)
(1) Georgia Health Sciences University

Epigenetic changes, especially DNA methylation at CpG loci has important implications in cancer and other complex diseases. With the development of next-generation sequencing (NGS), it is feasible to generate data to interrogate the

di erence in methylation status for genome-wide loci using case-controldesign. However, a proper and efficient statistical test is lacking. There are several challenges. First, unlike microarrays methylation data, here we have the counts of methylation allele and unmethylation allele for each individual. Second, due to the nature of sample preparation, the measured methylation re ects the methylation status of a mixture of cells involved in sample preparation. Therefore, the underlying distribution of the measured methylation level is unknown, and a robust test is more desirable than parametric approach. Third, currently NGS measures mehtylation at over 2 million CpG sites. Any statistical tests have to be computationally efficient in order to be applied to the NGS data.Taking these challenges into account, we propose a robust test for di erential methylation based clustered data analysis by modeling the mehtylation counts. We performed simulations to show that it is robust under several distributions for the measured methylation levels. It has good power and is computationally efficient. Finally, we apply the test to our NGS data on chronic lymphocytic leukemia. The results indicate that it is a promising and practical test.

15

Lessons Learned from Analysis of DNA Methylation Array Data

Brooke L. Fridley (1) Sebastian M. Armasu (1) Melissa C. Larson (1) Mine S. Cicek (1) Robert A. Vierkant (1) Viji Shridhar (1) Janet E. Olson (1) Julie M. Cunningham (1) Kimberly R. Kalli (1) Ellen L. Goode (1)
(1) Mayo Clinic

In the post-GWAS era, many genetic studies are augmenting SNP data with genome-wide DNA methylation data. Often, methylation data are integrated with mRNA expression and SNP data to understand the possible functional mechanism by which the SNP might influence the trait. These types of methylation studies often involve: Infinium HumanMethylation27 (27K) or HumanMethylation450 Beadchips (450K); DNA extracted from lymphocytes in blood or tumor tissue. Over the course of the last year, we have gained experience in the analysis of methylation data from both the 27K (N∼500) and the 450K (N∼750) arrays using DNA from ovarian tumor tissue (N∼340) and lymphocytes (N∼775). During these analyses, we learned several meaningful lessons and therefore recommend the following: (1) a carefully considered plate design; (2) inclusion of duplicates and negative/positive controls; (3) assessment of artifacts, such as plate effects, and normalization; (4) analysis of the methylation beta value on the 0 to 100 percent scale; (5) assessment of covariates that differ between cases and controls (as these can impact blood based methylation); and (6) validation of methylation levels with a second assay. We will provide specific examples of our analysis approach. With appropriate attention to the analysis and interpretation of methylation data, it can provide added layers of information to our understanding of the underlying genomic make-up of complex phenotypes.

16

Association Testing of Mitochondrial Genome Using Pedigree Data

Chunyu Liu (1) Josee Dupuis (2) Martin G Larson (2) Daniel Levy (1)
(1) Framingham Heart Study/NHLBI (2) Boston University

In humans, mitochondria (mt) contain their own DNA (mtDNA) that is inherited exclusively from the mother. The mtDNA encodes thirteen polypeptides that are components of the major metabolic pathways of energy production. Any disruption in these genes might interfere with energy production and thus contribute to metabolic derangement. Mitochondria also regulate several important cellular activities including cell death and calcium homeostasis. As a result of drastic declines in sequencing cost, the mtDNA will soon be sequenced in several cohort studies. Association testing of mitochondrial variants with disease traits raises unique challenges because of the difficulty in separating the effects of nuclear and mitochondrial genomes, which display different modes of inheritance. Failing to correctly account for these effects might decrease power or inflate type I error in association tests. We proposed several strategies to account for polygenic effects of the nuclear and mitochondrial genomes and performed extensive simulation studies to evaluate the type I error and power of these strategies. In addition, we described a permutation test that can be used to obtain empirical p-values for these strategies. Furthermore, we applied two of the analytical strategies to association analysis of 196 mt variants with blood pressure and fasting blood glucose in the Framingham Heart Study data. Our results provide guidance for association testing of mt variants in large pedigrees.

17

Association of X Chromosomal Variants With Coronary Heart Disease: Results from a Meta-analysis

Christina Loley (1) Heribert Schunkert (2) Jeanette Erdmann (2) Inke R Konig (3)
(1) Institut fur Medizinische Biometrie und Statistik, Universitat zu Lubeck, and Medizinische Klinik II, UKSH, Lubeck, Germany (2) Medizinische Klinik II, Universitatsklinikum Schleswig-Holstein, Campus Lubeck, Lubeck, Germany (3) Institut fur Medizinische Biometrie und Statistik, Universitat zu Lubeck, Lubeck, Germany

In many genome-wide association studies, the focus is on the analysis of autosomal markers, regions on the X chromosome are often neglected. Thus, potentially important genetic information remains undetected. The likely reason for this is that no standard statistical method exists to adequately deal with X chromosomal data.
Specific suggestions to analyze association on the X chromosome have been made and compared systematically in simulation studies (Loley et al. 2011). However, most approaches do not estimate effects and standard errors, so that a meta-analysis cannot be performed.
In this project, we analyze X chromosomal associations with coronary heart disease, where gender specific effects point at an important role of the sex chromosomes. Based on data from the CARDIoGRAM consortium (Schunkert et al. 2011) with genotypes imputed on the 1000 genomes data, we used logistic regression models to perform meta-analyses on the X chromosome. The models chosen account for the specific data structure of chromosome X, including the phenomenon of inactivation of one female X chromosome. This approach easily allows for adjustments for environmental factors and

population stratification as well as calculation of fixed and random effect estimates. Results on the association with coronary heart disease will be presented.
Loley C, Ziegler A, Konig IR, Hum Hered, 2011, 71, 23–36
Schunkert H, Konig IR, Kathiresan S et al. Nat Genet, 2011, 43, 333–8.

18

Trans-ethnic Mapping in Admixed African Americans

Daniel Shriner (1)
(1) National Human Genome Research Institute

Admixed populations represent a special case for multi-ethnic genome-wide association studies. Whereas in conventional multi-ethnic studies an individual is ancestrally homogeneous, an admixed individual is ancestrally heterogeneous. We illustrate methods accounting for local ancestry when assessing transferability of association across ethnicities. Two major conclusions are that weaker linkage disequilibrium levels in individuals of African ancestry can substantially enhance localization of association signals and the identification of more strongly associated variants can explain some of the missing heritability. Admixed populations also provide an opportunity to map trait loci using joint admixture mapping and association testing in the same admixed individuals. Analysis of 29 traits in admixed African Americans reveals 14 genome-wide significant admixture peaks and 95 genome-wide significant loci using the joint test compared to 28 loci using conventional association analysis. Ancestry effects reflect markers with differentiated allele frequencies and/or heterogeneous effect sizes between ethnicities. Many genetic variants in African Americans are segregating in Africans and absent in Europeans. These findings suggest that heritability in admixed African Americans can be explained via population-specific variants inherited through African ancestry, a smaller number of population-specific variants inherited through European ancestry, and cosmopolitan variants.

19

Evidence of Recent Positive Selection in Africans at Known and Novel BMI Loci

Todd L Edwards (1) Digna R Velez Edwards (1) George J Papanicolaou (2) Kari E North (3) The African American BMI GWAS consortium NA (4)
(1) Vanderbilt University (2) NHLBI (3) University of North Carolina (4) NA

The obesity epidemic in the US is a major public health concern with notable disparity between Europeans (EA) and persons of recent African ancestry (AA). AAs have higher rates of obesity, higher risk of sequelae, and more acute sequelae of obesity compared with EAs. We participated in a meta-analysis of 39,144 AA adults with GWAS data from 19 studies, and replication in 10,000 participants to discover the genetic determinants of body mass index (BMI) in AAs. We evaluated the novel and previously discovered gene regions for evidence of recent selection in the Human Genome Diversity Panel, HapMap, Perlegen, and the BioVU GWAS data. The region near the novel BMI gene region containing *KLHL32* had evidence of recent selection in Africans from all statistics (iHS=4.05, XP-EHH=1.3, Tajima's D=2.7, Treeselect $p=4\times10^{-4}$, Max $F^{st}=0.30$), and has been previously associated with cardiac traits and mitochondrial complex 1 deficiency. *ADCY3* also lies in a region of recent selection in Africans (iHS=5.2, XP-EHH=1.8, Tajima's D=2.8, Treeselect $p=2.6\times10^{-3}$, Max $F^{st}=0.27$), and is associated with obesity in humans and mice. *ADCY3* -/- mice also lack olfaction, research has shown that ADCY3 functions in olfaction in higher mammals, and previous olfaction research in humans has described statistically significant differences in olfaction between AAs and EAs. These results suggest that selective pressure at BMI loci may determine in part the disparity in BMI between AAs and EAs.

20

Over-adjustment of Population Stratification and a Model-averaging Based Solution

Dalin Li (1) Jerome I Rotter (1) Xiuqing Guo (1)
(1) Cedars-Sinai Medical Center

Population stratification can introduce confounding and lead to inflated type 1 error rates in genetic association studies. Adjustment based on inferred genetic ancestry, typically using a few top principle components derived from population stratification analysis, is a standard approach that is usually applied to all SNPs in the genome indifferently. However, it may not be true that all SNPs are confounded by the population substructure (PS). Our simulation studies demonstrated that for the un-confounded SNPs, over-adjustment using the top principle components can lead to greatly reduced power. We developed here a flexible population-stratification adjustment approach, in which two associations, 1) association between PS and genetic variants (G) and 2) association between PS and the outcome (Y), are explicitly and jointly modeled. Then a model averaging approach is used to average over the models with and without adjustment for population substructure. The weights for the adjusted/unadjusted models are defined based on the probability of PS being a confounding factor, depending on the association between PS and Y/G in the joint model. We demonstrated using simulated data that the proposed approach can correctly control for the type I error and can be 10–20% more powerful than traditional population stratification analysis when PS is actually not confounding the association between Y and G.

21

Developmental Processes can Drive Inter-chromosomal Linkage Disequilibrium

Alexander M Kulminski (1) Irina Culminskaya (1) Anatoli I Yashin (1)
(1) Duke University

Recently (Age, 2012, PMID: 22282054) we documented two clusters of extensive inter-chromosomal linkage disequilibrium (LD) in the Framingham Heart Study families. Both clusters were observed on two independent arrays, the Affymetrix 500K and 50K. The LD has unlikely been generated by stochasticity, population or family structure, or mis-genotyping. Here we report on significant enrichment of genes for SNPs in LD in Gene Ontology (GO) biological processes. Annotation of two (partly overlapping) clusters of SNPs in LD with minor allele frequency 0.009–0.499 results in two reference sets of 1697 and 4637 genes. These gene sets have been categorized into biological processes

using ArrayTrack and GoMiner tools. The analyses using the Fisher's test and the False Discovery Rate (FDR) approach reveal highly significant enrichment of GO terms linked to developmental process, particularly to nervous system development/axonogenesis ($p<10^{-4}$, $FDR<10^{-4}$) for both sets and cardiovascular system development for the largest set ($p\leq10^{-4}$, $FDR\leq0.007$). Both gene sets showed highly significant over-representation of biological processes tightly linked with neural development including, e.g., cell differentiation, morphogenesis, adhesion, projection organization, signaling, chemotaxis ($p<10^{-4}$, $FDR<10^{-4}$ for each category). The results provide evidences on significant role of functional interactions and developmental processes in generation of genome-wide LD.

22

Relaxing the Genetic Model to Identify Quantitative Trait Loci having Heterogeneous Effects

Hugues Aschard (1) Noah Zaitlen (1) Rulla M Tamimi (2) Sara Lindstrom (1) Peter Kraft (1)
(1) Department of Epidemiology – Harvard School of Public Health (2) Channing Laboratory, Brigham and Women Hospital – Harvard Medical School

Genome-wide association studies (GWAS) of quantitative phenotypes are usually conducted using linear regression assuming that risk alleles affecting natural variation in quantitative traits display a linear marginal effect. The utility of this assumption has been confirmed by the wide number of quantitative trait loci (QTL) identified using this approach. Still, summarizing the effect of a QTL with a single linear estimator may have limited statistical power when the magnitude and the direction of the effect depend on the genetic and environmental background of the individuals studied. To improve detection existing approaches have focused so far on strategies to allow for measured effect modifiers. We propose an alternative strategy which relies on a more flexible model for the marginal effect of QTLs. We developed a non-parametric score that compares phenotypic distributions by genotypes. This score can capture non-linear effects such as those displayed in the presence of interactions. We show via simulation that such approach can be more powerful than the standard tests for marginal effects when the effect of the QTL is heterogeneous. We further applied our test in a discovery context of a GWAS of mammographic density. We found a significant ($p=8.10^{-3}$) enrichment of genes differentially expressed in mammary epithelial cell or genes related to breast cancer among the top loci identified by our test, whereas no similar enrichment was observed for the marginal test ($p=0.27$).

23

Integrative Approaches for Genetic Association Studies via Bayesian Model Uncertainty

Melanie A. Quintana (1) David V. Conti (1)
(1) University of Southern California

Recent technological advances within the genetics community have lead to the collection of massive datasets involving numerous types of genomic data on a single set of individuals or across various levels of information. Traditional analyses concentrate on a single data type and rely heavily on an expert to build models based on contextual knowledge. However, with the amount of factors available for evaluation it is often impracticable to build a model in this way. With this in mind, we are interested in developing integrative approaches for high-dimensional genetic association studies that incorporate multiple-levels of data. In particular, we have developed iBMU, a Bayesian model uncertainty method that formally incorporates multiple sources of data via a multi-stage hierarchical probit model on the probability that each predictor is associated with the outcome of interest. Using simulations, we demonstrate that iBMU leads to a more efficient model search algorithm that yields an increase in power to detect true associations over more commonly used techniques. Finally, we demonstrate the power and flexibility of iBMU for several genome wide association studies involving: (1) the analysis of rare variants with incorporation of genetic annotation; (2) gene-set enrichment analysis for inference of gene- and pathway-level associations; (3) integrated genomic analysis incorporating functional information via biomarkers.

24

European and African Ancestry Interaction Increases the Susceptibility of Colorectal Cancer in Latino Populations

Gustavo A Hernandez-Suarez (1) Carolina Sanabria (2) Jovanny Zabaleta (3) Albert Tenesa (4)
(1) National Cancer Institute of Colombia (2) Nationa Cancer Insititute of Colombia (3) Louisiana State University, Louisiana, EU (4) Roslin Institute, The University of Edinburgh

Background: Latinos genetic background has a unique three way admixture: Amerindian, European and African. Countries from these reference populations show contrasting rates of Colorectal Cancer, mainly attributed to differences in diet lifestyle. Latin-American countries show rates just a third of those observed in the United States. The role of ancestry role in susceptibility to colorectal cancer in admixed South American populations has not been evaluated. Methods: We recruited 264 controls and 203 colorectal cancer cases from 5 Colombian cities and genotyped them for 678 independent SNPs cross matched with HapMap database. We estimated the admixture fractions using the STRUCTURE software assuming three distinct population origins from HapMap reference populations. Results: Mean ancestry fraction of Amerindian, European and African were 44.9%, 39.3% and 15.6% in the control group. After adjusting for all relevant risk factors (including educational level, physical activity and energy intake), an increase in African ancestry fraction was positive associated with colorectal cancer. (OR=3.75, CI95% 1.30–10.75). In cities with low African ancestry (mean 6% rank 1–20%), we found a positive interaction of African and European ancestry with colon and rectal risk (p value interaction 0.02 and 0.07 respectively).Conclusion: Higher African ancestry in Colombian population increases the susceptibility to colon and rectal cancer, especially among those with higher European ancestry.

25

Refining Association Mapping in a Heterogeneous Population

Amrita Ray (1) Laura C Lazzeroni (1)
(1) Stanford University

This presentation proposes a refined association mapping method to address the possibility of variable linkage disequilibrium (LD) structure within a heterogeneous population. We consider the situation in which markers associated with a trait fall into different linkage disequilibrium blocks in different ancestral populations. We propose a two-stage association statistic that combines aspects of principal component adjusted analyses and stratified analyses in order to address two issues simultaneously. Individuals are allowed to have mixed ancestry originating in more than one subpopulation and subpopulations are allowed to have different LD patterns. We will apply our methods to a genome-wide study of schizophrenia (GAIN) containing phenotype and genotype data on 2.8K European and 2.2K African American participants. We will use schizophrenia diagnosis and the total DSM score to analyze European and African American populations separately focusing on candidate genes previously characterized by the Consortium on the Genetics of Schizophenia (COGS) for association with schizophrenia and related endophenotypes. We will compare the association signals from our proposed two-stage association method in the heterogeneous GAIN populations to the relatively homogeneous Caucasian population in COGS and to one-stage association methods that allow less complex patterns of admixture and LD.

26

Models for Admixture Mapping in a Regression Framework

Jinghua Liu (1) Gary K. Chen (1) William J. Blot (2) Sara S. Strom (3) Sonja I. Berndt (4) Rick A. Kittles (5) Benjamin A. Rybicki (6) William Isaacs (7) Sue A. Ingles (1) Janet L. Stanford (8) W. RYAN Diver (9) John S. Witte (10) Ann W. Hsing (11) Barbara Nemesure (12) Timothy R. Rebbeck (13) Kathleen A. Cooney (14) Jianfeng Xu (15) Adam S. Kibel (16) Jennifer J. Hu (17) Esther M. John (18) Serigne M. Gueye (19) Stephen Watya (20) Lisa B. Signorello (2) Richard B. Hayes (21) Zhaoming Wang (22) Lisa Chu (18) Eric A. Klein (23) Phyllis Goodman (24) Edward Yeboah (25) Yao Tettey (25) Qiuyin Cai (26) Suzanne Kolb (8) Elaine A. Ostrander (27) Charnita Zeigler-Johnson (13) Yuko Yamamura (28) Christine Neslund-Dudas (6) Jennifer Haslag-Minoff (29) William Wu (29) Venetta Thomas (17) Glenn O. Allen (17) Adma Murphy (30) Bao-Li Chang (13) S. Lilly Zheng (15) M. Cristina Leske (12) Suh-Yuh Wu (12) Anna M. Ray (14) Anselm JM Hennis (12) Michael J. Thun (9) John Carpten (31) Graham Casey (1) Stephen J. Chanock (11) Daniel O. Stram (1) Brian E. Henderson (1) Christopher A. Haiman (1) David V. Conti (1)
(1) University of Southern California (2) Vanderbilt University (3) The University of Texas M. D. Anderson Cancer Center (4) National Cancer Institute (5) University of Illinois at Chicago (6) Henry Ford Hospital (7) Johns Hopkins Hospital and Medical Institutions (8) Fred Hutchinson Cancer Research Center (9) Epidemiology Research Program, American Cancer Society (10) University of California San Francisco (11) Division of Cancer Epidemiology and Genetics, National Cancer Institute (12) Stony Brook University (13) University of Pennsylvania School of Medicine and the Abramson Cancer Center (14) University of Michigan Medical School (15) Wake Forest University School of Medicine (16) Washington University, St. Louis (17) University of Miami Miller School of Medicine (18) Cancer Prevention Institute of California (19) Hopital General de Grand Yoff, Dakar, Senegal (20) Mulago Hospital/Makerere University, Department of Surgery, Urology Unit, Kampala, Uganda (21) New York University Langone Medical Center (22) Division of Cancer Epidemiology and Genetics, National Cancer Institute (23) Glickman Urological & Kidney Institute, Cleveland Clinic (24) Fred Hutch Cancer Center (25) University of Ghana Medical School, Legon, Ghana (26) Vanderbilt University and the Vanderbilt-Ingram Cancer Center (27) Cancer Genetics Branch, National Human Genome Research Institute (28) The University of Texas M. D. Anderson Cancer Center (29) Washington University (30) Northwestern University (31) The Translational Genomics Research Institute

We propose a novel flexible regression framework to perform admixture mapping for both case-only and case-control study designs. The method tests if the mean of the local ancestry among the cases significantly diverges from the mean of the local ancestry among the controls or, for case-only designs, if the local mean diverges from the genome-wide mean among the same set of cases. We demonstrate via simulations that this approach is more powerful than alternative admixture mapping techniques while allowing for the inclusion of additional covariates, if necessary. While the case-only design is more powerful it is also more susceptible to increased type I error from subtle biases in local ancestry estimation. In addition, since current admixture mapping studies use genome-wide SNP arrays, we compare this approach to conventional SNP association tests and find that the additional information from a test of admixture association rarely adds power for discovery. However, extensions of the regression framework capturing heterogeneity of SNP effects by local ancestry do offer a potential increase in power. We apply the various models to a real data set consisting of African Americans from a prostate cancer genome-wide association study in men of African ancestry and discuss the implications for genome-wide discovery in admixed populations.

27

Overestimation of Relatedness in Admixed and Ancestrally Heterogeneous Populations

Jean Morrison (1)
(1) University of Washington

It is common to estimate the proportion of the genome which is identical by descent (IBD) between pairs of individuals in studies involving genome-wide SNP data. These estimates can be used to check pedigrees, estimate heritability and adjust association analyses. One of the most commonly used tools for inference of genome-wide allele sharing probabilities is the method of moments technique implemented in PLINK and other software which estimates the proportions of the genome at which two individuals share 0, 1, or 2 alleles IBD. This technique is computationally efficient but requires that several strong assumptions hold in order to yield accurate estimates. The most problematic of these assumptions is that the study sample is drawn from a single, homogeneous, randomly mating population. This assumption is violated in the case of small samples with no close population match in publicly available data sets, bi-ancestral pedigrees and admixed populations. I used publicly available genome-wide SNP data to simulate pedigrees under these different conditions.

I am able to demonstrate that the method of moments estimator is biased in these conditions most often leading to overestimation of relatedness between ancestrally similar individuals. Finally, I propose a simple method easily implemented without additional software for improving genome-wide IBD estimates when the assumption of a single, homogeneous population is violated.

28

Race-dependent Associations Between Variations in ADRB1 and Ventricular Arrhythmias

Indrani Halder (1) Haider Mehdi (1) Madhurmeet Singh (1) Rebecca Gutman (1) Elisabeth Barrington (1) Ryan Aleong (2) Heather Bloom (3) Samuel Dudley (4) Patrick T Ellinor (5) Samir Saba (1) Alaa Shalaaby (1) Raul Weiss (6) Barry London (1)
(1) University of Pittsburgh (2) University of Colorado (3) Emory University (4) University of Illinois at Chicago (5) Massachusetts General Hospital (6) Ohio State University

Ventricular Arrhythmias (VA) remain a major cause of mortality in heart failure (HF) patients. Sympathetic stimulation and adrenergic signaling plays important roles in arrhythmogenesis. We investigated whether *ADRB1* polymorphisms are associated with VA in HF patients. Enrolled subjects had ejection fraction (EF) 0.30, implantable cardioverter defibrillators (ICD), genotyped for *ADRB1* polymorphisms S49G (A>G) and R389G (C>G) and prospectively followed for 5 years. Freedom from appropriate ICD shocks were compared by Kaplan Meyer tests and Cox regression. Median follow up on 1021 patients (80% white, 17% African American (AA), 80% male, 62.5± 12.5 years, EF 0.29±0.06, NYHA class 2.1±0.6) was 32 months. AAs were more likely to get shocks (P=0.005). Genotype frequencies for both loci varied between races; neither predicted shocks in the total sample. In AAs, significantly more appropriate shocks were observed in S49G Gly carriers (P=0.018) and marginally more in R389G Ser/Ser homozygotes (P=0.06). Cox models adjusting for age, gender, history of diabetes, medications, EF and individual admixture showed 2.2 times higher hazards of shock for S49G Gly carriers (P=0.016); R389G Ser/Ser homozygotes (P=0.025) and 2.8 times higher hazards (P=0.002) for those with both risk genotypes. No associations were observed in Whites. Common ADRB1 polymorphisms are associated with VAs in AAs and may partly explain racial differences in prevalence of VAs.

29

Mapping of a Blood Pressure QTL on Chromosome 17 in American Indians of the Strong Heart Family Study

Nora Franceschini (1) Ran Tao (2) Sue Rutherford (3) Lan Liu (4) Shelley A Cole (5) Laura Almasy (5) Harald HH Goring (5) Sandra Laston (5) Karin Haack (5) Lyle G Best (6) Kari E North (1)
(1) Epidemiology, University of North Carolina at Chapel Hill (2) Biostatistic, University of North Carolina at Chapel Hill (3) Biochemistry and Molecular Biology, Penn State University (4) Bostatistics, University of North Carolina at Chapel Hill (5) Genetics, Texas Biomedical Research Institute (6) Missouri Breaks Industries Research, Inc at Chapel Hill

Blood pressure (BP) is a complex trait, with a heritability of 30 to 40%. Recently identified BP genome wide association loci explain a small fraction of its phenotypic variation. Family studies can help gene discovery by utilizing trait and genetic transmission information among relative-pairs. We previously described a systolic BP quantitative trait locus (QTL) in Strong Heart Family Study (SHFS) American Indians at 17q25.3, a locus also reported in BP linkage studies of Europeans, African Americans and Hispanics. To follow-up persuasive linkage findings at this locus, we investigated the effect of common variants in the 1-LOD score interval region using a two-step strategy. We first genotyped a panel of 1,334 SNPs in 928 individuals belonging to families that showed strong evidence of linkage for BP. We then genotyped a second panel of 306 SNPs in all SHFS participants (N=3,807) for the most prominent associated genes in the region. We performed association analyses in the region. Three genes had multiple SNPs marginally associated with systolic BP (*TBC1D16*, *HRNBP3* and *AZI1*). Using the *Bayesian quantitative trait nucleotide* (*BQTN*) method to estimate the posterior probability (PP) that any variant in each gene had an effect on the phenotype, *AZI1* showed the most prominent findings (PP 0.66). Our findings provide some preliminary evidence for three genes at 17q25.3 for BP, which we are currently pursuing for replication in another large American Indian sample.

30

Race and Subtype Differences in the Replication of Previously Identified Breast Cancer Susceptibility Loci

Katie M O'Brien (1) Robert C Millikan (1)
(1) University of North Carolina at Chapel Hill

Genome-wide association studies (GWAS) and candidate gene analyses have led to the discovery of several dozen genetic polymorphisms associated with breast cancer susceptibility. While many of these loci are now considered well-established risk factors for the disease, discrepancies in risk by race and breast cancer subtype are poorly understood. Here, we estimated associations between 83 previously identified single nucleotide polymorphisms (SNPs) and white and African-American cases and controls using both frequentist and hierarchical methods. We also assessed whether these associations varied by breast cancer subtypes, as defined by 5 immunohistochemical markers. Twenty-six of the candidate SNPs replicated in whites ($p<0.05$), including several SNPs in *FGFR2*, *TNRC9/TOX3* and *MRPS30*. *FGFR2* and *TNRC9/TOX3* were also strongly associated with breast cancer in African-Americans, though only 14 of the 83 SNPs successfully replicated. Within subtype strata, *TNRC9/TOX3* SNPs were associated with luminal A and basal-like disease, while SNPs in *FGFR2* were associated with luminal A, human epidermal growth factor receptor positive/ estrogen receptor negative, but not non basal-like disease. The effects of several other candidate SNPs also varied by breast cancer subtype. As breast carcinogenesis may differ by race and tumor type, analyses such as these are necessary to identify race and subtype-specific genetic risk factors and advance our understanding of disease etiology.

31

Random Mating – Fact or Fiction? Evidence for Multigenerational Non-random Mating

Ronnie Sebro (1) Gina Peloso (2) Josee Dupuis (3) Neil Risch (1)
(1) Institute for Human Genetics, University of California San Francisco (2) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA (3) Department of Biostatistics, Boston University School of Public Health, Boston MA

The factors that influence spouse selection are important to geneticists because the mating pattern determines the genetic structure of a population. There has been evidence of positive assortative mating (PAM) due to several phenotypic traits like height. It has been shown theoretically that ancestrally-related PAM is necessary for population stratification, which means spouses are more likely to share genes of common ancestry. 885 Caucasian spouse pairs from the Framingham Heart Study (FHS) Original and Offspring cohorts genotyped on Affymetrix 500K were analyzed using principal component (PC) analysis. Data from the HapMap and Human Genome Diversity Project (HGDP) were used to identify the principal components. The first principal component separates North-Western from South Eastern Europeans, probably separating English/Irish/German from Italians. The second separates Europeans from Middle-Eastern Caucasians. There is strong correlation between the spouses' first (r=0.73, p=2e-22) and second (r=0.80, p=4e-29) principal components in the Original cohort. In the Offspring cohort the correlation between the spouses' first (r=0.38, p=3e-28) and second (r=0.45, p=9e-40) principal components are decreased but remain significant suggesting admixture. The magnitude of the ancestry correlations is higher than that between spouses' heights (r=0.27, p=2e-14) or weights (r=0.16, p=4e-6). Ancestry may be the most significant factor in spouse selection.

32
Variation in Heritability due to Population Stratification and Implications for Heritability Studies
Ronnie Sebro (1) Neil Risch (1)
(1) Institute for Human Genetics

The heritability of a trait is the ratio of the additive genetic variance to the total variance of the trait. Comparison of the correlation between monozygotic (MZ) and dizygotic (DZ) twin trait values has been used to estimate the heritability of a trait to assess whether a trait is monogenic or polygenic. These calculations assume a random mating population. It has been recently shown that ancestrally-related positive assortative mating (PAM) results in population stratification. Unlike other forms of PAM which are based on phenotypic similarity between spouse-pairs, ancestrally-related PAM occurs between spouses with similar genetic ancestries. The aim of this study is to assess how stratification affects trait heritability and the correlation between relatives. First a generalization for calculation of the mating type frequencies at a single nucleotide polymorphism in a population with structure was derived. The trait values for genotypes AA, AB and BB were assumed to be a, d and -a respectively. Theoretical calculations of the heritability and correlation between trait values of MZ, DZ, parent-offspring and second degree relative-pairs were then calculated. If there is no dominance variance and population structure, then the heritability of the trait increases, primarily due to an increase in the additive genetic variance of the trait. These findings suggest population stratification may explain some of the variability in heritability estimates seen across studies.

33
Clustering of Crohn's Disease Patients: Identification of Sub-phenotypes and Population Stratification
Barbel Maus (1) Emmanuelle Genin (2) Jestinah M Mahachie John (1) Kristel Van Steen (1)
(1) University of Liege (2) Inserm

The identification of genetically based patient subgroups is useful to develop personalized medicine and to reclassify diseases. Cleynen et al. [2010] applied latent class analysis (LCA) to detect different sub-phenotypes in Crohn's disease (CD) cases. However, due to the unavailability of sufficient null markers, the overlap between the identified case subgroups and population strata could not be assessed. Systematic differences in allele frequencies between population subgroups can bias the definition of case subgroups. Here, we analyzed available case-control data on CD to discover genetic case subgroups. Using a latent class analysis we identified several clusters within CD cases. We then compared the clustering results to commonly applied population stratification techniques i.e. structured association and principal component analysis (PCA). Some similarities were found between the results of the LCA and the PCA with regard to the indicated structure and influential SNPs. Thus, it seems necessary to correct for population stratification in a cluster analysis of patients.
The inclusion of principal components (PCs) calculated on genome wide data as covariates in a latent class model on disease associated SNP may present a possibility to correct for population stratification. Therefore in a further analysis, PCs were included as covariates and the resulting case clusters were compared to the earlier obtained ones.
Cleynen et al. 2010. PLoS One 5(9):e12952.

34
Efficient Association Analysis for Groups of Genetic Markers that Avoids Confounding by Genetic Structure
Jennifer Listgarten (1) Christoph Lippert (1) David Heckerman (1)
(1) Microsoft Research

Approaches for testing groups of variants for association with complex traits are becoming critical. Examples of groups typically include a set of rare or common variants within a gene, but could also be variants within a pathway or any other set. These tests are important for aggregation of weak signal within a group, allow interplay among variants to be captured, and also reduce the problem of multiple hypothesis testing. Unfortunately, these approaches do not address confounding by, for example, family relatedness and population structure, a problem that is becoming more important as larger data sets are used to increase power. We introduce a new approach for group tests that can handle confounding, based on Bayesian linear regression, which is equivalent to the linear mixed model. The approach uses two sets of covariates (equivalently, two random effects), one to capture the group association signal and one to capture confounding. We also introduce a computational speedup for the two-random-effects model that makes this approach feasible even for extremely large cohorts, whereas

it otherwise would not be. Application of our approach to richly structured GAW14 data, comprising over eight ethnicities and many related family members, demonstrates that our method successfully corrects for population structure, and application of our method to WTCCC Crohn's disease and hypertension data demonstrates that our method recovers genes not recoverable by univariate analysis.

35

Principal Component Analysis Corrects for Population Stratification in Studies of Gene-Environment Interactions

Elena Viktorova (1) Melanie Sohns (1) Heike Bickeboeller (1)
(1) Department of Genetic Epidemiology, University Medical Center, Georg-August University of Goettingen, Goettingen, Germany

Uncovered population stratification (PS) in large scale genetic association studies may lead to false positive results or masks true association signals via under(over)estimation of the effects in the analysis that fails to account for it. This is well studied for main effects. The extent of PS bias for GxE interactions depends on specific characteristics of the cohort, particularly on the number of admixed ethnicities, differences in genotype, exposure frequencies and baseline disease risks across the strata. We considered admixture of two or more discrete subpopulations. PS bias for interactions appears to be the largest when there are two subgroups. We investigated the magnitude of the bias due to PS for GxE interactions in case-control, case-only, Albert's and Murkray's two step, Mukherjee empirical Bayes and empirical hierarchical Bayes analysis. Our simulations show that PS bias can reach intolerable level in all methods, especially for the case-only design. Importantly, bias for GxE interactions is greater compared to main effects, when large variation of exposure frequencies is observed across the strata. Therefore, we propose principal component analysis (PCA) to correct for the hidden PS and to improve precision of GxE effect estimates. We applied PCA in our simulations and showed that in all scenarios PCA reduces bias to almost zero. We conclude that PCA adjustment yields a good appropriate correction not only for main effects but also for GxE interaction analysis.

36

Rare Variants in Adjusting for Population Structure

Omar De la Cruz Cabrera (1) Paola Raska (1)
(1) Case Western Reserve University

The characterization of population structure from genomewide genotype data is important when adjusting for the effect of the structure in a genetic association test. A very useful approach to the study of population structure is the use of multivariate statistical methods, like Principal Components Analysis (PCA). It turns out that, due to the high dimensionality of genotype data, the statistical behavior of these methods is similar to that produced by the use of kernel methods. We have found that, just like the result of a kernel-based procedure depends on the bandwidth used, the inferred population structure can change depending on the population frequency of the variants used.

We investigate the effect on the inferred structure of using sets of variants of approximately the same frequency. The underlying hypothesis is that rare variants, which tend to be shared only between individuals with more recent common ancestry, will tend to reveal structure at a smaller scale, while common variants, which can be shared by unrelated individuals, reveal differences at a larger scale. The structures found at different scales can be quite different.
We perform simulations based on data from the 1000 Genomes project, and use some theoretical considerations to explain some of the effects found, and describe how the adjustment for population structure in GWA studies is affected by the choice of variants, especially when testing rare variants.

37

BioBin: A Bioinformatics Tool for Biologically Inspired Collapsing of Rare Variants

Carrie B Moore (1) John R Wallace (2) Alex T Frase (2) Sarah A Pendergrass (2) Marylyn D Ritchie (2)
(1) Vanderbilt University (2) Pennsylvania State University

There has been increasing interest in rare variants (RVs) and methods to detect their association to disease. We have developed a powerful, flexible collapsing method inspired by biological knowledge using database resources such as NCBI, KEGG, GO, PharmGKB, Entrez, ECRbase, and ORegAnno. Variants can be collapsed according to functional regions, evolutionary conserved regions, regulatory regions, genes, and/or pathways without the need for external files. We conducted a pairwise comparison of RV burden differences (MAF <0.03) between ancestry groups (15 populations) of the 1000 Genomes Project and found marked differences. For example, between Yoruba (YRI) and European descent (CEU) individuals, we found that 56.8% of gene bins, 78.9% of intergenic bins, and 83.5% of pathway bins have significant differences in RV burden. Comparing two related populations, YRI and ASW (African ancestry in Southwest USA), we identified 0.39% gene bins, 0.60% of intergenic bins, and 0.98% of pathway bins had significant differences in RV burden. Ongoing efforts are examining regional characteristics using BioBin (evolutionarily conserved and/or regulatory). We are also exploring ancestry correction, the magnitude of potential stratification in RVs is large and inadequate correction could have large implications for sequence data analysis. BioBin is a useful, powerful, and flexible tool in analyzing sequence data and will be successful at uncovering novel associations with complex disease.

38

Identification of Grouped Rare and Common Variants via Penalized Logistic Regression

Kristin L Ayers (1) Heather J Cordell (1)
(1) Institute of Genetic Medicine, Newcastle University, UK

In spite of the success of genome-wide association studies in finding many common variants associated with disease, these variants seem to explain only a small proportion of the estimated heritability. Data collection has turned toward exome and whole genome sequencing, but it is well known that frequently used single marker methods have low power to detect rare variants associated with disease,

even with very large sample sizes. In response, methods have been developed which attempt to cluster rare variants so that they may gather strength from one another under the premise that there may be multiple causal variants within a gene. Most of these methods group variants by gene or proximity, and test one gene or marker window at a time. We propose a penalized regression method that analyzes all genes at once, allowing grouping of all (rare and common) variants within a gene, along with subgrouping of the rare variants, thus borrowing strength from both rare and common variants within the same gene. In simulations, our method performs favorably when compared to many previously-proposed approaches, including its predecessor, the sparse group lasso,[1,2] although we find SKAT[3] generally outperforms our method.

[1]Friedman J, Hastie T, Tibshirani R. Technical report, Department of Statistics, Stanford University. 2010
[2]Zhou H, Sehl ME, Sinsheimer JS, Lange K. Bioinformatics 2010; 26: 2375–2382.
[3]Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. AJHG 2011; 89:82–93.

39

Detecting Association for Low-frequency Variants in Case-control Studies

Chang-Yun Lin (1) Guan Xing (2) Chao Xing (1)
(1) University of Texas Southwestern Medical Center (2) Bristol-Myers Squibb

In testing association between genetic variants and a disease, there exist popular metrics such as odds ratio and Phi coefficient to measure the strength of association, and there exist popular tests such as Pearson's chi-squared test, Fisher's exact test, and the likelihood ratio test to test the strength of association. However, these measures and tests are not sensitive to detect association for low-frequency exposures. In this study, we propose a new statistic for the purpose of detecting low-frequency variants associated with the disease. The association test based on the new statistic is more sensitive than existing methods in detecting association for low-frequency variants. We are able to numerically prove that the new statistic offers higher power while maintains the same type I error rate as Pearson's chi-squared test in a balanced study design. Application of the new method in genetic studies of detecting association of rare genetic variants with diseases is shown.

40

Estimating and Testing Genetic Effects for Complex Traits in Sequence-Based Association Studies and Power Comparisons

Jin J Zhou (1) Nan M Laird (1)
(1) Harvard School of Public Health

Genetic studies have been haunted by the mystery of "missing heritability". Although genome-wide association studies (GWAS) have identified many variants associated with some common diseases and traits, these variants typically explain only a small fraction of the heritability. With emergence of the next generation sequencing (NGS) technology, many novel statistical methodologies have been proposed to assess the contribution of rare variations to complex disease etiology. In this paper we explore the application of linear mixed model (LMM), which has been recently used to estimate genetic variance using variants in current GWAS platform, to the estimation and testing of genetic variance based on SNP panels of NGS data. In general, LMM provides an association testing strategy that can detect both rare and common variants, deal both additive and interaction effects, handle both quantitative and dichotomous traits, and incorporate non-genetic covariates. We explore the theoretical connection between LMM and kernel-based association testing such as SKAT and carry out extensive simulation studies to evaluate the power of LMM under various distribution assumptions of traits, proportions of the causal variants and the quality control (QC) stringent thresholds. Our studies show superior performance of LMM under various conditions compared to other existing methods.

41

Genome-wide, Permutation-based Rare Variant Association Analysis with INTERSNP-RARE

Dmitriy Drichel (1) Andre Lacour (1) Christine Herold (1) Tatsiana Vaitakhovich (2) Vitalia Schueller (2) Tim Becker (1)
(1) DZNE – German Center for Neurodegenerative Diseases (2) Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn

INTERSNP-RARE is an implementation of three rare-variant testing procedures and their respective variable threshold (VT) extensions: CMAT (cumulative minor allele test), COLL (collapsing test) and FR (Fisher-Rare), a variation of the Fisher combination test. All tests operate on a number of rare (MAF=<MAF_T) variants in physical proximity (bins).
Deciding on a MAF_T choice of binning is not straightforward in genome-wide data. Distribution of 'small' MAFs across the frequency spectrum and physical distance has to be taken into account. Ideally, resulting bins are non-overlapping and consist of a 'reasonable' number of rare variants.
Bins can be generated using either an algorithmic method or based on a priori information on genomic intervals (genes, LD blocks, results from a rare variant association analysis . . .) that can be modified using a combination of interval modification and QC-procedures to create a set of intervals suitable for binning.
We conducted a case-control power simulation study on bins containing 20–60 rare SNPs using disease models with different proportions of protective and causal variants. The single-marker analysis outperforms other approaches in a few scenarios with large MAF thresholds and few causal markers (∼10%), while CMAT and COLL are well-powered with ∼30–50% damaging and up to 20–30% protective variants. FR offers superior performance and robustness in most modes with many protective variants and >10% causal SNPs.

42

Statistical Tests for Disease Association with Rare Variants in Next-generation Sequence Data

Saonli Basu (1) Wei Pan (1)
(1) University of Minnesota

With the advent of high-throughput sequencing technologies, there is an increasing interest for developing methods

to detect association between complex traits and rare variants (RVs). It has been argued that collections of rare variants hold promise as a source of heritability, which is not explained by common variants. Several new tests have been proposed to analyze RVs, most of which are based on the idea of pooling or collapsing the RVs. Here we propose a partitioning model for the detection of joint effect of a group of variants, where we classify the RVs into groups based on their direction of association with the trait. We adopt a sequential search for classification and selection of RVs from a group of variants, which significantly reduces the computational cost and cost due to multiple testing adjustment. We propose a test with flexible scoring schemes to capture interaction among the variants, while maintaining the same degrees of freedom of the test. We illustrate and compare our model with several existing approaches through extensive simulations. We consider both independent and correlated RVs; causal and non-causal RVs and representative tests for both common variants and rare variants and provide a comprehensive comparison of various statistical tests using simulated data under different alternatives.

43

Incorporating the Gene Genealogy in Rare Variant Mapping Methodology

Kelly M. Burkett (1) Brad McNeney (2) Celia M.T. Greenwood (1) Jinko Graham (2)
(1) Lady Davis Institute, Jewish General Hospital; McGill University (2) Simon Fraser University

For understanding genetic associations with disease outcomes, it is useful to model the latent gene genealogies that give rise to the sample's genetic variability. To this end, we consider the genealogy of a target locus in genetic sequences from a random sample of unrelated individuals. Though the true genealogy is unknown, we can model its distribution conditional on the observed genotype data. However, to sample from this distribution requires Monte Carlo methods. We describe our implementation of a sampler that uses Markov chain Monte Carlo to sample genealogies conditional on phased or unphased genotype data and its application to mapping disease predisposing variants. In particular, we focus on the application of the genealogy sampler to the discovery of rare variants. If many rare, disease-predisposing variants are present in a genomic region, we expect individuals with disease to appear together in clusters near the tips of the genealogical tree of that region. We can measure the degree to which those affected with disease cluster on the sampled trees relative to those who are not affected with disease. We evaluate existing and novel tree-based statistics that measure case clustering using both simulated and real datasets.

44

Deep Targeted Sequencing of 12 Breast Cancer Loci in 4,700 Women Across Four Different Ethnicities

Sara Lindstrom (1) Brad Chapman (1) Gary Chen (2) Constance Chen (1) Oliver Hofmann (1) Daniel B Mirel (3) Christopher A Haiman (2) Peter Kraft (1)
(1) Harvard School of Public Health (2) University of Southern California (3) Broad Institute of Harvard and Massachusetts Institute of Technology

Genome-wide association studies (GWAS) have identified multiple loci associated with breast cancer risk. However, the underlying genetic structure is not fully understood and it is likely that the GWAS signal originates from one or more as yet unidentified causal variants. We used next-generation sequencing to characterize 12 GWAS-discovered breast cancer loci in a total of 2,335 breast cancer cases and 2,365 controls across four ethnic populations. Our primary aims were to identify sets of putative causal alleles and assess whether these regions are enriched for rare variants in cases (or controls). Region boundaries were defined by the nearest recombination hotspot downstream and upstream from the index GWAS signal. In total we sequenced 5,500 kb. On average, we captured 82% of the non-repetitive sequence in the targeted regions, and the average fraction of captured bases sequenced with a depth >20x was over 98%. Single Nucleotide Variant (SNV) genotypes were called using the GATK pipeline and were over 99.5% concordant with GWAS data. For a subset of samples, we also have access to data from the Illumina HumanExome Beadchip allowing us to calculated concordance rates for rare non-synonymous variants as well. We will present association results across ethnicities with emphasis on functional variants such as non-synonymous, regulating and splicing variants. We will discuss practical issues in targeted sequencing and the importance of thorough quality control procedures.

45

Performance of Statistical Tools on Testing CHARGE-S Targeted Sequencing Data

Chuanhua Julia Xing (1) Josee Dupuis (1) L. Adrienne Cupples (1)
(1) Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA; Framingham Heart Study, Framingham, MA

The CHARGE-S (Cohorts for Heart and Aging Research in Genomic Epidemiology-Sequencing) project is a collaborative effort from Framingham Heart Study (FHS), Cardiovascular Health Study (CHS), and Atherosclerosis Risk in Communities (ARIC). CHARGE-S used a case-cohort based design, whereby a random sample of study participants is enriched by subjects in extremes of traits. Statistical tools were developed to investigate the role of rare variants, upon the observation that a large portion of rare variants emerged from sequencing data. Here, we tested several including SKAT, Score-Seq and Step-Up. Using genotypes from CHARGE-S targeted-sequencing data for FHS (n=1096), we simulated phenotypes to evaluate type I error and power in unrelated individuals. We derived phenotypes in a large sample for 11 correlated traits, mimicking traits used to select targets. We sampled individuals from this population to simulate the study design in CHARGE-S. We evaluated type I error and power for all regions. For variants with minor allele frequency (maf) < 1%, we observed proper type 1 error in all regions for SKAT, except for inflation in 3 regions. Observed power for this sample size was generally low, but larger in some regions. Power is expected to improve when the threshold for selecting rare variants increases (eg. 5%), if variants in the 1% to 5% maf range are causal. We provide some guidelines on the performance of some popular methods to detect rare variants.

46

Comparison of Two Next-generation Sequencing Technologies on the Genomes of a Trio Family

Remi Kazma (1) Eric Jorgenson (2)
(1) Department of Epidemiology and Biostatistics and Institute for Human Genetics, University of California San Francisco (2) Department of Neurology, Ernest Gallo Clinic and Research Center, University of California San Francisco

Next-generation sequencing provides a new level of scrutiny on the human genome, and its rapidly decreasing cost is making the investigation of the effect of rare genomic variants on complex diseases and traits increasingly affordable. Yet, our knowledge of the relative quality of the various competing technologies is scarce, and only one comparison of a single individual has been published to date (Lam et al., Nature Biotech 2012;30:78–82). Here, for the first time, we compare the two most widely used sequencing technologies, Illumina and Complete Genomics using the genomes of a trio family. We sequenced the genomes of two parents and one male child from Costa Rica using two technology pipelines: Complete Genomics version 1.10.0.28 and Illumina version 1.3.0.0 (MARS). Both pipelines were also used to call single nucleotide variants and small insertion/deletions using the reference genome (NCBI build 37) for alignment. Additionally, the Complete Genomics pipeline called copy number variants and larger structural variants, and provided gene annotations. The average read depth was about 55x for Complete Genomics and 50x for Illumina. Furthermore, the three genomes were genotyped on two Illumina chips: HumanHap650y and HumanOmni 2.5, providing another level of comparison. The use of a trio family for this comparison allowed for crosschecking calls and better estimation than a single genome of the coverage, concordance, and false positives of both pipelines.

47

ATOMIC – Assess Genotype Calling Quality Using R or Affymetrix' Genotyping Console Software

Andreas Ziegler (1) Arne Schillert (1)
(1) Institut fur Medizinische Biometrie und Statistik, Universitat zu Lubeck

Single nucleotide polymorphism (SNP) chips are used for the analysis of exomes, targeted SNP typing, and genome-wide association analysis. High data quality is required for especially for SNP chips focusing on rare variants because misclassified genotypes for only a few individuals can result in highly inflated type I errors. Several statistical approaches, such as collapsing methods for rare variants, gene-gene interaction studies, or data mining approaches also require extremely clean input data to avoid false positives. The best approach for quality control is the investigation of signal intensities, but this approach is computationally intensive and has therefore been limited to SNPs with positive association signals. We present the R package ATOMIC, an acronym for AuToMatic Inspection of Cluster plots. Its main features are 1) splitting of signal intensities into smaller, manageable datasets, 2) assessment of the clusterings utilizing the theory of cluster analysis, including tests for unimodality [1,2], an improved implementation of the ACPA [3] algorithm, and 3) modern graphical display of the resulting signal intensity plots [4]. The approaches implemented in the R package will also be available soon in Affymetrix' genotyping console software.

[1] Larkin 1979, Behav Res Meth Instr 11:467–468
[2] Engelman and Hartigan 1969, J Am Stat Assoc 64:1647–1648
[3] Schillert et al. 2009, BMC Proc 3:S58
[4] Wickham, 2009, Springer New York

48

Assessing Error Rates for Low maf SNPs on the HumanExome Beadchip Array

Elizabeth W Pugh (1) Hua Ling (1) Jane M Romm (1) Ivy A McMullen (1) Jeroen R Huyghe (2) Michael Boehnke (2) Kimberly F Doheny (1)
(1) Center for Inherited Disease Research, Johns Hopkins University (2) Department of Biostatistics and Center for Statistical Genetics, University of Michigan

We compared 4 measures of error for very rare, rare and more common minor allele frequency SNPs on the Illumina HumanExome Beadchip Array, reproducibility, Mendelian inheritance, consistency with next genome sequence data and manual review of SNP cluster plots. We used GenomeStudio version 2011.1, Genotyping Module version 1.9.4, GenTrain Version 1.0 to cluster the 87 unrelated HapMap samples and one pair of HapMap duplicates together with 4111 randomly selected investigator samples from a large exome project of over 11,000 European ancestry samples. This created a 4200 sample project, which is our median project size for exome content array projects. We then applied the clustering from the 4200 samples to 142 additional HapMap samples, run with the project and computed reproducibility and Mendelian inheritance rates. We manually reviewed cluster plots for subsets of the SNPs to determine the proportion that appeared poorly clustered such that a genotype would be misspecified. We also compared the genotypes for 62 of the 87 samples which were present in the 1000 Genomes Sequencing Project (Oct 2011 release).

49

Genetic Simulation Resources (GSR): A Website for the Registration and Discovery of Genetic Data Simulators

Bo Peng (1) Ben Racine (2) Huann-Sheng Chen (3) Leah Mechanic (3) Lauren Clarke (2) Elizabeth Gillanders (3) Eric Feuer (3)
(1) The University of Texas MD Anderson Cancer Center (2) Cornerstone Systems NorthWest, Inc. (3) Division of Cancer Control and Population Sciences, National Cancer Institute

Many genetic simulation programs have been developed to simulate evolutionary processes under realistic ecological and genetic scenarios and generate genetic data resulting from such processes. Such simulations have been used, for example, to predict properties of populations retrospectively or prospectively according to mathematically intractable genetic models, and to assist the validation, statistical inference and power analysis of a variety of statistical models. However, due to the differences in type of genetic data of interest, simulation methods, evolutionary features, input and output formats, terminologies and assumptions for different applications, choosing the right tool for a particular study can be a troublesome process that usually involves searching, downloading and testing many

different tools. Here we present Genetic Simulation Resources (http://popmodel.nci.nih.gov/geneticsimulation), a web service provided by the National Cancer Institute that aims to help researchers compare and choose the right simulation tools for their studies. This service allows authors of simulation software to register their applications and describe them with well-defined attributes, and users to search and compare simulators according to specified features. Activities such as citations, updates from authors, and number of visitors are tracked and provide updated information about the applications to help users weed through a large number applications.

50

compreheNGSive – A Visualization Tool for Prioritizing Variants from Next Generation Sequence Data

Alex Bigelow (1) Miriah Meyer (2) Nicola J Camp (3)
(1) University of Utah Division of Genetic Epidemiology, Scientific Computing and Imaging Institute (2) University of Utah Scientific Computing and Imaging Institute (3) University of Utah Division of Genetic Epidemiology

In complex traits, the underlying critical risk variants will likely include some involved in regulation and/or that influence risk via novel mechanisms. Such variants will not be conducive to discovery using more standard annotation techniques or formal algorithm-based methods based on well-described mechanisms. Simple filtering on differences in allele frequencies between groups and sharing within groups will be useful in identifying these more obscure risk variants; however, black-box hard-filters may inadvertently remove variants of potential importance. Data visualization can be extremely useful for informed heuristic prioritization. The observed structure of the data can instruct and guide prioritization of variants. We have developed an interactive visualization tool, *compreheNGSive*, to support investigation and prioritization of next generation sequence variants. The tool requires .VCF file/s for variants and, optionally, additional variant-level data (e.g. annotations, correlations with known associated risk variants) in columnar text format and/or feature-level data in .BED or .GFF file format. The viewer includes a scatterplot, parallel coordinates and a genome browser, each with interchangeable axes, variant selection mechanisms, and methods for coping with missing data. The software is written using Python and the Qt framework, which is compatible with Windows, Linux, OS X, and potentially mobile operating systems and is fully scalable to whole genome data.

51

Biofilter 2.0 for Advanced Model Development, Testing, and Hypothesis Generation Using Expert Domain Knowledge Resources

Sarah A Pendergrass (1) Alex Frase (1) John Wallace (1) Carrie Moore (2) Neerja Katiyar (1) Marylyn D Ritchie (1)
(1) The Pennsylvania State University (2) The Pennsylvania State University, Vanderbilt University

Leveraging the wealth of biological information collected from the genome, transcriptome, proteome, and other -omic data types, is a key element for advanced predictive model development. Biofilter is a software tool for using biological information from databases to direct analyses. Biofilter 1.0 was specifically for use with gene or single-nucleotide polymorphism based data and used 7 database sources: the National Center for Biotechnology (NCBI) dbSNP and gene databases, the Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, Gene Ontology (GO), the Protein families database (Pfam), and NetPath – signal transduction pathways. Biofilter 2.0 now includes 5 additional sources: the Molecular INTeraction database (MINT), the Biological General Repository for Interaction Datasets (BioGrid), the Pharmacogenomics Knowledge Base (PharmGKB), the database of Evolutionary Conserved Regions (ECRBase), and the Open Regulatory Annotation Database (ORegAnno). Biofilter also accepts a greater range of data types, such as SNP, single-nucleotide variant (SNV), rare variant, copy-number variation (CNV), and evolutionary conserved region (ECR) data. Biofilter provides a flexible way to use the ever-expanding expert biological knowledge that exists to direct complex predictive model development for elucidating the etiology of complex phenotypic outcomes. The software is freely available for non-commercial use at *http://ritchielab.psu.edu/*.

52

A New Approach to Maximally Select Unrelated Individuals for Genetic Analysis

Jennifer E Below (1) Jeffrey Staples (1) Deborah A Nickerson (1)
(1) The University of Washington

Many statistical analyses of genetic data rely on the assumption of independence among samples. Consequently, relatedness is either modeled in the analysis or samples are removed to "clean" the data of any pairwise relatedness above a tolerated threshold. Current methods do not maximize the number of unrelated individuals retained for further analysis, and this is a needless loss of resources. We report a novel application of graph theory that identifies the maximum set of unrelated samples in any dataset given a user-defined threshold of relatedness, as well as all networks of related samples. We have implemented this method into an open source program, PRIMUS. We show that PRIMUS outperforms three existing methods, allowing researchers to retain up to 50% more unrelated samples. A unique strength of PRIMUS is its ability to weight the maximum clique selection using additional criteria (e.g. affected status and data missingness). Similar to the maximum unrelated set identification, PRIMUS always identifies the largest set of unrelated affected individuals. PRIMUS retained up to 75% more affected individuals in the weighted comparisons between PRIMUS and each of the other methods. PRIMUS is a permanent solution to identifying the maximum number of unrelated samples for a genetic analysis.

53

Two-Phase Design to Follow-up Genome Wide Association Signals with DNA Resequencing Studies

Daniel J Schaid (1) Gregory D Jenkins (1) James N Ingle (1) Richard M Weinshilboum (1)
(1) Mayo Clinic

Genome wide association studies (GWAS) of complex traits have generated many association signals, in the form of

extreme p-values of single nucleotide polymorphisms (SNPs). To understand the underlying causal genetic variant(s), focused DNA resequencing of targeted genomic regions is commonly used, yet the current cost of resequencing per sample limits sample sizes for resequencing studies. Information from the large GWAS can be used to guide choice of samples for resequencing, such as the SNP genotypes in the targeted genomic region. Study design questions were motivated by a GWAS of breast cancer occurrence among high-risk women treated with drugs to prevent breast cancer. Viewing the GWAS tag-SNPs as imperfect surrogates for the underlying causal variants, yet expecting that the tag-SNPs are correlated with the causal variants, a reasonable approach is a two-phase case-control design, with the GWAS serving as the first-phase and the resequencing study serving as the second-phase. Using stratified sampling based on both tag-SNP genotypes and case-control status, we explore the gains in power of a two-phase design relative to randomly sampling cases and controls for resequencing (i.e., ignoring tag-SNP genotypes). Simulations are be used to illustrate the strengths and limitations of this approach in the context of genomic studies, particularly the impact of linkage disequilibrium between tag-SNPs and causal SNPs.

54

Genetic Case-control Matching Strategies via Bipartite Graphs in Genome-wide Association Studies

Andre Lacour (1) Tim Becker (1)
(1) German Center for Neurodegenerative Deseases (DZNE), Bonn, Germany

Population stratification in samples of GWAS give rise to large obliterations in the results of statistical tests. In order to correct for stratification effects we have developed three strategies for structured association testing. Taking the IBS matrix as a measure for genetic kinship, we obtain pairwise case-control matchings, groupings with respectively at least one control and one case, and clusters. All strategies are based on a 'maximum weighted bipartite matching' by making use of the hungarian algorithm which solves the assignment problem in polynomial time. We also present a set of quality control strategies on the found matchings.
Association P-values are obtained by within-structure case-control permutations. As it turns out from simulation studies, the empirical niveau for within-structure permutation becomes close to the nominal level. Thus, our methods considerably reduces false-positive rates compared to unstratified analyses and possibly leads to an increase of power.
The matching strategies can be performed both genome-wide and localized. Local window sizes of a few thousand SNPs are enough to guarantee identification of strata. As a byproduct, our implementation strongly outperforms common covariate approaches in runtime, and makes genome-wide application possible. Our method for stratified analyses is implemented in the genome-wide interaction analysis software INTERSNP.

55

A Theoretical Comparison of the Sample Size Requirements for Obtaining Equivalent Powers in Case-control and Trio Designs

Tanushree Haldar (1) Saurabh Ghosh (1)
(1) Indian Statistical Institute

Case control studies have been the traditional method of choice for association mapping of complex genetic traits. However, the Transmission Disequilibrium Test (TDT) has turned out to be a suitable alternative since it circumvents the problem of population stratification encountered in case-control studies. The aim of this study is to develop a theoretical comparison of the sample size requirements in the two study designs to obtain equivalent power to detect association in the absence of population stratification. We derive the distributions of both the test statistics under the alternative hypothesis for different genetic models and hence, determine the number of cases (or controls) in a case-control design and the number of transmissions from heterozygous parents to affected offspring in a trio design that yield a pre-assigned power. We also estimate the number of trios to be sampled to get the required number of informative transmissions. We find that for a given power, the larger the value of the coefficient of the linkage disequilibrium between the disease locus and the marker locus, the smaller is the number of cases (or controls) required in a case-control design compared to the number of transmissions required in the TDT design. Our computations clearly show that in order to attain a given power, the TDT design requires a larger sample compared to the case-control design.

56

Detecting the Remaining 80% of Genome-wide Associations with Improved Microarray Coverage and Larger Sample Sizes

Karla J Lindquist (1) Eric Jorgenson (1) Thomas J Hoffmann (1) John S Witte (1)
UCSF

In the previous five years, genome-wide associations studies (GWAS) have identified many genetic variants associated with complex traits, contributing to our understanding of the biology and heritability underlying these traits and to advances in medicine. Recently, microarrays that improve coverage of variants in diverse populations have been developed, and declining genotyping costs allow for larger samples. We estimate the extent to which these improvements can lead to the detection of additional associations by GWAS. We focus on single nucleotide polymorphisms (SNPs) with minor allele frequencies of >1% underlying complex diseases. Overall, we found that *GWAS have detected less than one fifth of all common SNPs underlying disease.* Microarrays over four times larger than those used by previous GWAS could only detect about one quarter of these associations without increasing sample sizes. Conversely, quadrupling sample sizes alone would allow for the detection of about three quarters of these associations. Next-generation sequencing may be required to find the rare variants underlying complex disease, and polygenic, epigenetic, and other models will be required to explain all heritability. However, our results show that many novel SNP-disease associations can still be detected by future GWAS.

57

Power Considerations for a GWA of Linear Mixed Effects Models

Kelly S Benke (1) YanYan Wu (1) Lyle J Palmer (2)

(1) Samuel Lunenfeld Research Institute (2) Ontario Institute for Cancer Research

Tools to compute power for binary and quantitative traits for a cross-sectional GWA are widely available, as there are closed form solutions for these types of data. There is no closed form solution, however, to compute power when using linear mixed models for quantitative, repeated measures traits. Additionally, interest for a GWA that employs the linear mixed model often centers on both the main SNP effect as well as the SNP-by-time effect. We show in principle that interpreting both the SNP and the SNP-by-time effects at an alpha threshold established for the interpretation of a single SNP parameter is inflated, as would be expected. Further, we use simulation (mimicking a total cholesterol trait in 1269 older adults) to compute power when controlling the false positive rate by 1) employing a Bonferroni correction, and 2) using a two degree of freedom likelihood ratio test (2dfLRT). We show that in situations where both a SNP and SNP-by-time effect are present, the 2dfLRT is more powerful (the 2dfLRT 18% to 92% percent more powerful compared to the Bonferroni correction across the range of effect sizes that we explored). In situations where only a SNP or SNP-by-time effect is present, the 2dfLRT is only slightly less powerful, such that the Bonferroni correction never realizes a power gain of more than 6 percent. Given these findings, we recommend the 2dfLRT in a hypothesis-free GWA setting.

58

Polygenes and Estimated Heritability of Prostate Cancer in an African American Sample Using GWAS Data

Jing He (1) Gary K Chen (1) William J Blot (2) Sara S Strom (3) Sonja I Berndt (4) Rick A Kittles (5) Benjamin A Rybicki (6) William Isaacs (7) Sue A Ingles (1) Janet L Stanford (8) Ryan W Diver (9) John S Witte (10) Ann W Hsing (4) Barbara Nemesure (11) Timothy R Rebbeck (12) Kathleen A Cooney (13) Jianfeng Xu (14) Adam S Kibel (15) Jennifer J Hu (16) Esther M John (17) Serigne M Gueye (18) Stephen Watya (19) Lisa B Signorello (2) Richard B Hayes (20) Zhaoming Wang (4) Lisa W Chu (4) Eric A Klein (21) Phyllis Goodman (22) Edward Yeboah (23) Yao Tettey (23) Qiuyin Cai (24) Suzanne Kolb (8) Elaine A Ostrander (25) Charnita Zeigler-Johnson (12) Yuko Yamamura (3) Christine Neslund-Dudas (6) Jennifer Haslag-Minoff (15) William Wu (15) Venetta Thomas (26) Glenn O Allen (26) Adam Murphy (27) Bao-Li Chang (12) Lilly S Zheng (14) Cristina M Leske (11) Suh-Yuh Wu (11) Anna M Ray (13) Anselm JM Hennis (11) Michael J Thun (9) John Carpten (28) Graham Casey (1) Stephen J Chanock (4) Brian E Henderson (1) Christopher A Haiman (1) Daniel O Stram (1)

(1) Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center (2) International Epidemiology Institute, Rockville, MD (3) Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, TX (4) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD (5) Department of Medicine, University of Illinois at Chicago, Chicago, IL (6) Department of Biostatistics and Research Epidemiology, Henry Ford Hospital, Detroit, MI (7) James Buchanan Brady Urological Institute, Johns Hopkins Hospital and Medical Institutions, Baltimore, MD (8) Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA (9) Epidemiology Research Program, American Cancer Society, Atlanta, GA (10) Institute for Human Genetics, Departments of Epidemiology and Biostatistics and Urology, University of California, San Francisco (11) Department of Preventive Medicine, Stony Brook University, Stony Brook, NY (12) University of Pennsylvania School of Medicine and the Abramson Cancer Center, Philadelphia, PA (13) Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI (14) Center for Cancer Genomics, Wake Forest University School of Medicine, Winston-Salem, NC (15) Division of Urology, Department of Surgery, Washington University, St. Louis, MO (16) Sylvester Comprehensive Cancer Center and Department of Epidemiology and Public Health, Univ of Miami Miller School of Medicine (17) Cancer Prevention Institute of California, Fremont, CA (18) Hopital General de Grand Yoff, Dakar, Senegal (19) Mulago Hospital/Makerere University, Department of Surgery, Urology Unit, Kampala, Uganda (20) Division of Epidemiology, Department of Environmental Medicine, New York University Langone Medical Center, New York, NY (21) Glickman Urological and Kidney Institute, Cleveland Clinic, Cleveland, OH (22) The Fred Hutchinson Cancer Research Center, Seattle (23) University of Ghana Medical School, Legon, Ghana (24) Div. of Epidemiology, Dept. of Medicine, Vanderbilt Epidemiology Ctr., Vanderbilt Univ. and the Vanderbilt-Ingram Cancer Ctr. (25) Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD (26) Sylvester Comprehensive Cancer Ctr. and Dept. of Epidemiology and Public Health, Univ. of Miami Miller School of Medicine (27) Department of Urology, Northwestern University, Chicago, IL (28) The Translational Genomics Research Institute, Phoenix, AZ

There is a gap between the phenotypic variance explained by Genome Wide Association Study (GWAS) hits and those estimated from classical methods. Here we adopted score analysis (Purcell et al) and linear mixed effects model (Yang et al) to estimate narrow sense heritability of Prostate Cancer in African Americans, trying to find the missing heritability in GWAS. In score analysis, two independent populations were used as discovery and target samples; each locus is tested in the discovery sample, the variation across nominally associated loci is summarized into quantitative scores and the scores are related to disease in the target sample. Linear mixed model is an alternative approach to fit all the GWAS SNPs where the effects of the SNPs are treated statistically as random, and the variance explained by all the SNPs together is estimated. In order to determine the sensitivity of both methods to distant relatedness we simulated phenotypes and genotypes that have low pairwise correlation, to see whether small correlations between individuals have a major effect on the estimate. In the score analysis phenotypic variance explained increased as a greater number of SNPs were considered; on a liability scale 29.5% of variance of Prostate Cancer in African Americans can be explained by considering 873,644 SNPs in linear mixed model. However, we concluded that the estimates may be biased upward due to unmeasured causal alleles and confounding effects of distant relatedness.

59

Heritability of John Henryism, and Correlation Between John Henryism and Hypertension in the Jackson Heart Study

Sarah G Buxbaum (1) Pooja Goel (1) Wendy White (2) Mathew Gregoski (3) Sandra H Dunn (4)
(1) Jackson State University (2) Tougaloo College (3) Medical University of South Carolina (4) University of Pittsburgh

Purpose: The John Henryism Scale of Active Coping is measured in the Jackson Heart Study, a cohort of African Americans in the Jackson, MS area. The purpose of this study is to assess the heritability of John Henryism (JH) and to determine whether or not it is correlated with hypertension. The null hypothesis for this study was twofold: 1) John Henryism is not associated with hypertension; 2) John Henryism is not genetic. The initial data set comprised 4,567 participants who answered questions in the third JHS annual follow-up questionnaire and 64.5% are women. Method: After recoding the JH survey responses, a score was created by summing the values. Spearman correlations between JH and both hypertension and age were determined. Heritability of the JH score within the family subset, excluding singletons (N=1040), was determined using S.A.G.E. and SOLAR, with adjustment for hypertension and age. Results: 3,987 participants answered all 12 questions in the John Henryism questionnaire. Among these, the correlation between JH and hypertension was not significant (p-value=0.39). However, JH is inversely correlated with age ($r=-0.12$, $p<0.0001$). The heritability of John Henryism was 10% with adjustment for hypertension and 9% with further adjustment for age, and not quite statistically significant (p=0.07). Conclusion: This study failed to reject the null hypotheses. Unlike a previous study, we did not find strong evidence of heritability of John Henryism.

60

Estimation of Heritability of Survival Outcomes Using Residuals from Proportional Hazards Models

Joel A Mefford (1) John S Witte (1)
(1) Department of Epidemiology and Biostatistics, University of California San Francisco

We are interested in estimating the heritability of survival outcomes that is explained by the SNPs genotyped for genome-wide association studies.
The linear mixed model (LLM) approach to heritability estimation has been used for analyses of continuous traits, and for binary traits after translation of the outcomes from the binary scale to a continuous liability scale. The data from survival analyses cannot be used directly in the LMM framework because of the prevalence of censored observations.
Here we examine the use of residuals from Cox proportional hazards models as quantitative traits for a subsequent estimation of the heritability of a survival outcome using the LMM approach. Simulated datasets are produced using genotypes from GWAS and event times generated by specified disease models. Event times are coarsened at random to yield right-censored observations. Cox proportional hazards models are fit to the datasets. Then, the martingale and deviance residuals are extracted from the fit survival models.
The residuals are used in estimation of the heritability of the survival phenotype. The performance and interpretability of the two types of residuals are compared.
The residual analyses are applied to data from the paclitaxel arm of the CALGB 40101 breast cancer clinical trial. This is a genome-wide association study where the outcome examined is cumulative dose of paclitaxel until occurrence of peripheral neuropathy.

61

Refined IBD: A New Method for Detecting Identity by Descent in Population Samples

Brian L Browning (1) Sharon R Browning (1)
(1) University of Washington

Identity-by-descent (IBD) from a recent common ancestor is ubiquitous in large population samples. IBD manifests as a long haplotype shared by two or more individuals, and IBD can be detected with high power when the recent common ancestor is within the past 25 generations. The power to detect IBD can be increased by modeling haplotype phase uncertainty, but this modeling has been computationally expensive.
We have developed a new method for IBD detection called Refined IBD that solves the computational problem of incorporating haplotype phase uncertainty. First, candidate IBD tracts are identified by searching for shared haplotypes that exceed a length threshold. Candidate tracts are efficiently identified using IBD detection methods such as GERMLINE or fastIBD that do not model haplotype phase uncertainty. We then compare the likelihood of the genotypes in each candidate tract under IBD and non-IBD models using a likelihood ratio test. This test incorporates haplotype phase uncertainty and identifies the subset of candidate tracts with the greatest evidence for IBD.
We have applied Refined IBD to the WTCCC2 control data (n=5200). After tuning methods to achieve equal false positive rates, Refined IBD detected fourfold more IBD than the GERMLINE and fastIBD algorithms.
Refined IBD is implemented in BEAGLE version 4 (http://faculty.washington.edu/browning/beagle/beagle.html).

62

Detection of Excess Homozygosity in Association with Rheumatoid Arthritis Using SNPs

Chih-Chieh Wu (1) Sanjay Shete (1) Eun-Ji Jo (2) Yue E Lu (1) Yaji Xu (3) Wei V Chen (1) Christopher I Amos (1)
(1) MD Anderson Cancer Center (2) Baylor College of Medicine (3) Yale University

Because deletions are abundant in humans and because known rheumatoid arthritis (RA) susceptibility loci explain only a small portion of familial clustering, we performed genome-wide study of association of deletion or excess homozygosity with RA using high-density 550K SNPs. We applied genome-wide statistical method that we recently developed and PennCNV to test each contiguous SNP locus between 868 cases and 1197 controls for detecting deletions or homozygosity that influence susceptibility. Our method is designed to detect statistically significant evidence of deletions at individual SNPs for SNP-by-SNP analyses and combine information among neighboring significant SNPs for cluster analyses. In addition to successfully

detecting known deletion variants on HLA-DRB1 and C4 in HMC, we identified respective 4.3-kb and 28-kb clusters on chromosomes 10p and 13q, significant at a corrected 0.05 nominal significance level. Independently, we ran PennCNV to identify cases and controls having segments with copy number=0 or 1. Using Fisher's exact test for comparing numbers of cases and controls per SNP, we identified 26 significant SNPs (protective; more controls than cases) aggregating on chromosome 14 with p-values $<10^{-8}$ and 49 SNPs on chromosomes 2, 14, and 20 with p-values between 10^{-5} and 10^{-8}. Twelve cases and 1 control commonly shared a 6.6-kb segment with copy number=1, which lay between 2 adjacent significant SNPs on chromosome 19p detected by our cluster method.

63

Homozygosity by Descent Detection and Mapping

Elisa Sheng (1) Brian L Browning (1) Sharon R Browning (1)
(1) University of Washington

An individual is homozygous by descent (HBD) for a segment on a chromosome if the individual inherited both chromosome copies from a single ancestral haplotype. Detected HBD is useful for HBD mapping, which involves searching for regions harboring recessive causal variants by looking for positions at which rates of HBD are higher in cases than in controls.
HBD may be identified from genotype data by looking for long segments over which the genotypes are homozygous. GERMLINE uses a length-based threshold based on genetic (centiMorgan) distance, while PLINK uses a length-based threshold based on physical (basepair) distance. In contrast, BEAGLE models linkage disequilibrium (LD) and uses a probabilistic approach. Each of these approaches has parameters that can be varied. We applied these methods with varied parameter settings to real data. We compared false-positive rates by masking a proportion of the genotypes and measuring the proportion of masked genotypes that are heterozygous in inferred HBD tracts. When holding this discordance rate fixed, BEAGLE has more power to detect HBD than the other two methods. Thus use of BEAGLE should make HBD mapping more powerful. We conduct HBD mapping using BEAGLE in data from the Autism Genome Project, and we compare power of HBD mapping using the three methods on simulated data.

64

Identity-by-Descent Analysis of Sequence Data

Steven M Smith (1) Sharon R Browning (1) Brian L Browning (1)
(1) University of Washington

Individuals are identical by descent (IBD) if they share a haplotype by inheritance from a recent common ancestor. A shared haplotype with length greater than 2 centimorgans corresponds to a common ancestor within approximately the past 25 generations, and these long shared haplotypes can be detected with high power using existing methods for SNP array data.
IBD detection methods can also be applied to sequence data, and the detected IBD can be used to phase rare variants and to detect genotype errors. In theory, the complete marker coverage in sequence data will maximize power to detect IBD. However in practice, alignment and sequencing artifacts, reduced phase accuracy at low frequency markers, and recent mutations pose challenges to IBD detection in sequence data. We will present and evaluate several methods for improving IBD detection in sequence data using 1000 Genomes Project and simulated data.

65

Identity-by-Descent-Based Heritability Analysis in the Northern Finland Birth Cohort

Sharon R Browning (1) Brian L Browning (1)
(1) University of Washington

For most complex traits only a small proportion of heritability is explained by statistically significant associations from genome-wide association studies (GWAS). In order to determine how much heritability can potentially be explained through larger GWAS, several different approaches for estimating total narrow-sense heritability from GWAS data have recently been proposed. These methods include variance components with relatedness estimates from allele sharing, variance components with relatedness estimates from identity by descent (IBD) segments, and regression of phenotypic correlation on relatedness estimates from IBD segments. These methods have not previously been compared on real or simulated data. We analyzed the narrow-sense heritability of nine metabolic traits in the Northern Finland Birth Cohort using these methods. We found substantial estimated heritability for several traits. When we used a variance components approach, IBD-based estimates of heritability were higher than allele-sharing based estimates for almost all traits examined, which is consistent with a significant role of rare variants in these traits because rare variants are generally not well tagged by SNPs but are captured by long IBD segments. Estimates of heritability from the regression-based approach are much lower than variance components estimates in these data, which may be explained by the presence of strong population structure in these data.

66

Using Whole Exome Sequencing to Identify Rare Causal Variants for Oral Clefts in Multiplex Families with a Focus on Syrian Families

Joan E Bailey-Wilson (1) Margaret M Parker (2) Silke Szymczak (1) Qing Li (1) Cheryl D Cropp (1) Markus M Nothen (3) Jacqueline B Hetmanski (2) Hua Ling (4) Elizabeth W Pugh (4) Priya Duggal (2) Margaret A Taub (2) Ingo Ruczinski (2) Alan F Scott (4) Mary L Marazita (5) Jeffrey C Murray (6) Elisabeth Mangold (3) Terri H Beaty (2)
(1) National Human Genome Research Institute, National Institutes of Health (2) Johns Hopkins Bloomberg School of Public Health (3) Institute of Human Genetics, University of Bonn (4) Johns Hopkins School of Medicine (5) School of Dental Medicine, University of Pittsburgh (6) University of Iowa Children's Hospital

Oral clefts (cleft lip, cleft palate and cleft lip & palate) are common birth defects with a complex and heterogeneous etiology. Some genes and chromosomal regions have been associated with risk in GWAS and linkage studies. This whole exome sequencing (WES) study used 108 affected 2^o and 3^o relatives drawn from 52 multiplex families (4

families with 3 relatives and 48 families with 2 relatives) ascertained for linkage. WES was done by CIDR using the Agilent SureSelect v.4 capture reagents & Illumina HiSeq 2000 sequencers. Initially, we focused on truly novel single nucleotide variants (SNVs), i.e. not previously reported, shared by affected relatives within a family (exact genotype matches only) and predicted damaging by SIFT score ($\leq$0.05). A total of 516 novel SNVs were shared between affected relatives in these 52 families. Only one novel SNV (A to G at position 3056632 in hg19) in *ZNF764* was shared by affected relatives across 2 families, and these may be distantly related. When analyzing 10 inbred families from the Syrian Arab Republic, after QC, we found 10 rare variants (MAF<1% in 1000 Genomes) that were homozygous in all 21 affected, distantly related individuals and were predicted to be damaging, with 9 predicted to result in loss of gene function. Additional sequencing studies of more families and more affected individuals in these families are ongoing to determine which genes segregate with oral clefts in these Syrian families.

67

Characterization of Rare Variants in Melanoma-associated Genes in Melanoma-prone Families without CDKN2A/CDK4 Mutations using Exome Sequencing Data

Xiaohong Yang (1) Kevin Jacobs (1) Michael Cullen (1) Joseph Boland (1) Laurie Burdett (1) Michael Malasky (1) Melissa Rotunno (1) Meredith Yeager (1) Stephen Chanock (1) Margaret Tucker (1) Alisa Goldstein (1)
(1) National Institutes of Health

Recent genome-wide association studies (GWAS) have identified common genetic variants in a number of genes that are associated with melanoma risk. The goal of this study is to examine whether rare variants in these genes may influence melanoma risk in melanoma-prone families without known mutations. Methods: We conducted exome sequencing in blood-derived DNA in 50 melanoma cases from 17 families (number of cases 1–5 for each family). We evaluated rare non-synonymous variants, defined as allele frequency less than 5% among our internal controls (N$\sim$200) or reported in public databases (dbSNP, 1000 Genomes, or NHLBI's Exome Variant Server), in *CDKN2A, CDK4, BAP1, MITF, MC1R, ATM, TERT, MTAP, VDR, ASIP, TYR, TYTP1, OCA2, SLC45A2, PLA2G6, MX2, PARP1, and CASP8*. Results: We identified 27 rare variants in 10 genes. Among them, 9 variants occur in multiple families and 17 variants occur in multiple cases within families (4 in all three cases within respective families). Seven variants were novel (not seen in controls or reported before), but only 2 of them occur in multile cases within a family. The majority (16 out of 27) of all identified rare variants were predicted as damaging by either PolyPhen or SIFT or both. Our findings suggest that rare variants in known melanoma-associated genes may influence familial melanoma risk, however, these varaints need to be validated and evaluated in a large number of cases and controls to determine their role in disease associations.

68

Association of Variants near-NLRP1 with High Density-lipoprotein Cholesterol in the Long Life Study Family

Mary F Feitosa (1) Aldi T Kraja (1) Joseph Lee (2) Kaare Christensen (3) Judy Wang (1) Candace Kammerer (4) Michael A Province (1) Ingrid B Borecki (1)
(1) Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine (2) Taub Institute, Columbia University (3) Institute of Public Health at the University of Southern Denmark (4) Department of Human Genetics, University of Pittsburgh Graduate School of Public Health

We sought to identify novel loci that influence levels of high-density lipoprotein (HDL) cholesterol in the Long Life Family Study (LLFS), which was designed to characterize families of exceptionally healthy, elderly people. For subjects taking lowering-lipid medications, the HDL levels were corrected for the effect of these medications. HDL was adjusted by field centers, age, age^2, sex, and twenty principal components using a stepwise regression analysis. We performed genome-wide association scans on HDL using a mixed model approach and accounting for family structure with the kinship correlations. A total of 4,114 European-American subjects (480 families) was genotyped at $\sim$2.3 million SNPs (Illumina Omni chip). We identified novel variants near-*NLRP1* (17p13, $p<3.2\times10^{-8}$) associated with an increase of HDL levels. Additionally, several *CETP* (16q21) variants associated with HDL at genome-wide significant levels ($p<5.0\times10^{-8}$) were found, replicating those previously reported in the literature. A possible regulatory variant upstream of *NLRP1* is associated with higher levels of HDL in LLFS data, which may contribute to the longevity and health of these subjects. *NLRP1* plays an important role in the induction of apoptosis, and its inflammasome is critical for mediating innate immune responses. An *NLRP1*-variant has been reported to be more frequent in the long-lived than control samples; however, the connection with HDL levels has not been previously reported.

69

Natural and Orthogonal Association Framework to Detect Parent-of-Origin Effects

Feifei Xiao (1) Jianzhong Ma (1) Christopher I. Amos (1)
(1) UT M.D Anderson Cancer Center

Many human diseases, such as cancer, diabetes and obesity may be related to imprinted genes. Traditional association studies always assume equal contribution of the paternal and maternal allele to the trait. Recently many studies have reveals that the power of GWAS could be improved when the parent-of-origin effect (POE) is considered in the analysis model, while searching for complex disorders related genes. In the natural and orthogonal interaction (NOIA) framework which was developed for the quantitative trait analysis by Alvarez-Castro and Carlborg [Genetics, 176 1151–1167 (2007)], the estimates of genetic effects are unrelated, also called orthogonal. Therefore I proposed that the power of NOIA could be improved when POE is combined to the analysis model while keeping its orthogonality. We conducted simulations to evaluate the performance of the statistical models in comparison with the usual functional models. We also compared the usual NOIA model with and without POE considered in the model. The results showed that the power for testing associations while allowing for POE considered using the statistical model is much higher than using usual functional models. And as expected, even

with the POE considered, the statistical model is still orthogonal for detecting the overall genetic effect and POE effect. We also applied our approach to family data of oral cleft disorder.

70

Sex-specific and X-chromosome Association Studies of Venous Thromboembolism (VTE)

Mariza de Andrade (1) Sebastian M Armasu (1) Landon L Chan (1) John A Heit (1)
(1) Mayo Clinic

Background: VTE is a complex disease resulting from multigenic action and environmental exposures.
Objective: To identify autosomal and X-chromosome VTE-susceptibility genes, both overall and by sex. Methods: Genome-wide scan genotypes from 1270 non-Hispanic adults of European ancestry with objectively-diagnosed VTE and 1302 controls (frequency-matched on case age, sex, race, MI/stroke status) were imputed to 9.2 million SNPs using BEAGLE and EUR 1000G. Genome-wide association analyses were performed on autosomal and X-chromosome SNPs, both overall and by sex, using PLINK. Results: For the autosomal females analysis, after adjusting for covariates, 3 SNPs on chromosome 1q24.2 (including *F5* [Factor V Leiden] rs6025, odds ratio [OR]=3.2, p=2.5E-9) and 38 on chromosome 9q34.2 (including *ABO* rs2519093, OR=1.7, p=3.7E-9; and rs495828, OR=1.7, p=3.5E-9) exceeded genome-wide significance. Similar results were observed for males (*F5* rs6025, OR=4.1, p=3.0E-12;*ABO* rs2519093, OR=1.9, p=1.4E-10;rs495828, OR=1.7, p=2.7E-8). For the X-chromosome overall analyses, the most significant SNPs were near 4 pseudo-genes and *FAM46D* (rs12688059, OR=1.48, p=7.07E-5); for females only, *FAM47A* (chrX:34138682, OR=0.048, p=2.54E-5) and *TEX11* (chrX:69992070, OR=2.81, p=6.20E-5); and for males only, *SCML2* (rs58926087, OR=0.74, p=8.14E-5). Conclusion: Autosomal and X-chromosome VTE-susceptibility genes vary, overall and by sex, consistent with the hypothesized multigenic action.

71

Merging Genomic Data for Research in the Electronic MEdical Records and GEnomics Network: Lessons Learned in eMERGE

Marylyn D Ritchie (1) Shefali Z Setia (1) Gretta D Armstrong (1) Loren Armstrong (2) Yuki Bradford (3) Dana C Crawford (3) David R Crosslin (4) Mariza de Andrade (5) Kimberly F. Doheny (6) M. Geoffrey Hayes (2) Gail P. Jarvik (7) Iftiklhar Kullo (5) Rongling Li (8) Cathy A. McCarty (9) Daniel Mirel (10) Lana Olson (3) Shaun Purcell (11) Elizabeth W Pugh (6) Gerard Tromp (12) Helena Kuivaniemi (12) Vaneet Lotay (11) Omri Gottesman (11) Jonathan L Haines (3)
(1) The Pennsylvania State University, University Park, PA (2) Feinberg School of Medicine, Northwestern University, Chicago, IL (3) Vanderbilt University, Nashville, TN (4) University of Washington, Seattle, WA (5) Mayo Clinic, Rochester, MN (6) Center for Inherited Disease Research, Johns Hopkins University (7) University of Washington Seattle, WA (8) NHGRI, NIH (9) Essentia Rural Health, Duluth, MN (10) The Broad Institute (11) Mt Sinai School of Medicine, New York, NY (12) Geisinger Health System

Biobanks linked to electronic health records (EHR) is an emerging area of research for dissecting the architecture of complex traits.Electronic phenotyping algorithms are deployed in large EHR systems to "ascertain" samples for analysis. The eMERGE network, an NHGRI funded initiative, has developed a pipeline for merging genomic data generated on a single platform as well as a new pipeline for merging data from different genotyping arrays based on imputation to maximize sample size and power. eMERGE consists of seven sites each with DNA databanks linked to EHRs.Over 42,000 samples have been genotyped using one of the available Affymetrix or Illumina genome-wide genotyping arrays. These data have been imputed using BEAGLE and October, 2011 release of 1000 Genomes cosmopolitan reference panel. Because of the computational complexity, a distributed imputation pipeline was implemented. In this scheme, the genome was divided by SNPs into 30,000 marker "SNPlets" with 700 markers of overlap on each side, resulting in 510 SNPlets. This parallelized pipeline resulted in over 556 billion SNPs (more than 13 million per individual) based on hundreds of thousands of CPU hours. The dataset generated consists of genome-wide SNPs on thousands of individuals all linked to EHR systems where numerous phenotypes can be explored.The lessons learned by this group of investigators will be valuable for the genomics community also dealing with the combining of large-scale genomic datasets.

72

Genome-wide Association Analysis of Rare Variants Identifies Potential Novel Susceptibility Genes for Type 1 Diabetes and Coronary Artery Disease

Andrew P Morris (1) Reedik Magi (2)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford (2) Estonian Genome Center, University of Tartu

We investigated association of seven diseases with rare genetic variation (MAF<1%) within genes via imputation into genome-wide genotype data (Affymetrix GeneChip 500K) in 13,241 cases and 2,938 controls of European descent from the Wellcome Trust Case Control Consortium. We performed imputation using IMPUTE2 up to the 1000 Genomes Phase I (interim) reference panel, and tested for association of each disease with accumulations of minor alleles at high-quality rare variants (info>0.4) within genes (boundaries defined from the UCSC Human Genome database) using GRANVIL. We observed genome-wide significant evidence of rare variant association (Bonferroni correction for 30,000 genes, $p<1.7\times10^{-6}$) with two diseases. For coronary artery disease, we observed association with rare variants in *PRDM10* ($p=4.9\times10^{-8}$). We also observed evidence of rare variant association with type 1 diabetes in 10 MHC genes. The strongest signals were observed for *HLA-DRA* ($p=2.0\times10^{-13}$), *HLA-DRB5* ($p=1.6\times10^{-10}$) and *SLC44A4* ($p=1.7\times10^{-10}$), all of which remained significant after adjustment for the previously described common variant association (rs9268645) in this region. Our results highlight the potential for the identification of rare variant associations using existing GWAS genotyping data, supplemented with imputation from publicly available high-density reference panels, without the need for costly re-sequencing experiments.

73

Re-sequencing of the 5p15.3 Region Identifies Novel Rare Variants from Lung Cancer Cases and Controls

Hagit Katzov-Eckert (1) Gord Fehringer (1) Robert E. Denroche (2) Philip Zuzarte (2) John McLaughlin (1) John D. McPherson (2) Rayjean J. Hung (1)
(1) Samuel Lunenfeld Research Institute of Mount Sinai Hospital (2) Ontario Institute for Cancer Research

Genome-wide associations have identified the 5p15.3 locus to be associated with lung cancer susceptibility. However, the causal variants remain to be elucidated. To identify novel variants, we re-sequenced a 250kb region on 5p15.3 spanning *hTERT*, *CPTM1L*, *SLC6A3* and *LPCAT1* in 576 cases and controls from European and Asian descent. On average, in depth coverage was obtained (>6,000x) using Agilent SureSelect capture array and Illumina HiSeq 2000. Validation on Ion PGM sequencing confirmed the presence of 96.2% of the variants. A total of 2,019 variants (67 coding and 1,952 noncoding) were detected. Of these variants 1,200 (59%) were novel based on a cut-off of MAF>1%. In addition, 1,030 variants were unique to the European population and 86 variants were unique to the Asian population. Computational tools were used to evaluate the functional effects of the variants and to prioritize them for further fine-mapping studies. In the coding region, 8 non-synonymous variants were predicted to be damaging. The noncoding region included 105 highly conserved variants predicted to affect transcription and splics sites and 40 variants predicted to affect miRNA binding sites. Selected variants will be genotyped in a large cohort of lung cancer cases and controls to narrow down the susceptibility region and identify variants that influence the risk of developing lung cancer.

74

Identification of Risk Genes through Whole Exome Sequencing in the Colon Cancer Family Registry

Melissa S DeRycke (1) Shanaka R Gunawardena (2) Shannon K McDonnell (3) Shaun M Riska (3) Yan W Asmann (3) Sumit Middha (3) Daniel J Schaid (3) Bruce W Eckloff (4) Stephen N Thibodeau (2) Noralane M Lindor (5) John L Hopper (6) Mark A Jenkins (6) Daniel D Buchanen (7) Michael Woods (8) Robert W Haile (9) Graham Casey (9) John A Baron (10) Steve Gallinger (11) Ellen L Goode (1)
(1) Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA (2) Department of Laboratory Medicine & Pathology, Mayo Clinic, Rochester, MN, USA (3) Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA (4) Advanced Genomics Technology Center, Mayo Clinic, Rochester, MN, USA (5) Department of Health Sciences Research, Mayo Clinic, Scottsdale, Arizona, USA (6) Center for Molecular, Environmental, Genetic, and Analytic Epidemiology, The University of Melbourne, Melbourne, Australia (7) Cancer and Population Studies Group, Queensland Institute of Medical Research, Herston, Australia (8) Discipline of Genetics, Memorial University, St. John's, Newfoundland, Canada (9) Department of Preventive Medicine, University of Southern California, Los Angeles, California, USA (10) University of North Carolina, Chapel Hill, NC (11) Cancer Care Ontario, Mount Sinai Hospital, Samuel Lunenfeld Research Institute, University of Toronto, Ontario, Canada

Several genes have been implicated in familial forms of colorectal cancer (CRC); however, they only explain a small fraction of the familial risk. Identifying additional CRC risk genes may explain families without inherited mutations in known genes and illuminate new therapeutic targets.
Whole exome sequencing (WES) of 40 individuals from 16 high-risk families was conducted using Agilent's SureSelect Exon kit and either an Illumina GAIIx or HiSeq2000. Called variants were grouped based on their predicted severity. Filtering excluded common, tolerated, or benign variants and variants not shared among CRC cases within a family.
Thirty-three nonsense, intronic splice-site, or frameshift SNPs remained (Tier I); only four occurred in multiple families (in CDC27, KIR3DL1, SHROOM3, and TMC2), and all were in unique genes.
There were 344 missense, exonic or splice-site SNPs (Tier II); nine were in multiple families and all but 16 were in unique genes.
Synonymous, intronic, or intergenic SNPs were the most abundant (Tier III). We observed 1,394 of these with 1,203 in only a single family. Most Tier III SNPs were in unique genes (n=837); the remaining variants were in 152 genes (2–30 variants/gene).
Fifty indels remained after filtering; 45 were unique to one family and five were in >1 family.
WES identified many genes with rare variants in CRC families that may influence risk. Validation by re-sequencing of selected genes is underway in an additional 1,500 familial CRC cases.

75

NF-kB Polymorphisms and Ovarian Cancer Risk

Bridget Charbonneau (1) William R Bamlet (1) Robert A Vierkant (1) Julie M Cunningham (1) Kimberly R Kalli (1) David N Rider (1) Brooke L Fridley (1) Jonathan Tyrer (2) Joe Dennis (2) Susan Ramus (3) Paul D Pharoah (2) Catherine M Phelan (4) Ellen L Goode (1)
(1) Mayo Clinic (2) University of Cambridge (3) University of Southern California (4) Moffitt Cancer Center

NF-kB, a transcription factor family that activates several pro-inflammatory genes, is thought to be an important mediator in carcinogenesis related to chronic inflammation. We tagged single nucleotide polymorphisms (SNPs) in over 200 genes in the NF-kB pathway, resulting in 2,282 SNPs selected based on $r^2 \geq 0.8$, MAF ≥ 0.05, 5 kb upstream and downstream from the gene, and passing genotype quality control. Data was available for ovarian cancer risk assessment with these SNPs from 15,604 combined invasive cases, including 9,888 serous, 2,100 endometrioid, 1,591 mucinous, and 1,034 clear cell cases and 23,235 controls of European descent across multiple studies from the Ovarian Cancer Association Consortium (OCAC). After adjustment for age, study site and population stratification using the first five principal components, the top five SNPs associated with invasive ovarian cancer risk ($p<0.001$) included two SNPs from *CARD11* and individual SNPs from *PRKCA*, *TAF3*, and *NFKB1*. The top SNPs associated with each subtype were found in *TNFSF10*, *CD3E*, *F2R*, and *PRKCA* for serous, clear cell, endometrioid, and mucinous ovarian cancer, respectively. Our results suggest polymorphisms in several NF-kB pathway genes may be related to ovarian cancer risk, providing further support for the role of this pathway in ovarian cancer carcinogenesis. Additionally, we found that

within this pathway, individual genes may have differing roles in risk for each subtype.

76

Field Synopses of Genetic Variation in Colorectal Neoplasia

Julian Little (1) Harry Campbell (2) Gillian Gresham (1) Zahra Montazeri (1) Shanya Sivakumaran (2) Evropi Theodoratou (2)
(1) University of Ottawa (2) The University of Edinburgh

The scientific community and policy makers appear to be developing an increasing appetite for comprehensive overviews of broad areas within specific fields.
We have recently completed a field synopsis for colorectal cancer (CRC) and are in progress of developing one on colorectal polyps. For the CRC synopsis, we reviewed over 10,000 titles, then collated and extracted data from >600 publications reporting on >400 polymorphisms in >100 different genes. We carried out meta-analyses to derive summary effect estimates for >90 polymorphisms in >60 genes. To assess the credibility of associations, we applied the Venice criteria and added consideration of Bayesian False Discovery Probability (BFDP). We considered four genetic models (two additive, one dominant, and one recessive).
We will present illustrative results from the CRC field synopsis. We will report on work in progress on colorectal polyps, and comment on differences in the nature of the evidence on these precursor lesions for CRC, such as volume and quality of evidence, and issues including manner of detection of polyps (symptomatic vs. asymptomatic), investigation of initially detected vs. recurrent polyps, and subtype (defined by histology, size and multiplicity). In addition, we wish to stimulate discussion about (a) updating of field synopses and (b) operationalization of the Venice criteria, for both of which there appear to be differences across field synopses.

77

Gastrointestinal Stromal Tumors, Somatic Mutations, and Candidate Genetic Risk Factors

Katie M O'Brien (1) Irene Orlow (2) Cristina R Antonescu (3) Karla Ballman (4) Linda McCall (5) Ronald DeMatteo (6) Lawrence S Engel (1)
(1) Department of Epidemiology, University of North Carolina at Chapel Hill (2) Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center (3) Department of Pathology, Memorial Sloan-Kettering Cancer Center (4) Department of Health Sciences Research, Mayo Clinic (5) American College of Surgeons Oncology Group (6) Department of Surgery, Memorial Sloan-Kettering Cancer Center

Gastrointestinal stromal tumors (GISTs) are rare but treatable soft tissue sarcomas. Nearly all GISTs have somatic mutations in either the *KIT* or *PDGFRA* gene, but there are no known inherited genetic risk factors. We assessed the relationship between *KIT*/*PDGFRA* mutations and select single nucleotide polymorphisms (SNPs) in 331 participants from a clinical trial of adjuvant imatinib mesylate. As previous epidemiologic studies suggest dioxin and radiation exposure may be linked to the disease, we chose 216 SNPs in 39 candidate genes related to DNA repair and dioxin metabolism or response. We calculated adjusted odds ratios (ORs) and 95% confidence intervals (CIs) for the association between SNPs and 7 categories of tumor mutation using logistic regression. We also evaluated gene-level effects using the sequence kernel association test (SKAT). SNPs in *CYP1B1* were strongly associated with *KIT* exon 11 codon 557 deletions (OR=1.9, 95% CI: 1.3–2.8 for rs2855658 and OR=1.8, 95% CI: 1.2–2.6 for rs1056836) and wildtype GISTs (OR=2.7, 95% CI: 1.5–4.9 for rs1800440 and OR=0.5, 95% CI: 0.3–0.9 for rs1056836). *CYP1B1* was associated with these mutations types in the SKAT analysis (p=0.002 and p=0.003, respectively). Other potential risk variants include *GSTM1, RAD23B* and *ERCC2.* This preliminary analysis of inherited genetic risk factors for GIST offers some clues about the disease's genetic origins and provides a starting point for future candidate gene or gene-environment research.

78

Breast Cancer in Women with Family History of Breast Cancer who have Tested Negative for BRCA1 or BRCA2 Mutation

Nataliya Kitsera (1) Yaroslav Shparyk (2)
(1) Institute of Hereditary Pathology (2) Lviv regional diagnostic cancer center, Lviv, Ukraine

Our work has **aimed** to estimate frequency of breast cancer (BC) in women with a family history and who tested negative for a mutation in *BRCA1/ BRCA2*. **Methods of study -** genealogical, molecular genetics.
Results 110 women with BC were examined, whose relatives had undergone treatment of breast cancer. The women's age was between 25 and 75 years (median is 52.2 years), which were treated in the Lviv regional diagnostic cancer center from 2008 till 2012.
For the first time in Ukraine, studied a wide range of mutations BRCA1/2 mutations in seven genes *BRCA1* (185delAG,4153delA,5382InsC,188del11, 5396+ 1G>A,185InsA,5331 G>A) and three gene mutations in *BRCA2* (6174delT, 6293S> G, 6024delTA) in patients with BC. In 2 cases diagnosed mutation of the gene BRCA1 (5382 Ins C), and one – *BRCA* (185delAG).
A total of 107 women with BC who were tested negative for a mutation in *BRCA1/BRCA2* included in the analyses. We study 82 (74.5%) probands where first-degree relatives had the same diagnosis: mother-58, sisters-18,daughter-6. The same diagnosis as 12 (10.9%) probands had their second-degree relatives (grandmother, aunt), 16 (14.5%) women and their three-degree relatives (great-grandmother, cousin). There were 92 (83.6%) probands with BC of families on mother's side, 10 (9.1%) – of families on father's side and 8 (7.3%) families of both side.
Conclusion Our data suggest that such women in the West Ukraine should adhere to population based guidelines for breast cancer screening.

79

Inference Based on Distribution-Comparisons of BP Candidate Gene-Association Results among LLFS versus FHS (CVD High Risk and Randoms)

Aldi T. Kraja (1) Candace Kammerer (2) Haley J. Abel (1) Robert Straka (3) Joseph H. Lee (4) Michael A. Province (1)
(1) Div. of Statistical Genomics, Dep. of Gen., Washington U. School of Med., St. Louis, MO, USA (2) Dep. of Human

Genetics, U. Of Pittsburgh Graduate School of Public Health, PA, USA (3) Dep. of Experimental and Clinical Pharmacology, U. of Minnesota, MN, USA (4) Taub Institute, Columbia U., NY, USA

A large blood pressure, hypertension (BP/HTN) gene-network was built from human, mouse and/or rat literature. We used this gene-set in the Long Life Family Study (LLFS), to evaluate subsets of genes that contribute / are protective to BP/HTN. The LLFS is a family-based cohort study designed to characterize exceptional aging in the elderly (N>4,600, genotyped with ∼2.5M SNPs; 52% after medication corrections classified with HTN). Because of unknown exact validity of BP genes, we tested the same gene-set in the Family Heart Study (FHS) split in a high risk sample for cardiovascular disease (FHS-N) and a random one (FHS-RQ) representing general population (N>5,000, with ∼2.5M "hybrid" SNPs; 66% after medication correction classified with HTN). An additive genetic model with Linear Mixed Effects with Kinship was implemented on the gene-set SNPs for systolic, diastolic BP, pulse pressure and hypertension. Individuals using antihypertensives were analyzed also with phenotypes corrected for classes of medication effects. Distributions of each of the >800 candidate gene results were compared by using a Kolmogorov-Smirnov bootstrapped test ignoring LD, a QQ-AUC assessing LD and significance via simulation, and graphically between LLFS vs FHS-N / FHS-RQ on all traits. As a result, two pools of BP candidate genes that relate or not as a proxy with CVD were inferred. BP candidates are further investigated bioinformatically for their importance in the BP/HTN genetic architecture.

80

Whole Exome Sequencing to Identify Genes Associated with Hypertension

Bamidele O. Tayo (1) Babatunde Salako (2) Amy Luke (1) Xiaofeng Zhu (3) Adesola Ogunniyi (2) Richard S. Cooper (1)
(1) Loyola University Chicago Stritch School of Medicine, Maywood, IL (2) University of Ibadan, Ibadan, Nigeria (3) Case Western Reserve University, Cleveland, OH

BACKGROUND: Hypertension is the most common cardiovascular condition in the world and accounts for a substantial proportion of adult mortality. Although elevated blood pressure has similar heritability to many other traits related to cardiovascular risk, genetic susceptibility loci have been difficult to localize. **OBJECTIVE:** As part of a long-term study on genetics of hypertension in Blacks, we conducted whole exome sequencing in a sample of adult Nigerians to identify genes associated with hypertension in this population. **METHOD:** We sequenced DNA from 50 unrelated adults selected from both extremes of the blood pressure distribution using the Illumina Hi-Seq 2000 instrumentation using the 2 × 100 base pair paired-end sequencing protocol. Sequencing reads were aligned to the human genome build Hg19. Single Nucleotide Polymorphism analysis of the captured regions provided data within the targeted exon regions and these were used in association analysis of the 40 extreme cases and 10 extreme controls. **RESULTS:** Our findings indicate that mutations in B-cell translocation gene 4 (*BTG4*) and keratin 77 (*KRT77*) genes are associated with hypertension among Nigerians. **CONCLUSION:** This study provides preliminary data that these loci influence susceptibility to hypertension.

81

Histologic Types and Risk Factors in Familial Lung Cancer Cases from Southern Louisiana

Diptasri M. Mandal (1) Matthew Haskins (1) Angelle Bencaz (1) Jill Hutchinson (1) Jessica Chambliss (1) Henry Rothschild (1) Joan E. Bailey-Wilson (2)
(1) Department of Genetics, Louisiana State University Health Sciences Center, New Orleans, Louisiana (2) National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland

In the 1980's in the population in Louisiana (LA), squamous cell carcinoma was observed to be the most frequent type of lung cancer (LC)(39.3%), with nearly equal numbers of adenocarcinoma (25.0%) and small cell subtypes (25.5%). While some studies have shown similar proportions of familial lung cancer (FLC) cases in all histologic subtypes of LC, others have suggested that a higher proportion of patients with squamous cell carcinoma have a family history of LC.The objective of the present study was to analyze histologic subtypes and their association with smoking behaviors and other risk factorsamong FLC cases from southern LA. Histologic subtype (N=114) was abstracted from pathology reports of eligible subjects (N=148) with ≥2 relatives affected with primary LC. About 81% of cases had non-small cell LC, with adenocarcinoma (40%) being the most common histologic subtype; squamous cell was associated with higher mean pack-years of smoking and older age. A significant difference in the age of diagnosis was observed between cases with non-small cell and small cell LC; the mean number of pack years was about twice as high in cases with non-small cell LC. The frequency of adenocarcinoma in these FLC cases was higher than previously reported for the population in LA. This is consistent with a higher risk of adenocarcinoma in FLC cases, which may be related to our previous observations that less smoking appears to be necessary for LC in persons with a familial risk.

82

The Role of Genes and Life Course in Late Life Diseases

Alexander M Kulminski (1) Irina Culminskaya (1) Konstantin G Arbeev (1) Svetlana V Ukraintseva (1) Liubov Arbeeva (1) Anatoli I Yashin (1)
(1) Duke University

Genome-wide association studies (GWAS) of aging traits face a problem of missing genetic variance. This work supports the view that this problem can be overblown (Nat Rev Genet. 2010; **11**, 446–450). We use the Framingham Heart Study original (FHS) and Offspring (FHSO) cohorts to elucidate whether life course can substantially impact the role of lipid-related genes, the APOE e2/3/4 polymorphism and the APOB rs1042034 (C/T) SNP, in onset of cardiovascular disease (CVD). The APOE e4 allele and APOB CC genotype can play detrimental, neutral, and protective role in etiology of CVD in different ages and environments in sex-specific fashion. For example, the APOE e4 allele confers risk of CVD in younger-old women ($RR_{\leq 75yrs}=1.73$, $p=1.4\times10^{-3}$) but it can be cardio-protective in older women ($RR_{76+yrs}=0.68$,

p=0.021) in the same FHS cohort. Disregarding the role of aging results in insignificant effect (RR=0.84, p=0.143). The role of the e4 allele can change through generations (proxy for environment); the e4 allele in the FHSO only confers risks of CVD in older women ($RR_{>64yrs}$=1.57, p=6.6×10^{-3}). Just increasing sample of women by pooling data from the FHS and FHSO merely scales the life course heterogeneity in parallel (RR=1.03; p=0.701). The effects are stable longitudinally ensuring that they are not of stochastic nature and robust to longitudinal attrition of the samples at CVD risks. The results suggest that life course can play a key role in genetic predisposition to healthspan.

83

An Efficient Multiple Testing Strategy in Genome-Wide Association Studies Using a Multivariate Normal Block Design

Arunabha Majumdar (1) Saurabh Ghosh (1)
(1) Indian Statistical Institute

Genome-wide association studies have been partially successful in identifying novel variants involved in complex disorders. However, correcting for multiple testing in such studies becomes inevitable to maintain the appropriate overall false positive error rate. While different methods have been developed that control for family wise error rate (FWER) in genome-wide association studies, two prudent strategies have gained popularity because of their reduction in the computational burden: one based on the asymptotic multivariate normal distribution framework of the test statistics at the correlated SNPs as implemented in the method *SLIDE* (Han et al. 2009) and the other estimates the effective number of independent SNPs based on the principal components analysis (PCA) of the correlation matrix of the correlated SNPs as implemented in the method *simpleM* (Gao et al., 2008). We develop a block wise strategy *MVNblock* of multiple testing correction based on the asymptotic multivariate normal framework and investigate few of its important theoretical properties. We compare *MVNblock* with *simpleM* using extensive simulations and find that *simpleM* behaves more conservatively than *MVNblock* with respect to controlling for FWER. Moreover, *MVNblock* consistently produces a lower estimate of the effective number of independent SNPs compared to *simpleM*, subject to controlling FWER at the desired level, indicating that, *MVNblock* is expected to produce higher power compared to *simpleM*.

84

Multiple Testing Methods for Analyzing Rare Genetic Variants

Marinela Capanu (1) Venkatraman Seshan (1) Colin B Begg (1)
(1) Memorial Sloan-Kettering Cancer Center

With the advent of next generation sequencing, the identification of disease loci when there are multiple rare variants is the pre-eminent current technical challenge in statistical genetics. In previous work we have used hierarchical modeling techniques to estimate the relative risks of individual rare variants from a known risk gene. We now address the challenge of interpreting these results from the perspective of multiple testing, recognizing that our proposed hierarchical modeling approach implicitly aggregates the information from variants that share higher-order characteristics (e.g. bioinformatic characteristics). Consequently, the results for individual variants can be very strongly correlated, thus violating a crucial assumption of conventional multiple testing approaches. Nevertheless, since each specific variant is of crucial interest to the individuals and their family members who possess this specific variant, classifying each of these variants as deleterious versus neutral is a particularly important goal. Using simulations we examine the properties of different false discovery controlling procedures in this setting with the goal of optimizing the classification of rare variants as deleterious versus neutral. We illustrate the methods with an application to a real study of breast cancer.

85

An effective Multiple Testing Procedure for Association Studies Incorporating Admixture Mapping Information

Wenan Chen (1) Guimin Gao (1)
(1) Department of Biostatistics, Virginia Commonwealth University

Admixed populations (such as African Americans) denote the populations formed by recent admixture of two or more ancestral populations. For gene mapping in admixed populations, admixture mapping tests use admixture linkage disequilibrium (LD) can only identify a causal variant in a large chromosomal region (several Mbs). To identify a causal variant in a small region (< a few hundred Kbs), association tests that correct for local ancestry have been developed. However, these tests can have relatively low power. Recently, joint association tests that combine information from admixture mapping tests and association tests that correct for local ancestry have been proposed, but these methods may have inflated type I error rates when a test marker is located in a region with admixture signal but far from causal variants. A main reason for this is that these joint methods may use too much information from the admixture mapping at a test marker. In this study, we adapt the generalized sequential Bonferroni procedure (GSB) to association studies to incorporate appropriate amount of information from admixture mapping into association tests that correct for ancestry. Our simulation studies indicates that the GSB procedure not only controls family-wise error rates, but also can have improved power compared to the association tests that correct for local ancestry. We applied the GSB procedure to association studies in a data set from the Multi-Ethnic Study of Atherosclerosis project.

86

A General Bias-reduction Procedure to Address the Winner's Curse in Genetic Association Analysis: Evaluation for Time-to-Event Phenotypes

Julia Taleban (1) Laura L Faye (2) A Dimitromanolakis (1) Andrew D Paterson (3) Lei Sun (2) Shelley B Bull (4)
(1) Samuel Lunenfeld Research Institute of Mount Sinai Hospital, Toronto, Canada (2) Dalla Lana School of Public Health, University of Toronto (3) Genetics and Genome Biology, Hospital for Sick Children and University of Toronto (4) Samuel Lunenfeld Research Institute and Dalla Lana School of Public Health, University of Toronto

With the shift in genome-wide analyses from simple binary or quantitative traits to more complex phenotypes, existing methods are not well equipped to address the winner's curse. Motivated by investigation of the genetics of complications of type 1 diabetes in a rigorously followed cohort of participants in the DCCT/EDIC study (Al-Kateb et al. 2008 Diabetes 57;218), we investigate bias reduction methods for effect estimation in studies with time-to-event phenotypes, building on previous work in non-parametric bootstrap resampling (Faye et al. 2011 StatMed 30;1898), and incorporate survival models into the computationally feasible BR-squared software (Sun et al. 2011 HumGenet 129; 545). In analysis of time to nephropathy in 1362 white probands, examining 1213 SNPs in 201 candidate genes, reduction in log hazard ratio estimates ranges from 42.8 to 79.7%, depending on the SNP effect and minor allele frequency. In simulations based on the observed genotype data, bootstrap estimates are usually closer to the true effect size than uncorrected estimates, but the method tends to over-correct for SNPs with moderate power. Among false positives, the bootstrap shrinks estimates appropriately toward the null. This implementation reduces genome-wide selection bias in log hazard ratio estimates for analysis of time-to-event data under the Cox proportional hazards model, and demonstrates potential for extensions to other phenotypes and to methods adapted to low frequency variants.

87

Multiple Regression Strategies for Detecting Disease Associations with Poorly Tagged Causal Variants

Richard Howey (1) Heather J Cordell (1)
(1) Institute of Genetic Medicine, Newcastle University, UK

A popular method for the detection of disease-causing variants in genome-wide association studies (GWAS) is to individually test each SNP using a test derived from a logistic regression model. This approach relies on the assumption that a casual variant will be in strong linkage disequilibrium (LD) with at least one of the tested SNPs. Here we present a new method (and accompanying software) for improving the association signal in regions where there are combinations of SNPs, but no single genotyped SNP, in strong LD with the causal variant. Our approach is much quicker and less fiddly than competing methods such as haplotype-based methods or imputation, thus we hope it may be useful for performing a first-pass analysis across the genome, before proceeding to more complicated/computer-intensive analysis if deemed worthwhile. Our approach proceeds by selecting, for each genotyped "anchor" SNP, a nearby genotyped "partner" SNP (chosen, on the basis of a specific algorithm we have developed, to be the optimal partner). These two SNPs are then used as predictors in a logistic regression analysis, in order to generate a final significance test associated with the anchor SNP. The procedure is then repeated for each genotyped anchor SNP across the genome. We demonstrate via application to simulated and real data that our method provides a useful and complementary additional test for GWAS data.

88

Informed Conditioning on Clinical Covariates Increases Power in Case-control Association Studies

Noah Zaitlen (1) Sara Lindstrom (1) Bogdan Pasaniuc (1) Marilyn Cornelis (1) Giulio Genovese (2) Samuela Pollack (1) Anne Barton (3) Donald W Bowden (4) Steve Eyre (3) Barry I Freedman (4) John K Field (5) Leif Groop (6) Aage Haugen (7) Brian E Henderson (8) Pamela J Hicks (9) Lynne J Hocking (10) Laurence N Kolonel (11) Maria Teresa Landi (12) Carl D Langefeld (9) Loic Le Marchand (11) Michael Meister (13) Ann W Morgan (14) Olaide Y Raji (5) Angela Risch (15) David Scherf (15) Sophia Steer (16) Martin Walshaw (17) Kevin M Waters (18) Anthony G Wilson (19) Paul Wordsworth (20) Shanbeh Zienolddiny (21) Eric Tchetgen Tchetgen (1) Christopher Haiman (18) David J Hunter (1) Robert M Plenge (2) Jane Worthington (22) David C Christiani (1) Debra A Schaumberg (1) Daniel I Chasman (2) David Altshuler (23) Benjamin Voight (23) Peter Kraft (1) Nick Patterson (23) Alkes L Price (1)
(1) Harvard School of Public Health (2) Harvard Medical School (3) University of Manchester (4) Wake Forest School of Medicine (5) University of Liverpool (6) Scania University Hospital (7) National Institute for Occupational Health (8) University of Southern California (9) Wake Forest University (10) University of Aberdeen (11) University of Hawaii (12) NIH (13) Thoraxklinik am Universitatsklinikum (14) NIHR-Leeds Musculoskeletal Biomedical Research Unit (15) DKFZ-German Cancer Research Center (16) King's College Hospital (17) 23 Liverpool Heart and Chest Hospital (18) University of Southern California (19) University of Sheffield (20) Nuffield Orthopaedic Centre (21) National Institute of Occupational Health (22) The University of Manchester (23) Broad Institute

Genetic case-control association studies often include data on clinical covariates, such as body mass index, smoking status, or age, that may modify the underlying genetic risk of case or control samples. An unanswered question is how to optimally use this information to maximize statistical power in case-control association studies. Current approaches often lose power under case-control ascertainment and always fail to capture available power increases under case-control-covariate ascertainment such as age-matched designs. We show that an approach based on the liability threshold model with parameters informed by external epidemiological data, accounts for disease prevalence and ascertainment, and provides a substantial increase in power while maintaining a controlled false-positive rate. Our method outperforms standard case-control association tests and previously proposed tests for dealing with covariates in ascertained data. We investigate empirical case-control studies of type 2 diabetes, prostate cancer, lung cancer, breast cancer, rheumatoid arthritis, age-related macular degeneration, and end-stage kidney disease over a total of 89,726 samples. Informed conditioning outperformed logistic regression for 115 of the 157 known variants investigated ($pv=10^{-9}$), with a 16% median increase in χ^2 test statistics, and a commensurate increase in power. Applying our method to existing and future association studies of these diseases may identify novel disease loci.

89

Application of Multiple Regression with Singular Value Decomposition Method to Genome-wide Expression Quantitative Trait Loci Studies

Soonil Kwon (1) Jerome I. Rotter (1) Xiuqing Guo (1)
(1) Medical Genetics Institute, Cedars-Sinai Medical Center

Genome-wide expression quantitative trait loci (eQTL) studies attempt to identify single nucleotide polymorphisms (SNPs) that can be associated with expression levels. Up to millions of SNPs and thousands of gene expressions can be involved in eQTL studies. However, sample size is usually limited and much smaller than the number of SNPs. Most eQTL studies perform single SNP testing for each expression level, leading to multiple testing problems. We proposed here the multiple regression with singular value decomposition method which can analyze all SNPs simultaneously. To examine the validity of the method, we simulated 10 sets of data, each has 20 samples, 100 expression levels, and 1000 SNPs. We made the 10^{th} expression level to be associated with the 80^{th}, 90^{th}, 100^{th}, 110^{th}, and 120^{th} SNPs; the 50^{th} expression level associated with the 480^{th}, 490^{th}, 500^{th}, 510^{th}, and 520^{th} SNPs; the 90^{th} expression level associated with 880^{th}, 890^{th}, 900^{th}, 910^{th}, and 920^{th} SNPs. All associated SNPs were successfully identified, indicating that the proposed method might be a useful tool in eQTL studies when sample size is much smaller than the number of SNPs.

90

Adjustment of Covariates in Genetic Association Analysis using Propensity from Decision Trees

Yaji Xu (1) Epiphanie Nyirabahizi (1) Heping Zhang (1)
(1) Yale University

Propensity scores are usually estimated using parametric models such as logistic regression and then used as a matching or stratification criterion in an observational study. However, in practice, it is not always clear as to what variables we should include in computing propensity scores and what forms of the variables we should consider. Thus, nonparametric classifiers such as decision trees can offer a flexible method to overcome this practical problem by constructing strata to adjust for covariates in genetic association analysis. Simulation studies show that the tree-based stratification method is more robust than the parametric approach unless know and use the parametric model underlying the propensity scores, which is unrealistic in practice. We applied our method to Genetic Association Information Network-Major Depressive Disorder (GAIN-MDD) data, and our analysis demonstrates the usefulness of our proposed approach in assessing gene and environment factors for complex diseases.

91

Accounting for Control Mislabelling in Case-control Biomarker Studies

Mattias Rantalainen (1) Chris Holmes (1)
(1) Department of Statistics, University of Oxford

Uncertainty in case and control labels is often overlooked in biomarker discovery studies. Omitting to take into account label uncertainty can lead to bias in model parameter estimates and predictive risk that under some conditions is substantial. One common situation leading to label uncertainty is when the set of control subjects contain an unknown number of undiagnosed cases. This can occur in biomarker discovery studies in the context of genomic epidemiology when control subjects are sampled from the general population. Contamination of undiagnosed cases in the control group have the largest impact in situations where the model need to be well calibrated, for example, in the evaluation of biomarker panels. Failing to account for class label uncertainty may lead to underestimation of classification performance as well as bias in parameter estimates, which can also impact meta-analysis.
Here we describe how the commonly used logistic regression model can easily be modified to address class label uncertainty and fitted through Expectation-Maximization. Through a simulation study we evaluate how the conventional model and the proposed model perform in estimation of predictive risk, evaluation of classification performance and in meta-analysis under label uncertainty. Our results suggest that taking into account label uncertainty lead to well-calibrated prediction performance estimates and unbiased meta-analysis results under label uncertainty.

92

Testing for the Presence of Liability Models

Christine Herold (1) Tatsiana Vaitsiakhovich (2) Tim Becker (3)
(1) Harvard School of Public Health, Boston, USA; German Center for Neurodegenerative Diseases, Bonn, Germany (2) Institute for Medical Biometry, Informatics and Epidemiology, Bonn, Germany (3) German Center for Neurodegenerative Diseases; Institute for Medical Biometry, Informatics and Epidemiology, Bonn, Germany

It has recently been suggested that so-called liability models may define the architecture of the genetics of complex diseases [Zuk et al., 2012]. For individuals with less than k risk alleles a moderate baseline penetrance f_0 <=5% applies, whereas for individuals with an allele load above the threshold the penetrance $f_{>k}$ can be as high as 50%-80%. Such models in general do not lead to epistasis that is detectable by pair-wise testing but exhibit interaction of all higher orders. For a phenotype with m a priori confirmed SNPs, it is an important question whether they follow a threshold model. To this purpose, we present a simple 1 d.f. regression test. We conducted an extensive simulation for the threshold models with n SNPs, n$\in\{5,10,20,50\}$ and liability threshold corresponding to a portion of 70%–85% causal alleles. Our models lead to marginal effects that are compatible with typical GWAS effect sizes and pair-wise interaction that is hard to detect. However, with moderate sample size (3000 individuals) the power to detect the presence of a threshold model is effectively 100%, given that all SNPs of the threshold model are known. For the more realistic situation, in which the number of already detected SNPs m is smaller than the number n of SNPs in the unknown model, we can show that for $m>2/3n$ there is decent power to proof the involvement in a threshold model. We present results for Alzheimer's and Type II Diabetes GWAS data sets.

93

Unraveling Phenotype Heterogeneity in Prostate Cancer Susceptibility in Finland Utilizing Covariate-Based Analysis

Cheryl D. Cropp (1) Claire L. Simpson (1) Tiina Wahlfors (2) Asha George (3) MaryPat S. Jones (4) Ursula Harper (4) Damaris Ponciano-Jackson (4) Teuvo Tammela (5) Johanna Schleutker (6) Joan E. Bailey-Wilson (1)

(1) Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD (2) Institute of Biomedical Technology/BioMediTech, University of Tampere, Tampere, Finland (3) Fox Chase Cancer Center, Philadelphia, PA (4) Genomics Core/Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Rockville, MD (5) Department of Urology, Tampere University Hospital, University of Tampere, Tampere, Finland (6) Department of Medical Biochemistry and Genetics, University of Turku, Turku, Finland

Prostate cancer is the most common male cancer in developed countries. Previously, we reported a genome-wide linkage scan in 69 Finnish Hereditary Prostate Cancer (HPC) families, which replicated the *HPC9* locus on 17q21-q22 and identified a locus on 2q37. We used ordered subset analysis (OSA) to detect other loci linked to HPC in subsets of families to identify and detect other loci linked to HPC incorporating age of onset as a trait-related covariate to address genetic heterogeneity and strengthen the linkage findings previously reported. The overall mean age of onset across the families was 66.2 ± 8.8 years while the range of individual onset ages ranged from 46 to 98 years. Although the highest OSA LOD score with a ΔLOD (p=0.02) was 2.876 on 15q26.2-q26.3 in a subset of 40 families ascending by age at onset, no other ΔLOD scores were significant after permutation testing. Since OSA uses a single covariate value per family, we used mean age of onset for each pedigree. To better capture the effect of age on the linkage signal, we used LODPAL to perform a linkage analysis in affected relative pairs, while adjusting for the age of each individual family member as a single covariate. Preliminary results revealed strong evidence of linkage to HPC on chromosome 15q was (LOD=4.9, 132cM) and 8q (LOD=3.1, 157cM). Permutations are ongoing to determine empirical p-values for these LOD scores.

94

Sparse Principal Component Regression as a Tool to Detect Causal Regions in Genetic Studies

Stefan Konigorski (1) Rafal Kustra (1)
(1) Dalla Lana School of Public Health, University of Toronto

In genetic studies of phenotypes such as quantitative traits or disease status, it is of interest to identify regions containing causal loci. We propose a multivariate approach to identify such regions, which takes the linkage disequilibrium (LD) structure of the data into account. Regions are defined as LD blocks containing consecutive SNPs in high LD. In the first step of the method, the dimension of the data is reduced by LD block thresholding of SNPs based on the minP approach and by sparse principal component analysis. The latter allows to capture the signal of the regions in the sparse components in such a way that they can be found as significant predictors in a subsequent regression analysis. Results from analyzing simulated SNPs array datasets with up to 100,000 SNPs and binary outcome variables show that the proposed approach is able to identify almost all causal LD blocks with only a small number of false positives. Application to a real genome-wide association study, as well as limitations and modifications to improve performance are discussed.

95

Relationship Between the APOE Polymorphism and Risks of Cancer and Alzheimer's Disease: Application of the Genetic Stochastic Process Model

Konstantin G. Arbeev (1) Svetlana V. Ukraintseva (1) Alexander M. Kulminski (1) Liubov S. Arbeeva (1) Igor Akushevich (1) Irina V. Culminskaya (1) Deqing Wu (1) Anatoliy I. Yashin (1)
(1) Duke University

Negative dependence between mortality from and risks of cancer and Alzheimer's disease (AD) has been shown in our recent studies (Mech Ageing Dev 130: 98–104, 2009; Rejuv Res 13: 387–396, 2010). It is likely that genetic factors play a crucial role in this dependence affecting aging-related mechanisms involved in shaping the incidence rates of these diseases. To test this hypothesis, we applied the genetic stochastic process model (J Theor Biol 258: 103–111, 2009) to data on incidence of cancer and AD and age trajectories of physiological variables in carriers of different apolipoprotein E polymorphisms in the Framingham Heart Study. We performed new integrated analyses of data on individuals with and without genetic information. We found substantial difference in age patterns of aging-related biomarkers of carriers and non-carriers of the e4 allele. This difference can be explained by different patterns of aging-related decline in adaptive capacity and stress resistance (associated with the narrowing of the U-shape of risks as functions of physiological variables), mean allostatic trajectories (the trajectories of the variables that organisms are forced to follow by the process of allostatic adaptation) and physiological "norms" (the age-specific values of variables minimizing risks of onset of the diseases). These results indicate the presence of a strong genetic component in mechanisms of aging-related changes which contribute to the relationship between the two diseases.

96

Genetic Mapping Using the Theory of the Added Variable Plot in the Mixed Models

Nubia Duarte (1) Julia Soler (2) Mariza de Andrade (3) Suely Giolo (4) Alexandre Pereira (1)
(1) Heart Institute (INCOR) of University of Sao Paulo (2) Institute of Matematics and Statistics university of Sao Paulo (3) Division of Biostatistics, Mayo Clinic Cancer Center – Rochester, MN, USA (4) University of Curitiba Parana-Brazil

Recently, one of the most important problems in genetics is the identification of genes associated with complex diseases. A useful design for this proposal corresponds to collect data from extended families and molecular markers platforms SNPs (Single Nucleotide polymorphism). These platforms represent points of reference strategically placed along the genome of the individuals and are high dimensional. Analysis of these data brings analytical challenges as the problem of multiple testing and selection of predictive variables.
In this thesis, we propose a criterion for discriminating predictors of genetic effects due to random polygenic component and the residual component, under the framework of a linear mixed model. Also, considering that the individual effects of predictor variables is expected to be small, it is suggested a method for finding ordered subsets of these variables and study their simultaneous effect on the response

variable under study. In this context, is used the theory of the added variable plot under a mixed model framework. The proposals are validated through a simulation study, which is based on structures of families involved in the Project "Baependi Heart Study" (FAPESP Process 2007/58150-7), whose objective is to identify genes associated with cardiovascular risk factors in the Brazilian population. This proposal is implemented by using the R statistical environment and for the simulation of genetic predictors is adopted the SimPed application.

97

Discovery and Fine-mapping of Type 2 Diabetes Susceptibility Loci Through Trans-ethnic Meta-analysis

Anubha Mahajan (1) Jennifer E Below (2) Weihua Zhang (3) Min Jin Go (4) Esteban J Parra (5) Andrew P Morris (1) on behalf of the AGEN-T2D, DIAGRAM, MA-T2D and SA-T2D Consortia (6)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, UK (2) Departments of Human Genetics and Medicine, The University of Chicago, Chicago, Illinois, USA (3) Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, St Mary's Campus, London, UK (4) Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Chungcheongbuk-do, Korea (5) Department of Anthropology, University of Toronto at Mississauga, 3359 Mississauga Road North, Mississauga, ON, Canada (6)

We performed genome-wide trans-ethnic meta-analyses in 26,488 type 2 diabetes (T2D) cases and 83,964 controls from populations of European, East Asian, South Asian and Mexican American ancestry, at ~2.5 million autosomal SNPs. We identified a novel T2D susceptibility locus at genome-wide significance ($p<5\times10^{-8}$), mapping to *TMEM154* ($p=4.2\times10^{-9}$). We observed nominal evidence of heterogeneity in allelic effects (Cochran's *Q*-statistic $p<10^{-3}$) at lead SNPs at just 3 of the 55 established autosomal T2D susceptibility loci: *KLF14* ($p=1.4\times10^{-5}$), *HNF4A* ($p=4.0\times10^{-5}$) and *PEPD* ($p=7.1\times10^{-4}$). We constructed credible sets of SNPs that encompass 95% of the posterior probability of being causal (or tagging an unobserved causal variant) across 30 T2D susceptibility loci previously identified in European ancestry genome-wide association studies. We compared genomic interval covered by the credible set in the trans-ethnic and the European ancestry only (12,171 cases and 56,862 controls) MANTRA meta-analyses. Fine-mapping resolution was improved by the addition of non-European ancestry GWAS at 23 of the loci, most notably at *KCNJ11*, where the credible set of SNPs was reduced from 134 (959kb) to just 1. This SNP, rs5215, is in strong LD with the previously implicated E23K variant, not reported in the European ancestry meta-analysis. However, the trans-ethnic credible set excludes another implicated variant at this locus, A1369S, in the *ABCC8* gene.

98

Joint Statistical Modeling of Multiple Phenotypes in Samples with Related Individuals

Zuoheng Wang (1)
(1) Yale University

Genetic association studies have routinely been conducted to search for variants associated with diseases and quantitative phenotypes. Clinical and epidemiological studies typically collect data on a set of correlated phenotypes that may share common environmental and/or genetic factors. Such phenotypes contain more information than univariate phenotypes. Thus joint modeling of multiple phenotypes can potentially have increased power to detect association and increased precision of parameter estimation than univariate analysis. In this study, we develop novel statistical methods for multivariate association mapping in samples that contain arbitrarily related individuals. We address both common and rare variants association. The proposed methods are based on retrospective analysis that is less dependent on model assumptions on phenotypes, thus they are robust to trait model misspecification. The new methods can accommodate the external biological information by integrating graphical models and multivariate analysis, and are computationally affordable. We comprehensively evaluate the proposed methods using simulation studies and compare with existing methods. The proposed methods are applied to analysis of blood lipid levels in a whole genome sequence study on the Sardinian population.

99

Propensity Score Adjusted Association Tests for Multiple Traits Based on the Generalized Kendall's Tau

Yuan Jiang (1) Ni Li (2) Zhifa Liu (3) Heping Zhang (3)
(1) Oregon State University (2) Hainan Normal University (3) Yale University

Covariates such as environmental factors play an important yet complicated role in the association analysis between genetic risk factors and the outcomes. Although parametric methods have been developed to adjust for covariates in association analysis, difficulties arise when the traits are multivariate and of varying scales because there is no ready-to-use model for them. Recent nonparametric development includes U-statistics to measure the phenotype-genotype association weighted by a similarity score of covariates. However, it is still not clear how to optimize the similarity score. In this work, we propose a more natural and convenient U-statistic measurement adjusted by propensity scores using the idea of inverse probability weighting (IPW). The new U-statistic is shown to be unbiased under our null hypothesis while the previous ones are not. As a byproduct, we find that (consistently) estimated propensity scores are preferable to true propensity scores as the formers result in a smaller variance of our U-statistic. This finding agrees with a few previous studies in propensity score applications but it is the first time to be formalized in the framework of association tests. Simulation results show that our test improves power as opposed to the non-weighted and two weighted U-statistic methods. Finally, we apply our proposed test to the Study of Addiction: Genetics and Environment (SAGE) to demonstrate its usefulness in genome-wide association studies.

100

Construction of A Metabolic Syndrome Scale and Joint Phenotype Heritability Using Item Response Theory Models

Tiago M. Fragoso (1) Suely R. Giolo (2) Mariza de Andrade (3) Alexandre Pereira (4) Julia P Soler (1)
(1) Universidade de Sao Paulo (2) Universidade Federal do Parana (3) Mayo Clinic (4) InCor, Universidade de Sao Paulo

The metabolic syndrome is composed of a series of phenotypes that jointly contribute for cardiovascular disease and diabetes. Studies of its heritability are conducted analyzing it individually for each symptom, sometimes using dichotomized variables corresponding to defining criteria adopted by some health authority. In this work, we construct a metabolic syndrome scale using Item Response Theory (IRT) models for the dichotomized phenotypes that takes all measured phenotypes into consideration, allowing informative comparison between individuals. A welcome consequence of the IRT framework is that we derive a straightforward joint heritability parameter that can be easily interpretable as the heritability of the latent trait representing the individual metabolic syndrome level. Parameter estimation was conducted under a Bayesian statistics framework, using Markov Chain Monte Carlo (MCMC) type algorithms implemented as a script in the R Statistical Software. Simulation studies were conducted to evaluate recovery of the phenotype location and metabolic syndrome heritability parameters, indicating adequate recovery. The metabolic syndrome is first validated in comparison with a Factor Analysis, obtaining a high correlation between both constructs and then applied to a sample of 1.666 individuals obtained in the Baependi Heart Study described in Oliveira et al. (2008, BMC Medical Genetics 32 (9))

101

Joint Modeling of Disease and Endophenotype to Characterize the Effect of Genes and Their Interactions

Alexandre Bureau (1) Jordie Croteau (2)
(1) Universite Laval (2) Centre de recherche de l'Institut universitaire en sante mentale de Quebec

Endophenotypes are traits related to a disease and believed to be influenced by fewer genes. The presence of impairments on an endophenotype in affected and non-affected relatives can help tracing familial transmission of alleles of genes interacting to cause a disease. To exploit that information, we propose jointly modeling a disease and an endophenotype in function of two genes. When the endophenotype is coded as presence/absence of impairment, the disease and endophenotype form four phenotypic levels. We implemented a within-family association score test conditional on phenotype by extending the generalized disequilibrium test (GDT) to a polytomous phenotype. We compared various strategies to test a marker via simulation: a general test of the four phenotypic levels, a specific model of a locus conferring susceptibility to the disease only when an endophenotype impairment and a risk genotype at a known locus are present, and the GDT on the disease or endophenotype alone. Under two-gene scenarios, the tests of joint models of disease and endophenotype conditional on a known locus were more powerful than tests of association to the disease or endophenotype alone. Conditioning on a known locus provided more power than testing a marker alone. Power under joint models remained competitive in presence of strong main effects. The tests were applied to candidate SNPs, cognitive endophenotypes and schizophrenia and bipolar disorder in large kindreds from Eastern Quebec.

102

Genetic Association Analysis of Complex Diseases Incorporating Intermediate Phenotype Information

Yafang Li (1) Jian Huang (2) Chris Amos (1)
(1) MD Anderson Cancer Center (2) University of Iowa

Genetic researchers often collect disease related quantitative traits in addition to disease status in genome-wide association (GWA) studies, statistical tests combining both disease status and intermediate phenotypes information should be more powerful than case-control studies, as the former incorporates more information about the disease. We proposed a modified inverse-variance weighted meta-analysis method to combine disease status and quantitative intermediate phenotype information. The simulation results showed that when an intermediate phenotype was available, the inverse-variance weighted method had more power than did a case-control study in a GWA study of complex diseases, especially in identifying susceptibility loci having minor effects. We further applied this modified meta-analysis to a study of imputed lung cancer genotypes with smoking data in1154 cases and 1137 matched controls. The most significant SNPs came from the *CHRNA3-CHRNA5-CHRNB4* region on chromosome 15q24–25, our results confirm that this *CHRNA* region is associated with both lung cancer development and smoking behavior. We also detected three significant SNPs–rs1800469, rs1982072, and rs2241714–in the promoter region of the *TGFB1* gene on chromosome 19 (p=1.46e-5, 1.18e-5, and 6.57e-6, respectively). The SNP rs1800469 is reported to be associated with chronic obstructive pulmonary disease and lung cancer in cigarette smokers.

103

Comparison of Methods to Analyze Multiple Phenotypes and Longitudinal Data

Marianne Huebner (1) Mariza de Andrade (1) Syed H Arshad (2) Susan Ewart (3)
(1) Mayo Clinic (2) University of Southampton (3) Michigan State University

SNP selection procedures to identify genetic risk factors for a single phenotype include random forests, elastic net, two step procedures with mixed effects models, or repeated measures analysis. Methods for correlated phenotypes include testing predefined phenotype combinations or identifying phenotype sets using permutation tests or structural equation models. We compare several statistical methods for multiple phenotypes in an application to an allergy and asthma data set. Phenotypic information from the Isle of Wight Birth Cohort (n=1172 with genotype information) was collected at 4, 10, and 18 years of age. This included binary outcomes such as asthma, eczema, rhinitis, and skin prick test, as well as continuous outcomes such as lung function tests and total IgE at 10 and 18 years. For the methodswe focus on the cytokines IL-4 and IL-13, the cytokine receptor IL-4R, and the transcription factor GATA3, which are important in the allergic immune response that underlies asthma and other allergies.

104

Joint Modeling of Repeated Quantitative Trait Measures and Time to Event in Longitudinal Genetic Association Studies

Zhijian Chen (1) Andrew D Paterson (2) Angelo J Canty (3) Lei Sun (4) Shelley B Bull (1)
(1) Samuel Lunenfeld Research Institute (2) The Hospital for Sick Children (3) McMaster University (4) University of Toronto

In longitudinal study designs, it is typical to examine SNP associations in separate analyses of repeated QT measures and time to a disease-related event. As an alternative, we consider joint modeling, in which the repeated QT measure is also a time-dependent covariate in the survival model. Based on data from a recent GWAS of HbA1c, a quantitative measure of glycemia strongly related to the risk of diabetic retinopathy, a complication of type 1 diabetes (Paterson et al., 2010): we compare estimates and hypothesis testing results from joint likelihood maximization to those from separate analyses. The longitudinal component of the joint model, corresponding to genetic association with HbA1c, consists of a linear mixed effects model. It is linked to the survival component through the smoothed HbA1c trajectory function which is included as a time-dependent covariate associated with the risk of complications. Furthermore, because contribution to the current risk of complications may be larger for prior than for concurrent HbA1c measures, we extend the joint analysis method to incorporate a pre-specified weight function for the cumulative QT effect, and apply two-stage estimation. Advantages of joint modeling include inference for SNP association that can help to distinguish among alternative genetic architectures, and improved precision in genetic association estimates. The approach can be easily adapted to multiple-SNP association analysis across the genome or within a region.

105

GWAS of Repeated Lipid Measures in Type 1 Diabetes Identifies a Novel Locus for Low-density Lipoprotein Cholesterol

Angelo Canty (1) Tao Wang (1) Shelley B Bull (2) Lei Sun (3) Andrew B Boright (3) DCCT/EDIC Research Group (4) Andrew D Paterson (5)
(1) McMaster University (2) Lunenfeld (3) U Toronto (4) GWU (5) Sickkids

Abnormal cholesterol (C) levels are an important risk factor for cardiovascular disease in people with and without diabetes. Recent studies have identified numerous loci for lipid levels in people without diabetes, but it is not clear whether the same loci are relevant for those with diabetes. Participants (n=1,303) with T1D from the DCCT study had lipids measured annually for a mean of 6 years. We used linear mixed models to identify loci that are associated with repeated measures of lipids in T1D, after adjustment for covariates, including repeated measures of glycemia. Known loci (CEPT for HDL-C; APOE and LDLR for LDL-C) were identified at genome-wide criteria ($p<5\times10–8$). rs10866235 (chr 4; 181.4Mb, intergenic, MAF=38%) was associated with LDL-C ($p=6 \times 10–8$) in a model that adjusted for gender, retinopathy at baseline, intensive treatment, age at diagnosis and duration of diabetes. Results were modestly sensitive to excluding covariates or including time-dependent glycemia ($p=2 \times 10–7$ and $p=1 \times 10–6$, respectively). This locus was not associated with LDL-C (p=0.65) in meta-GWAS of 95,454 non-diabetic individuals (Teslovich et al., 2010 Nature, 466; 707). Intensive insulin treatment of people with type 1 diabetes (T1D), with the goal of normalizing glycemia, resulted in significantly lower LDL-C compared to conventional treatment. Further modeling will determine whether there is evidence for SNP*glycemia interaction on LDL.

Withdrawn abstract 106

107

An Alternative Analysis of Secondary Phenotype Data in Case-Control Genetic Association Studies

Sharon Marie Lutz (1) John E Hokanson (1) Christoph Lange (2)
(1) University of Colorado, Anschutz Medical Campus (2) Harvard School of Public Health

In case-control association studies, secondary phenotypes are often collected that are correlated with case-control status. Secondary phenotypes can provide valuable insight into the disease of interest. Standard statistical analysis of secondary phenotypes can be problematic due to unequal selection probabilities between cases and controls. Ling and Zeng present a method to properly analyze secondary phenotype data in case-control association studies using a likelihood based approach that models the probability of the secondary phenotype and the genetic information given case-control status. In this article, we present an alternative and simpler way to model the probability of the secondary phenotype and the genetic information given case-control status. This method makes no assumptions about the distribution of the secondary phenotype. As a result, this method works well on complex secondary phenotypes that are correlated with case-control status. The approach is illustrated by an application to a genome-wide association study for Chronic Obstructive Pulmonary Disease (COPDGene) that contains numerous secondary, quantitative phenotypes that are correlated with case-control status. The code to implement this method is readily available.

References:
[1] Lin, D.Y., Zeng, D. (2009). Proper Analysis of Secondary Phenotype Data in Case-Control Association Studies. Genet Epidemiol 33(3):256–265.

108

Visualizing High-dimensional Prediction Problems with Application to Automated Syndrome Classification

Brunilda Balliu (1) Stefan Boehringer (1)
(1) Leiden University Medical Centre

Clinical evaluation of children with developmental delay remains a challenge for clinicians due to a process based on experience involving informal criteria. The face provides important information to diagnose a syndrome. Computer-based techniques for analyzing 2D images of the face have been developed to help narrow down possible diagnoses. Here, we focus on simultaneous classification of 14 syndromes and analyze photographs of 205 individuals

based on 48 coordinate pairs (CP) available for each photograph.
Our strategy is to compute derived measures from raw data that are then used as predictors. We arrive at 1044 predictors: all possible Euclidean distances between CP, areas and angles of a triangulation and asymmetry scores for CP. Although we have created an n < p scenario, we can improve classifcation rates as compared to previous studies. We achieve an overall classification accuracy of 64% and a mean pairwise accuracy of 91%. Model selection is achieved through multinomial regression with elastic net penalty. A second advantage of our data transformations is the ability to visualize the decision process by means of visual maps corresponding to specific predictors with geometric interpretations. These techniques generalize to other high-dimensional prediction problems with geometric interpretations such as biological pathways, expression networks or genetic predictors combined with LD-distances.

109
Nonparametric Latent Variable Framework for Estimation and Prediction
Won H Lee (1) David V Conti (1)
(1) University of Southern California

Latent variable (LV) models estimate unmeasured constructs underlying observed measurements. A fundamental framework of three models is (Thomas, Lifetime Data Anal, 13(4), 2007, 565–81): 1) process (PM-LV regressed on a formative exposure); 2) measurement (MM-a reflective indicator regressed on LV); 3) disease (DM-the outcome regressed on LV). PM and MM have the potential to capture unmeasured effects from imperfect metrics and for dimension reduction. In DM, LV can be a more robust predictor of the outcome. This typically constrains LV to a single distribution. We present a nonparametric modification using the Dirichlet process (DP), a mixture of parametric distributions allocating observations into clusters with discrete distributions. Thus, LV can be flexibly estimated as a multimodal distribution of cluster effects. In simulations and using smoking cessation clinical trial data (Lerman, Neuropsychopharmacology, 31(1), 2006, 231–42) we make comparisons to conventional models to assess LV estimation, the LV-outcome association, and outcome prediction. CYP2A6 is a crucial gene in the metabolism of nicotine. Nicotine metabolite ratio (NMR) is a reliable marker of this process and associated with smoking abstinence. In simulations, our framework has better predictive abilities than conventional models, and reliably estimates parameters and LV. In the real data, our framework is comparable to conventional models, and able to identify underlying groups of nicotine metabolizers.

110
Can We Build the Model to Predict Cancer-associated Genes?
Ivan Gorlov (1) Christopher Amos (1) Olga Gorlova (1) Christopher Logothetis (1)
(1) The University of Texas MD Anderson Cancer Center

There was a substantial progress in identification of cancer related genes. It is also apparent that not all cancer related genes have been identified yet. Can we use the known cancer genes to develop and train a statistical model to predict novel cancer genes? We believe that the answer to this question is: "Yes, we can".
We focused on prostate cancer (PCa). First we have identified known PCa genes using literature mining. A forward stepwise binary logistic regression model was used to discriminate known PCa genes from non-PCa genes, based on such predictors as the level of evolutionary conservation of the gene, expression level in normal tissue, gene ontology annotation, posttranslational modifications, expression level in adjacent normal and tumor tissue, tissue specificity index and other characteristics (31 variables in total). We were able to correctly predict 96% of the non-PCa genes and 48% of PCa genes. The proportion of genes predicted to be PCa associated was significantly higher among putative prostate cancer genes compared to the genes without any evidence being PCa related. To validate the model we randomly chose 100 genes as "mock" PCa-related and built the predictive models using the same variables as we have used for "real" PCa genes. We found that we could correctly predict only 0.5% of the mock PCa genes on average. We found that the model also predicted known breast and lung cancer genes though the accuracy was not as good as for PCa genes.

111
Prostate Cancer Risk Prediction Using a Genetic Profile in an International Consortium (PRACTICAL)
Ali Amin Al Olama (1) Sara Benlloch (1) Daniel A. Leongamornlert (2) Edward J. Saunders (2) Malgorzata Tymrakiewicz (2) Michelle Guy (2) Koveela Govindasami (2) Zsofia Kote-Jarai (2) Rosalind A. Eeles (3) Douglas F. Easton (1)
(1) University of Cambridge (2) The Institute of Cancer Research (3) The Institute of Cancer Research & Royal Marsden NHS Foundation Trust

Genome-wide association studies have identified multiple genetic variants associated with prostate cancer risk, opening up the possibility for using genetic profiling in risk prediction. To evaluate the potential for risk prediction, we genotyped 25 known prostate cancer susceptibility SNPs in 31 case-control studies in an international consortium (PRACTICAL), and analysed data from 20,132 prostate cancer cases and 20,234 male controls of European ancestry. Odds ratios associated with each SNP genotype, and genotypes for pairs of SNPs, were estimated using unconditional logistic regression. Based on the assumption of a log-additive model, we constructed a risk score from the summed genotypes weighted by the per-allele log-odds ratios. Using this score, the top 1% of the population had an estimated increased risk of 49 fold compared with the bottom 1% of the population, and 4.7 fold compared with the average population risk, while the bottom 1% of the population had an estimated risk of 10% of the population risk. We used the effect sizes from each risk group to derive absolute risks of prostate cancer by age. Based on this analysis, the absolute risk by age 80 is 34% for a man in the top 1% of the risk distribution and 1% for a man in the lowest 1%. These results demonstrate that genetic risk profiling may play an important role in targeted prevention.

112
Using Random Forests for Consistent Probability Estimation in Whole Genome Association Studies
Jochen Kruppa (1) Inke R. Konig (1) Andreas Ziegler (1)
(1) Universitat zu Lubeck

Most machine learning approaches only provide a classification for binary responses. However, probabilities are required for risk estimation using individual patient characteristics. Specifically, the individual probability for a positive therapy response is of great interest in personalized medicine. Logistic regression is often utilized for individual probability estimation. However, logistic regression requires a correctly specified parametric model to consistently estimate probabilities. By embedding Random Forests into the class of nonparametric regression models, it has recently been proven that Random Forests provide consistent probability estimates for dichotomous outcome. In this work we show how regression random forests can be used for consistent estimation of individual probabilities.
We demonstrate probability estimation in Random Jungle (Schwarz 2010 Bioinformatics), a novel Random Forest application in C++ with a generalized framework. We illustrate probability estimation by real data analysis using GAW16 data from the North American Rheumatoid Arthritis Consortium (NARAC) and compare the performance of regression random forests with different other probability estimation rules like the logistic regression or the least absolute shrinkage and selection operator (lasso) approach.
Schwarz et al. (2010) Bioinformatics 26:14 1752–1758

113
A Study of Adaptive Boosting Methods in Presence of Large Number of Noise Signals in Genetic Association Studies
Wei Yang (1) Charles Gu (1)
(1) Department of Biostatistics, Washington University in St Louis

Complex diseases are caused by multi-factorial interplay of genetic and environmental factors. Findings from past genome-wide association studies (GWAS) are largely limited to marginal effects of individual SNPs that explain only a small fraction of disease heritability. Therefore, sophisticated methods are needed to test for collective genetic effects. Boosting is a machine learning method that gains predictive power through an ensemble of simple classifiers and increased weighting of harder samples. Previous experiments applying boosting to GWAS found it under-powered in presence of vast number of noise signals. Here, we study a novel localized adaptive boosting (aBoost) method that (1) construct each base learner using selected subset of variables, and (2) shaves off least important variables iteratively to reduce noise. Performance of aBoost is compared with model-based boosting (mboost) and random forest in a simulation study of 1,000 individuals and 1,000 SNPs, with 6 causal SNPs contributing to the disease risks based on various models of marginal and interaction effects. Initial results show that aBoost is capable to capture risk alleles as well as interaction effects with less false positives. Extensive evaluation is underway with different tuning parameters. Further improvement of aBoost will be achieved via integration with multi-locus and haplotype-based methods to improve power and to accommodate rare variants.

114
A Flexible Learning Classifier System for Classification and Data Mining in Genetic Epidemiology
Ryan J Urbanowicz (1) Jason H Moore (1)
(1) Dartmouth College

The nature of common disease etiology is now widely considered to be complex, multivariate, and heterogeneous. Certain statistical phenomena (i.e. epistasis and heterogeneity) confound our ability to connect predictive factors with disease phenotype. Previously, we explored the application of existing Learning Classifier Systems (LCS) as classification and data mining tools to address this need, and separately introduced heuristics to improve its performance within this epidemiological problem domain. We successfully demonstrated the unique potential of these systems to detect and characterize statistical patterns of epistasis and heterogeneity within simulated genetic data. In the present study we introduce and evaluate our own LCS algorithm designed specifically to address problems in genetic epidemiology. Specifically, we implement and evaluate a supervised learning LCS which combines expert knowledge weighting with attribute tracking with feedback. Additionally, we incorporate a flexible rule representation which confers the ability to learn from data having both discrete and continuous variables. In combination, this algorithm offers researchers a powerful and flexible tool for identifying predictive factors, complex patterns of association, and sample subgroups with heterogeneous patterns of association even in the context of variable data types and complex multivariate interaction.

115
A Two-stage Random Forest Approach to Identify Genetic Variants Using Recombination Hotspot Information
Silke Szymczak (1) Qing Li (1) Yoonhee Kim (2) Abhijit Dasgupta (3) James D Malley (4) Joan E Bailey-Wilson (1)
(1) Statistical Genetics Section, Inherited Disease Research Branch, National Human Genome Research Institute, NIH (2) Genometrics Section, Inherited Disease Research Branch, National Human Genome Research Institute, NIH (3) Clinical Sciences Section, Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH (4) Center for Computational Bioscience, Center for Information Technology, NIH

Random forest is a promising machine learning approach to jointly analyze many genetic variants for association with a disease. However, identification of true risk single-nucleotide polymorphisms (SNPs) that are in strong linkage disequilibrium (LD) with non-risk SNPs is challenging. Large regions of strong LD can lower the ranking of all SNPs in a causal region since they can serve as proxies for each other in different trees.
We propose a two-stage approach that uses information about recombination hotspots in the genome. In a first step, a random forest is trained for each region between two hotspots. Predicted case-control status or estimated probability of case/control status based on each region's SNPs then replaces the SNPs as predictor variables for a genome-wide random forest in the second stage.
We compare our approach with a random forest analysis using all SNPs simultaneously based on simulated GWAS data mimicking local LD patterns observed in European samples of the 1,000 genomes project. Genotype data for 500K SNPs

in 5,000 cases and 5,000 controls are generated using the software GWAsimulator. We compare the false-positive errors under a null model with no true genetic effects. For a power analysis several variants with different minor allele frequencies and in regions with different local LD structure are modeled as true risk SNPs with independent effects on the disease status.

116

A PCA-based Generalized Multifactor Reduction Method for Correcting Population Stratification

Xiang-Yang Lou (1) Guo-Bo Chen (1) Lei Yan (2) Nianjun Liu (1) Yann C. Klimentidis (1) Xiaofeng Zhu (3) Degui Zhi (1) Xujing Wang (1)
(1) University of Alabama at Birmingham (2) University of Tennessee Health Science Center (3) Case Western Reserve University

Gene-gene and gene-environment interactions govern a substantial portion of the variation in complex traits and diseases. Detection of multifactor interactions poses one of the great challenges to today's genetic study. The development of the multifactor dimensionality reduction (MDR) method and its extension, the generalized multifactor dimensionality reduction (GMDR) method, provides a useful solution for detection of multifactor interactions. However, as in population-based association studies, spurious results (false positives or false negatives) may arise for those methods in the presence of population structure, as they assume that the unrelated samples come from a homogeneous population. By integrating the theory of principal components analysis into the GMDR framework, we propose a new method to detect gene-gene interactions controlling false discover rate due to population admixture or stratification. Through comprehensive simulations, we demonstrate the desirable properties of the proposed method in correcting for population structure. In application to the Study of Addiction: Genetics and Environment (SAGE) sample, we detected two significant (after Bonferroni correction) tetragenic interactions among CHRNA4, CHRNB2, BDNF, and NTRK2 associated with Fagerstorm Test for Nicotine Dependence, suggesting the biological role of these genes in nicotine dependence development. (This study is being supported by NIH grant DA025095.)

117

Network-guided Sparse Regression Modeling for Detection of Gene by Gene Interactions

Chen Lu (1) Jeanne Latourelle (2) George T. O'Connor (3) Josee Dupuis (4) Eric D. Kolaczyk (5)
(1) Department of Biostatistics, Boston University School of Public Health (2) Pulmonary Center, Dept of Medicine and Dept of Neurology, Boston University School of Medicine; The NHLBI's Framingham Heart Study (3) Pulmonary Center, Department of Medicine, Boston University School of Medicine; The NHLBI's Framingham Heart Study (4) Dept of Biostatistics, Boston U. School of Public Health; Program in Bioinformatics, Boston U.; The NHLBI's Framingham Heart Study (5) Department of Mathematics and Statistics, Program in Bioinformatics, Boston University

Genetic variants identified by genome-wide association studies to date explain only a small fraction of total heritability. Gene by gene interaction is one important potential source of unexplained total heritability. We propose a novel approach to detect such interactions that utilizes penalized regression and sparse estimation principles, and incorporates outside biological knowledge through a network-based penalty. We tested our new method on simulated and real data. Simulation showed that with reasonable outside biological knowledge, our method performs noticeably better than stage-wise strategies (i.e., selecting main effects first, and interactions second, from among those main effects selected) in finding true interactions, especially when the marginal strength of main effects is weak. We applied our method to Framingham Heart Study data on total plasma Immunoglbulin E (IgE) concentrations and found a number of interactions among different classes of HLA genes that may interact to influence the risk of developing IgE dysregulation and allergy.

118

Integration of Biological Data in the Context of GWAI Studies

Kristel Van Steen (1) Francois Van Lishout (1) Elena Gusareva (1)
(1) University of Liege

From an evolutionary biology perspective, epistasis is a natural phenomenon: phenotypes can only be buffered against the effects of mutations, in the presence of an underlying genetic architecture of networks of genes that are redundant and robust. This is contrasted against the relatively poor detection or replication rate of epistasis discoveries, despite some of these studies being extremely thorough and involving adequate sample sizes. The sharp contrast clearly indicates the network complexity with which multifactorial traits are regulated. It also indicates our potential inability to appropriately apply the available methods. One of the problems is that their abundance and diversity, their differential aims and set-up, complicate comparative studies. Another is the lack of flexibility of software implementations; choosing one approach usually implies the same model assumption throughout the genome... We show that capturing complex phenomena such as epistasis using genome-wide data requires 1) the integration of biological information, 2) appropriate method selection criteria and 3) an ensemble view that exploits complementary benefits of several (equally powerful and Type I error controlling) methods, while reinforcing signals from "performance alike" epistasis detection methods. We illustrate our argumentation via simulated data and a variety of real-life genome-wide data applications.

119

On a Gene-based Test for Gene-Gene Interaction Using Similarity Measures Between Individuals

Indranil Mukhopadhyay (1)
(1) Indian Statistical Institute

Association study of common genetic variants using genome-wide association studies (GWAS) has been able to map some of the genes associated with diseases. However, a major part of heritability still remains unexplained. One

of the possible explanations for the mixed results of GWAS is that the linear modeling framework of GWAS considers only one single nucleotide polymorphism (SNP) at a time thus ignoring any possible interaction, be it gene-gene (G-G) and/or gene-environment (G-E) interactions. Recently much focus has been given to test such interaction. At the same time, instead of looking at a single SNP level, some multilocus association methods have been developed that combine information from multiple SNPs at a time, which also reduces multiple comparison burden. In this scenario we are working towards developing statistical tests for gene-gene interactions combining information from multiple loci at a time. Our method provides a more holistic approach to explore gene-gene interaction paradigm. It has the flexibility to use knowledge from other sources (biological pathways and protein-protein interactions, biological databases, epigenetic factors etc) and helps to prioritize which genetic variations should be analyzed for G-G interactions to interpret genetic association studies in a biologically meaningful manner to understand the complexity of genotype-phenotype relationship. Some optimum properties of the methods are shown using simulated data.

120

A Maximum Likelihood Approach to Prioritize SNPs for Interactions Using Variance per Genotype

Wei Q Deng (1) Senay Asma (1) Angelo J Canty (1) Guillaume Pare (1)
(1) McMaster University

Genetic interactions have been suggested as a source of "missing heritability" in complex trait genetics yet are difficult to identify because of low statistical power. Although an exhaustive search on a whole genome is feasible, most interactions fail to be significant after Bonferroni correction given the number of tests involved. We previously developed a method – variance prioritization – based on differences in quantitative trait variance among the three possible genotypes of biallelic genetic variants. We showed this method has increased power over exhaustive search under a variety of scenarios. Nevertheless, variance prioritization uses Levene's test, without taking into account that variance will increase or decrease with the number of minor alleles in the presence of interactions. We propose a maximum likelihood approach to test for variance heterogeneity among genotypes when ordering is expected. A closed-form representation of the likelihood ratio test (LRT) statistics was derived, which greatly improves the computational speed. We used simulations to investigate the accuracy and increase in power of LRT in detecting interactions. We show that variance prioritization using LRT is more powerful than both exhaustive search and variance prioritization using Levene's test. While developed to detect variance inequality pertaining to genetic interactions, LRT could be used to any other situation where ordered difference in variances is expected among groups.

121

A Powerful and Computationally Efficient Approach to Discovering Gene-Gene Interactions in Genome-Wide Association Studies

Alexander H Stram (1) Juan P Lewinger (1) Gary K Chen (1)
(1) University of Southern California

Seeing that genome-wide association studies have thus far explained only a small proportion of narrow-sense heritability, there is considerable interest in identifying genetic variants that contribute additively to disease risk. Zuk et al. has recently suggested that when estimating narrow-sense heritability, the impact of an assumed genetic risk factor is often inflated, as it captures both variants with additive effects, as well as variants acting in simple, biologically-plausible models that are only detectably via explicit testing for gene-gene or gene-environmental interactions. Traditional methods for detecting gene-gene interactions with no marginal effects generally suffer from low statistical power and high computational burden. Lewinger et al. proposes a two stage test for gene-gene interactions which is more powerful than a case-control test, and robust to spurious interactions in a general population, unlike case-only tests. While relatively efficient, the first stage of Lewinger et al's test still requires a computationally burdensome pairwise search. We introduce a software implementation of this first stage test which is optimized to run in a massively-parallel fashion on the sort of graphical processing units (GPUs) typically packaged with a consumer-grade desktop computer. By exploiting the large number of processors contained in such a GPU, we are able to run hundreds of searches simultaneously, hence reducing run-time by over one-hundred fold.

122

Recommendations for Genome-wide Search for Epistatic Loci

Tatsiana Vaitsiakhovich (1) Christine Herold (2) Dmitriy Drichel (2) Andre Lacour (2) Vitalia Schuller (1) Tim Becker (2)
(1) Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany (2) German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

Various strategies to identify causal SNPs in GWAS studies under assumption of possible interactions are evaluated. Series of "realistic" disease models are defined on a 2-SNP-genotype table by specification of allele frequencies, penetrances and a minimal distance between available and causal SNPs. We compare the standard single-marker analysis with multi-marker analysis; investigate relative performance of tests *for* interaction and tests including both marginal and interaction effects, so-called tests *allowing for* interaction. We compare case-only with case-control design and contrast allelic and genotypic models. Tests including marginal effects may become significant because of the marginal effect of just one SNP from an interaction pair. Since our goal is to detect **both** SNPs we embed tests allowing for interaction in a two-step hybrid strategy: genome-wide interaction analysis with a combined case-only-interaction/marginal-effects test; follow-up analysis of the significant pairs with a test for interaction, excluding the strongest marginal effect but allowing for marginal effect of the second SNP. For 5% of settings the most efficient strategy is single-marker analysis, typically when allele frequencies are high or causal variant tagging is poor. For another 5% of models the most suitable strategy is testing for interaction without marginal effects provided that a case-only test is used. In the remaining majority of scenarios a hybrid strategy is the most powerful.

123

SVM-based Generalized Multifactor Dimensionality Reduction Approaches for Detecting Gene-Gene Interactions in Family Studies

Yen-Feng Chiu (1) Yao-Hwei Fang (1)
(1) Institute of Population Health Sciences, National Health Research Institutes

Gene-gene interaction plays an important role in the etiology of complex diseases, which may exist without a genetic main effect. It would be helpful to develop methods that can detect not only the gene's main effects but also gene-gene interaction effects regardless of the existence of gene's main effects while adjusting for confounding factors. In addition, when a disease variant is rare or when the sample size is quite limited, the statistical asymptotic properties are not applicable; therefore, approaches based on a reasonable and applicable computational framework would be practical and frequently applied. In this study, we have developed an extended support vector machine (SVM) method and an SVM-based pedigree-based generalized multifactor dimensionality reduction (PGMDR) method to study interactions in the presence or absence of main effects of genes with an adjustment for covariates using limited samples of families. A new test statistic is proposed for classifying affecteds and unaffecteds in the SVM-based PGMDR approach to improve performance in detecting gene-gene interactions. The proposed and original approaches have been applied to the simulation study and a real data example for illustration and comparison. Both the simulation and real data studies show that the proposed SVM and SVM-based PGMDR methods have great prediction accuracies, consistencies, and power in detecting gene-gene interactions.

124

Test for Two-SNP Interaction Adjusting for Long Range Linkage Disequilibrium

Qing Li (1) Silke Szymczak (1) Joan E Bailey-Wilson (1)
(1) Inherited Disease Research Branch, National Human Genome Research Institute

For case-parent trios, we propose a new test for two-SNP interaction that adjusts for linkage disequilibrium (LD). Although several methods test interaction in family based studies, they depend either on haplotype reconstruction and likelihood ratio tests, or the assumption that the markers are in linkage equilibrium. However, in dense marker panels, we often observe two or more markers are correlated although they belong to different haplotype blocks, or are even further apart physically. Without adjustment for long range LD, tests for interaction can have inflated Type-I error rates. Our test utilizes the conditional logistic regression framework for the genotypic TDT test. Instead of using observed case genotypes and untransmitted genotypes generated for pseudo controls to fit the model, we use the probabilities that two alleles at a pair of loci were transmitted together to the case and the probability they were not simultaneously transmitted to weight the relative risk of the interaction between two alleles. To adjust for LD between the loci, we can choose to either rely on haplotype reconstruction or the conditional probability of one allele given the other based on data from founders. When relying on haplotype reconstruction, the test can be extended to multiple-marker interaction. We will present simulation results evaluating the Type-I error and power of the test compared to alternative tests, including PTDT, Cordell's test and case-only test implemented in PLINK.

125

Method to Analyze Multiple Mediators in Case-control Studies with Application to Detecting Mediating Effects of Smoking and Chronic Obstructive Pulmonary Disease on the Association Between CHRNA5-A3 Genetic Locus and Lung Cancer Risk

Jian Wang (1) Sanjay Shete (1)
(1) UT MD Anderson Cancer Center

Recently, there has been of interest using mediation analysis to dissect the direct and indirect effects of genetic variants on complex diseases using case-control studies. However, the bias could arise in the estimations of the genetic variant-mediator association due to the case-control study design. In this case, the mediation analysis might lead to biased estimates for the coefficients and indirect effects. We investigated a multiple mediation model involving a three-path mediating effect through two mediators using case-control study data. We proposed an approach to correct the biased coefficients and provided the accurate estimates for the indirect effects and percent mediated. We conducted simulation studies to investigate the performance of the proposed approach. Our approach can also be used when the original case-control study is frequency-matched on one of the mediators. We applied this approach to the multiple mediation study of simultaneous mediating effects of smoking and COPD on the association between CHRNA5-A3 genetic locus and lung cancer risk using lung cancer case-control study. The results of this analysis showed that the genetic variant influences the lung cancer risk indirectly through all three different pathways. The percent mediated was 18.3% through smoking alone, 30.2% through COPD alone and 20.6% through the path including both smoking and COPD, and the total genetic variant-lung cancer association explained by the two mediators was 69.1%.

126

Finding GWAS Signals in the Lower Manhattan by Testing GxE Interactions

Jim Gauderman (1) Juan Lewinger (1) Pingye Zhang (1) David Conti (1)
(1) University of Southern California

In a genomewide association study (GWAS), the Manhattan plot is used to display the −log10 (p-values) from tests of the main effect of each SNP on the trait. Less significant SNPs in the 'Lower Manhattan' may be truly associated with the trait but have only a modest overall effect and thus low power to be detected. However, some of these SNPs may be quite important in subgroups of the population defined by an environmental exposure (e.g. air pollution, smoking) or personal factor (e.g. sex, another gene). We address the question "Using a genomewide interaction scan (GWIS), can we find new genes that were not found in the main-effect GWAS scan?" We review commonly used approaches for testing gene-environment (GxE) interactions, and propose a new two-step method that makes more efficient use of available data. Our analytic framework can be applied to either case-control or case-parent trio data. We will show that for many models, our new method provides greater

power to detect GxE interaction than other available methods, including the powerful case-only analysis. For a variant with a weak main effect and a modest sized interaction, our method provides the best chance of a 'yes' answer to the above question.

127

Evidence of Gene-environment Interaction Between Common Breast Cancer Susceptibility Loci and Established Environmental Risk Factors

Stefan Nickels (1) Therese Truong (2) Rebecca Hein (3) Kristen Stevens (4) Katharina Buck (5) Mia M Gaudet (6) Lothar Haeberle (7) Amanda Spurdle (8) Sabine Behrens (1) Ursula Eilber (1) Jonine Figueora (9) Alison M Dunning (10) Georgia Chenevix-Trench (11) Paul Pharoah (10) Douglas F Easton (10) Per Hall (12) Marjanka Schmidt (13) Montserrat Garcia-Closas (14) Roger L Milne (15) Jenny Chang-Claude (1)

(1) German Cancer Research Center (DKFZ), Division of Cancer Epidemiology (2) Inserm (National Institute of Health and Medical Research), CESP (Center for Research in Epidemiology and Population Health) (3) PMV Research Group at the Department of Child and Adolescent Psychiatry and Psychotherapy, University of Cologne (4) Department of Health Sciences Research, Mayo Clinic, Rochester (5) German Cancer Research Center (DKFZ) National Center for Tumor Diseases (6) Epidemiology Research Program, American Cancer Society, Atlanta (7) Department of Obstetrics and Gynecology, Erlangen University Hospital (8) Queensland Institute of Medical Research, Molecular Cancer Epidemiology (9) National Cancer Institute, Division of Cancer Epidemiology and Genetics (10) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge (11) Genetics and Population Health Division, Queensland Institute of Medical Research (12) Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm (13) Netherlands Cancer Institute, Antoni van Leeuwenhoek hospital, Amsterdam (14) Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey (15) Genetic and Molecular Epidemiology Group, Human Cancer Genetics Program, Spanish National Cancer Research Centre [CNIO]

There is limited information on the combined effects on breast cancer of established common genetic susceptibility variants and life-style/environmental risk factors.Using up to 34,793 invasive breast cancers and 41,099 unaffected controls, we examined potential effect modification of single nucleotide polymorphisms at 23 susceptibility loci by 10 established environmental risk factors (with 18 variables) in women of European ancestry. We used a novel empirical-Bayes estimator to test for gene-environment interactions on a log-additive scale with increased efficiency compared to standard methods for breast cancer overall and by estrogen receptor status. Five gene-environment interactions were statistically significant at Bonferroni corrected P-value$\leq$, or $\leq 1.21\times10^{-4}$. We replicated the previously reported interaction between rs3817198 (*LSP1*) and number of full-term births. The per allele odds ratio (95% confidence interval) for rs3817198 was 1.08 (1.01–1.16) in nulliparous women and ranged from 1.03 (0.97–1.10) in parous women with one birth to 1.26 (1.15–1.37) with $\geq$ 4 births ($P_{interaction}=3.6\times10^{-10}$). Additional interactions were found between rs11249433 (1p11.2) and parity ($P_{int}=6.1\times10^{-9}$), rs999737 (*RAD51L1*) and duration of current estrogen-only therapy use, rs13387042 (2q35) and current estrogen-progestagen therapy use, and rs1292011 (12q24) and lifetime alcohol intake. The impact of such results on risk prediction models for breast cancer needs to be considered in future studies.

128

Application of the Empirical Hierarchical Bayes Approach for Gene-Environment Interaction to TRICL Lung Cancer GWAS

Melanie Sohns (1) Elena Viktorova* (presenting author) (1) Gord Fehringer (2) Valerie Gaborieau (3) Younghun Han (4) Jenny Chang-Claude (5) Joachim Heinrich (6) Angela Risch (7) Christopher I. Amos (4) Paul Brennan (3) Rayjean J. Hung (2) Heike Bickeboeller (1)

(1) Department of Genetic Epidemiology, University Medical Center, Georg-August University of Goettingen, Goettingen, Germany (2) Prosserman Centre for Health Research, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada (3) International Agency for Research on Cancer (IARC), Lyon, France (4) Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA (5) Unit of Genetic Epidemiology, Division of Cancer Epidemiology, Deutsches Krebsforschungszentrum, Heidelberg, Germany (6) Institute of Epidemiology I, Helmholtz Center Munich, Neuherberg, Germany (7) Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center, Heidelberg, Germany

Lung cancer is the most common cause of the cancer mortality and smoking its major risk factor. Detecting susceptibility genes and their interactions with smoking and other environmental factors are expected to further elucidate disease mechanism. Notably, in such studies underlying G-E association may greatly influence the results, producing false positive findings. Therefore, we propose a novel empirical hierarchical Bayes approach (EHB-GE) for GWAS to identify GxE interactions in presence of population based G-E association. We adapt the hierarchical Bayes prioritization of Lewinger borrowing G-E information across all SNPs, to obtain reduced variance estimates of G-E associations in controls. In simulations EHB-GE outperforms other GxE methods, being promising test to detect GxE interactions. For real data applications, we investigated SNP-smoking interactions in four lung cancer GWAS for the TRICL/ILCCO consortia. We compared EHB-GE with case-control, case-only, Albert's and Murkray's two step and Mukherjee empirical Bayes approaches. EHB-GE reaches high power as the case-only test, while adjusting for G-E associations. A strong SNP-smoking association was not observed. Thus EHB-GE and case-only agree in SNPs ranking. Overall, the EHB-GE demonstrated higher ranking power under various simulated scenarios, accounting for G-E association. Therefore, it should be preferable method for SNP selection for further investigations particularly when G-E associations are suspected.

129

Empirical-Bayes Approach to Investigate Gene-environment Interactions in Large Consortia

Katharina Buck (1) Rebecca Hein (2) Lothar Haeberle (3) Montserrat Garcia-Closas (4) Nilanjan Chatterjee (5) Jenny Chang-Claude (1)
(1) German Cancer Research Center (DKFZ), Heidelberg (2) University of Cologne (3) University Hospital Erlangen (4) Institute of Cancer Research and Breakthrough Breast Cancer Research Centre, London (5) National Cancer Institute, NIH

Pooling data from many studies in large consortia has become a common practice for the analysis of gene-environment interactions (GxE). This involves combining studies of varying designs and quality, which can be problematic particularly for estimation of E main effects. Using simulations we compared the degree of bias in E and GxE effects when using logistic regression and empirical-Bayes (EB) analysis to analyze pooled case-control data from studies with different degrees of bias in E and no bias in G effects. As shown previously, logistic regression was less powerful than the EB method to detect GxE effects, and was robust to departures from the G-E independence assumption. Under G-E independence in the controls, both methods resulted in unbiased GxE effects, however the E main effects were biased by design. To address this problem, we introduced an interaction term for study type (studies with unbiased vs. biased E) and E main effect to the EB model. This allowed us to obtain unbiased E main effects from the subset of E-unbiased studies while using data from all studies to estimate G and GxE effects and maintaining the gain in power for GxE effects of the EB-method. In summary, we showed that, under G-E independence, large pooled analyses including a mixture of E-biased and E-unbiased studies can obtain unbiased GxE estimates using data from all studies, and E effects from the subset of E-unbiased studies. Also, the EB-method is robust to departures from G-E assumption.

130

The Impact of Exposure Misclassification and Exposure-biased Sampling on Power for Detecting Gene-by-Environment Interactions in Case-control Studies

Stephanie Stenzel (1,2) Jaeil Ahn (3) Bhramar Mukherjee (4)
(1) Department of Epidemiology, School of Public Health, University of Michigan (2) Department of Statistics, College of Literature, Science, and the Arts, University of Michigan (3) Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center (4) Department of Biostatistics, School of Public Health, Univesity of Michgan

Background-With limited funding and sample availability, choosing an optimal sampling design to maximize power for detecting gene-by-environment (G*E) effects is critical. Exposure enriched sampling is often used to select subjects for genotyping to enhance power. However, exposure misclassification (MC) combined with biased sampling can affect the characteristics of relevant statistical tests.
Methods-We assessed the power, Type I error, and bias properties of case-only, case-control, and empirical Bayes methods for testing G*E interaction and a joint marginal genetic (or environmental) effect and G*E interaction under different scenarios of biased sampling and exposure MC. Properties of these methods were examined across a range of exposure sensitivity and specificity values and various biased designs. Properties were assessed via an extensive simulation study with several case-control sample sizes. We also considered the role of gene-environment (G-E) independence.
Results-Exposure-enriched sampling schemes enhance power as compared to random selection of subjects but yield bias in estimation. Exposure MC leads to severe depression of power for these tests. G-E independence affects the relative properties of the 3 methods, with the case-only approach suffering most severely.
Conclusion-Those conducting G*E studies should be aware of exposure measurement MC properties and the prevalence of exposure when choosing methods for G*E interaction detection and sampling scheme.

131

Application of Genome-wide Gene-environment Interaction Methods: The SEED Autism Study

Christine Ladd-Acosta (1) Brian K Lee (2) Joseph Bonner (3) Brooke Sheppard (1) Nicole Gidaya (2) Lauren Weiss (4) Jeffrey Quinn (4) Gayle Windham (5) Ann Reynolds (6) Lisa Croen (5) Diana Schendel (7) Craig Newschaffer (2) Daniele Fallin (1)
(1) Johns Hopkins Bloomberg School of Public Health (2) Drexel University School of Public Health (3) Michigan State University Biomedical Research Informatics Core (4) University of California, San Francisco (5) Kaiser Permanente Northern California Division of Research (6) University of Colorado Denver School of Medicine (7) Centers for Disease Control and Prevention

There is increasing interest in gene-environment-wide interaction studies (GEWIS) for complex disease, particularly for autism as few definitive genetic variants have been identified and recent evidence supports a stronger role for environmental factors in disease risk than previously thought. We performed the first GEWIS in autism using several novel GEWIS methods. Using Illumina Omni and Affymetrix Axiom arrays we measured over 1 million SNPs for 614 cases and 742 controls enrolled in the Study to Explore Early Development (SEED). We examined prenatal exposures across four domains including maternal use of tobacco, alcohol, and medication (B2ARs and SSRIs) as well as maternal infection. After applying data quality control measures, and performing imputation to obtain $>$ 6 million genotypes per person, initial analysis was performed using a new joint likelihood ratio test for marginal genetic main effects and gene-environment interaction among 873 SEED children. We detected genome-wide significant ($P < 5\times10^{-7}$) interaction between genotype and smoking for several neighboring SNPs on chromosome 2. We will present empirical results comparing multiple GEWIS methods based on analyses of 1,356 children across the four exposure domains. Our initial analysis suggests coupling genetic and environmental exposure information to determine autism risk is more fruitful than looking for genetic marginal effects alone.

132

Prioritizing SNPs for Gene-Environment and Gene-Gene Interactions: A method to Meta-analyze Levene's Test of Homogeneity of Variance

Guillaume Pare (1) Wei Q Deng (1) Senay Asma (1)
(1) McMaster University

Meta-analysis is frequently employed as a mean to increase sample size for genome-wide association and interaction searches when individual studies are otherwise underpowered. We have previously developed a prioritization scheme to select potentially interacting SNPs for gene-environment and gene-gene interaction testing. This method, Variance Prioritization, leverages the observation that quantitative trait variance conditional on genotype will vary when an interaction is present. Variance prioritization first tests for heterogeneity of variance between genotype groups using Levene's test. SNPs with nominally significant Levene's p-value are then tested for interaction. We now propose to extend Variance Prioritization to meta-analysis of multiple studies and herein develop a framework to meta-analyze heterogeneity of variance using Levene's test across studies without exchanging individual-level data. Meta-analysis presents a promising strategy in the discovery of novel genetic interactions, and implementation of Variance Prioritization for meta-analysis of genetic interactions offers the opportunity to combine the considerable sample size of genome-wide meta-analysis with the increase in power gained through Variance Prioritization.

133

Heavy Metals, Organic Solvents and Multiple Sclerosis: An Exploratory Analysis of Gene-environment Interactions

Melanie D. Napier, M.S.P.H. (1) Charles Poole, Sc.D. (1) Glen A. Satten, Ph.D. (2) Allison Ashley-Koch, Ph.D. (3) Ruth Ann Marrie, M.D., Ph.D. (4) Dhelia M. Williamson, Ph.D. (2)
(1) Department of Epidemiology, University of North Carolina Gillings School of Global Public Health, Chapel Hill, NC, USA (2) Division of Reproductive Health, Centers for Disease Control and Prevention, Atlanta, GA, USA (3) Center for Human Genetics, Duke University Medical Center, Durham, NC, USA (4) Department of Internal Medicine, University of Manitoba, Winnipeg, Manitoba, Canada

Multiple sclerosis (MS) is a multifactorial disease with environmental and genetic influences. Exposures to heavy metals and organic solvents are potential etiologic factors, but their interaction with genes associated with MS has not been studied. We examined the role of exposure to lead, mercury, zinc, and solvents and 60 single nucleotide polymorphisms (SNPs) in eight genes associated with MS: human leukocyte antigen (HLA) DRB1*1501; interleukin 2 receptor alpha (IL2RA); interleukin-7 receptor alpha chain (IL7RA); T-cell receptor alpha (PTCRA); vitamin D receptor (VDR); myelin basic protein (MBP); tumor necrosis factor alpha and beta (TNFα/β); apolipoprotein E (APOE). Data from a population-based case-control study of 224 prevalent MS cases and 501 age-, race-, gender- and location-matched controls were used to construct conditional logistic regression models, additionally controlling for education and population stratification. MS cases were more likely to report exposure to lead, mercury, and zinc than controls, and less likely to report solvent exposure. Statistically significant environmental interactions ($p<0.05$) were identified with SNPs in TNFα, APOE, VDR, MBP, and PTCRA. We believe this is the first epidemiologic study to describe interactions between MS genes and exposure to lead, mercury, and zinc. Results from this study are a crucial starting point for exploring the validity of immunological hypotheses of MS pathogenesis for metal and solvent exposures.

134

Effects of Waterpipe Smoking on DNA Modification and Gene Expression

Zahra Montazeri (1) Hoda El katerji (1) James Gomes (1) Julian Little (1)
(1) University of Ottawa

Recently, a sharp rise in waterpipe smoking has been observed in North, especially among young adults. In 2006, 4% of Canadians aged 15 years and older reported ever trying waterpipe, and 1% reported use in the previous month. There is a gap in the evidence as to the potential health effects of waterpipe smoking. We are investigating the effects of waterpipe smoking on gene expression and methylation patterns among young waterpipe smokers in Ottawa, Canada. Our findings could be providing evidence for epigenetic modification and hence predict adverse health effects from waterpipe smoking.
The main purpose of this study is to investigate whether smoking waterpipe causes alterations in gene expression through DNA methylation known to be associated with smoking-related diseases, such as cancer and cardiovascular diseases. We identified almost 30 smoking-related diseases, and validated SNPs associated with these on the basis of genome-wide association studies, thereby identifying about 260 SNPs.
We recruit ever and never users of waterpipes from community engagement in young people. These participants are asked to respond to a questionnaire and to provide a saliva sample. DNA extracted from these samples is being analyzed using RT-PCR.
The data from assays of samples from waterpipe smokers and non-smokers will be analysed by using advanced bioinformatic and statistical techniques for high throughput data.

135

Gene-by-Smoking Interaction Analyses of Kidney Traits in the NHLBI Candidate-gene Association Resources CARe Cohorts

Nora Franceschini (1) Albert Dreisbach (2) W Linda Kao (3) Kari E North (4) Michael Nalls (5) Caroline Fox (6) Adrienne Cupples (7)
(1) Epidemiology, University of North Carolina at Chapel Hill (2) Nephrology, University of Mississippi, Jackson, MS (3) Epidemiology, John Hopkins University, Baltimore, MD (4) Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC (5) Laboratory of Neurogenetics, National Institute of Aging, National Institutes of Health, Bethesda, MD (6) NHLBI's Framingham Heart Study and the Center for Population Studies and Endocrinology, Harvard Medical School, Boston, MA (7) Biostatistics, Boston University School of Public Health, Boston, MA and Framingham Heart Study, Framingham, MA

Smoking is a risk factor for chronic kidney disease (CKD) but little is known about its interaction with genetic factors. We examined the evidence for interaction of current smoking on the association of SNPs with kidney traits (estimated glomerular filtration rate, eGFR, and albuminuria, UACR) in 23,767 white CARe participants from five studies, genotyped using the custom IBC array. We obtained study-specific residuals of natural log-transformed eGFR or UACR regressed on age, sex and study site. We then

performed smoking strata-specific genome wide association analyses (GWAS) using additive models adjusted for population stratification and family structure, if needed. Meta-analyses was performed across smoking-strata using inverse variance weighted fixed effect methods. We assessed smoking interaction using a heterogeneity test ($P<0.10$) and the I^2 metric. Among SNPs reaching the array wide significance threshold (2.0E-6) for marginal association, there was significant evidence for smoking-strata heterogeneity for rs7422339 (*CPS1*, $P_{het}=0.04$, $I^2=77.7\%$) with eGFR, with the A allele having larger decreases in eGFR among smokers (beta=-0.020) than in non-smokers (beta=–0.009). For UACR, the rs1801239 (*CUBN* missense variant, $P_{het}=0.07$, $I^2=64.8\%$) T allele showed less protective effect among smokers (beta=–0.030) than non-smokers (–0.126). Our findings suggest important modification of known gene effects on kidney traits by smoking exposure.

136

Associations and Interactions of Genetic Polymorphisms in Innate Immunity Genes with Early Viral Infections and Susceptibility to Asthma and Asthma-related Phenotypes

Denise Daley (1) Julie E Park (1) Jian-Quing He (1) Jin Yan (1) Loubna Akhabir (1) Dorota Stefanowicz (1) Allan B Becker (2) Moira Chan-Yeung (1) Yohan Bosse (3) Anita L Kozyrskyj (4) Alan L James (5) Arthur W Musk (5) Catherine Laprise (6) Richard G Hegele (7) Peter D Pare (1) Andrew J Sandford (1)
(1) University of British Columbia (2) University of Manitoba (3) Laval University (4) University of Alberta (5) Sir Charles Gairdner Hospital (6) Universite du Quebec a Chicoutimi (7) University of Toronto

The innate immune system is essential for host survival because it recognizes invading pathogens and mounts defensive responses. We used four large samples (n=5565) to investigate associations of 321 single nucleotide polymorphisms (SNPs) in 26 innate immunity genes with four phenotypes: atopy, asthma, atopic asthma and airway hyper-responsiveness (AHR) in three Canadian family-based studies and one Australian population-based case control study. Interactions between innate immunity genes and early viral exposure to three common viruses (parainfluenza, respiratory syncytial virus and picornavirus) were examined using data collected from the Canadian Asthma Primary Prevention Study. Analyses were conducted using both an affected-only family-based transmission disequilibrium test and case-control methods.
Results: Six SNPs (rs1519309 (*TLR3*), rs740044 (*ILIR2*), rs4543123 (*TLR1*), rs5741812 (*LBP*), rs917998 (*IL18RAP*), and rs3136641 (*NFKBIB*)) are significant ($P<0.05$, confirmed with 30,000 permutations) in both the combined analysis of main genetic effects and SNP*viral interaction analyses in both case-control and family based methods. The TLR1 variant (rs4543123) is associated with both multiple viruses (RSV and PIV) and multiple phenotypes.
Conclusion: We identified susceptibility genes for asthma and related traits and interactions between these genes and early life viral infections that may prove to be of predictive value for the development of allergic diseases.

137

A Comparison of Approaches for Genome-wide Gene-environment Interaction Analyses in the Risk of Estrogen Receptor (ER)-Negative Breast Cancer

Amit D Joshi (1) Sara Lindstrom (1) Regina G Ziegler (2) Jonine D Figueroa (2) Susan M Gapstur (3) Christopher A Haiman (4) Elio Riboli (5) Peter Kraft (1) Mia M Gaudet (3)
(1) Harvard School of Public Health (2) National Cancer Institute (3) American Cancer Society (4) Keck School of Medicine, University of Southern California (5) School of Public Health, Imperial College London

Established environmental risk factors can be used to identify novel susceptibility loci using genome-wide interaction scans. Although several approaches have been proposed for this purpose, few studies have used empirical data to compare their relative efficiency for binary phenotypes such as cancer. We performed genome-wide interaction analyses of ER-negative breast cancer risk in 1,998 cases and 3,263 controls, using data from six prospective cohorts within the NCI Breast and Prostate Cancer Cohort Consortium and a case control study. Environmental risk factors for breast cancer included were age at menarche, age at menopause, ever full-term pregnancy, ever use of oral contraceptives, ever use of post-menopausal hormones and body mass index. A meta-analysis approach is being used to combine study-specific test statistics and standard errors. We will present results for the top loci interacting with each environmental risk factor. Our study will discuss the relative strengths and limitations of the following approaches: the standard one degree of freedom (df) interaction test; the two df test for joint analysis of gene main effect and gene-environment interaction effect (Hum Hered 2007;63:111–9); the case-only analyses; and the semi-parametric maximum likelihood estimation approach (Biometrika 2005;92:399–418). Identifying optimal methods to examine high density genetic variation in the context of environmental exposures might yield novel loci associated with cancer.

138

External Sources of Vitamin D Modify the Effects of the GC and CYP2R1 Genes on 25-hydroxyvitamin D Concentrations: CAREDS

Corinne D Engelman (1) Kristin J Meyers (1) Sudha K Iyengar (2) Zhe Liu (1) Chitra Karki (1) Robert P Igo, Jr (2) Barbara Truitt (2) Jennifer Robinson (3) Gloria E Sarto (1) Robert Wallace (3) Lesley Tinker (4) Erin LeBlanc (5) Yiqing Song (6) Julie A Mares (7) Amy E Millen (8)
(1) University of Wisconsin-Madison (2) Case Western Reserve University (3) University of Iowa (4) Fred Hutchinson Cancer Research Center (5) Kaiser Permanente Center for Health Research (6) Brigham and Women's Hospital and Harvard Medical School (7) University at University of Wisconsin-Madison (8) University at Buffalo

Vitamin D deficiency (defined by the blood concentration of 25-hydroxyvitamin D [25(OH)D]) has been associated with many adverse health outcomes. Genetic and non-genetic factors account for variation in [25(OH)D], but the role of interactions between these factors is unknown. To assess this, we examined 1,204 women of European descent from the Carotenoids in Age-Related Eye Disease Study (CAREDS), an ancillary study of the Women's Health Initiative Observational Study. Twenty-nine SNPs in 4 genes, *GC*, *CYP2R1*, *DHCR7* and *CYP24A1*, from recent meta-analyses of genome-wide association studies of [25(OH)D], were genotyped. Associations between these SNPs and [25(OH)D] were tested using linear regression under an

additive genetic model adjusted for age, blood draw month and ancestry. Results were stratified by season of blood draw and, separately, vitamin D intake for 6 SNPs showing a significant association with [25(OH)D] ($p<0.01$). Two non-synonymous SNPs in *GC* and 4 SNPs in *CYP2R1* were strongly associated with [25(OH)D] in individuals whose blood was drawn in summer ($p\leq0.0001$) but not winter months and, independently, in individuals with vitamin D intake $\geq$400 ($p<0.0001$) but not <400 IU/day. This effect modification has important implications for the design of discovery and replication genetic studies for all health outcomes and for public health recommendations and clinical practice guidelines regarding achievement of adequate vitamin D status.

139

A Data-smoothing Approach to Graphical Displays and Testing of Gene-environment Interaction Using Data from Case-parent Trios

Ji-Hyung Shin (1) Claire Infante-Rivard (2) Brad McNeney (1) Jinko Graham (1)
(1) Simon Fraser University (2) McGill University

Complex diseases are thought to result from an interplay between genes (*G*) and environmental or non-genetic attributes (*E*). Several approaches have been proposed to assess gene-environment interactions (*GxE*) using data from case-parent trios. Under *GxE*, these approaches either explicitly specify or are designed to work well under a particular functional form of interaction, such as linearity. When this functional form is mis-specified, the approaches can lead to false-negative results. We present a data-smoothing approach with the flexibility to model non-linear *GxE*. Our approach offers the advantage of allowing the data to suggest the functional form of *GxE* rather than specifying it in advance. The resulting point and interval estimates may be displayed graphically to explore the form of *GxE*. For testing *GxE*, we adopt a permutation-based approach that accounts for the additional uncertainty introduced by the smoothing process. We investigate the statistical properties of the proposed permutation test through simulation. The simulation results demonstrate that the proposed test can detect non-linear (e.g., quadratic) *GxE* better than other available tests. If the functional form and/or the inheritance mode of *GxE* are not understood well, our proposed method has the flexibility to provide insight into both.

140

Joint Effect of Genetic and Lifestyle Risk Factors on Type 2 Diabetes Risk Among Chinese Men and Women

raquel villegas (1) Ryan Delahanty (1) Yu-Tang Gao (2) Jirong Long (1) Scott Williams (1) Yong-Bing Xiang (2) Hui Cai (1) Hong-Lang Li (2) Frank Hu (3) Qiuyin Cai (1) Wei Zheng (1) Xiao-Ou Shu (1)
(1) vanderbilt university (2) Shanghai Cancer Institute, Shanghai, China (3) Harvard University

We evaluated 36 type 2 diabetes (T2D) genome-wide association study identified SNPs in 2679 T2D cases and 3322 T2D controls middle-age Han Chinese from urban Shanghai. In analysis adjusted for age, body mass index (BMI) and sex, 14 SNPs were significantly associated with T2D. We calculated two genetic risk scores, GRS1 with all the 36 SNPs and GRS2 with the 14 SNPs associated with T2D and a lifestyle risk score by adding T2D high risk factors (no exercise participation, high BMI and high waist-to-hip ratio, WHR). The OR for T2D with each GRS1 point were 1.08 (95% CI:1.06–1.09) and 1.16 (95% CI:1.13–1.18) for GRS2. The OR for quintiles of GRS1 were 1.00, 1.26, 1.68, 1.95 and 2.18 ($P<0.0001$) and 1.00, 1.13, 1.68, 2.02 and 2.67 ($P<0.001$) for GRS2. Participants in the highest GRS1 tercile and higher BMI category had a higher risk of T2D (OR=11.08;95% CI:7.39–16.62) compared to those in the lower GRS1 tertile and low BMI. We found similar joint associations between GRS1 and WHR or exercise participation categories. Compared to participants with low GRS1 and no lifestyle risk factor, those with high GRS1 and 3 lifestyle factors had higher risk of T2D (OR=13.06;95%CI: 8.65–19.72). The association was accentuated in analysis with the GRS2. No significant interactions between the GRS and the lifestyle factors were observed. In conclusion an association between GRSs and lifestyle factors, alone and in combination contributed T2D risk among middle-age Chinese men and women.

141

HLA-DRB1*15 and Smoking as Risk Factors for Multiple Sclerosis in Serbia, Evidence of Interaction

Izaura Lima Bomfim (1) Dragana Obradovic (2) Gordana Toncev (3) Zorica Knezevic (3) Gordana Supic (2) Nemanja Borovcanin (2) Ingrid Kockum (1) Tomas Olsson (1)
(1) Karolinska Institutet, Stockholm, Sweden (2) Military Medical Academy, Belgrade, Serbia (3) University Hospital Kragujevac, Kragujevac, Serbia

Genetic as well as environmental factors contribute to onset of multiple sclerosis (MS). HLA-DRB1*15 and smoking have consistently been found to be associated with MS. Furthermore, HLA allele groups (HLA-DRB1*15 and absence of HLA-A*02) and smoking has recently been suggested to interact to cause disease in a Scandinavian population. We set out to investigate the potential effects of HLA-DRB1 and HLA-A, smoking as well as synergistic effects of these variables in a Serbian dataset. We tested for association between HLA-DRB1-, HLA-A alleles as well as smoking with MS risk in a Serbian dataset consisting of 580 MS patients and 620 controls. Patients were recruited from two clinics, one in Belgrade and one in Kragujevac. Controls were a combination of blood donors, hospital staff and volunteers. We found an association between HLA-DRB1*15 and risk of MS, OR=2.7 (95% CI: 2.0–3.7; $P=1\times10^{-10}$). No associations were found between HLA-A allele groups and susceptibility to MS. Smoking was found to increase risk of MS, OR=2.1 (95% CI: 1.6–2.7; $P=6\times10^{-7}$). We found a significant interaction between HLA-DRB*15 and smoking, attributable proportion=0.45 (95% CI: 0.2–0.7; P-value=8×10^{-4}). We conclude that HLA-DRB1*15 and smoking are risk factors for MS in Serbia, as in other populations. In line with previous observations in a Scandinavian dataset we found that presence of both HLA-DR15 and smoking confers a risk of MS which is over and above their individual effects.

142

Gene-environment Association Testing in Pedigrees and Mixed Study Designs

Karen Curtin (1) Jathine Wong (1) Nicola J Camp (1)
(1) Department of Medicine, University of Utah School of Medicine, Salt Lake City, Utah, USA

Logistic regression is used to investigate genotype-phenotype associations. To enhance the capabilities of our Genie software, we developed a logistic regression module (Logit) to perform association testing with capability to model gene-environment (G-E) interactions in pedigrees of unrestricted size/complexity and mixtures of family and independent subjects. The Logit module calculates initial parameter estimates using least squares. Maximum likelihood estimation is carried out with the iterative Newton-Raphson method. For genetic parameters or G-E interactions, significance is empirically assessed using Monte Carlo testing and a Mendelian gene-drop procedure to create an empirical null distribution. Interaction terms for any combination of genetic and environmental, discrete or continuous variables can be specified. Using simulated data assuming independent observations, we tested the accuracy of Logit point-estimates and confidence intervals compared to those from SAS(R) statistical software. Coefficients for main effects and two-variable interactions were identical. Confidence limits and *P*-values were consistent with that expected from an empirical estimate. To provide flexibility across an array of study designs, Genie Logit allows testing of candidate loci and interaction with environmental factors in mixtures of pedigree and unrelated subjects.

143

A Sparse Partial Least Squares Multivariate Model to Predict Rheumatoid Arthritis Erosive Joint Damage by Selecting Key Variables from a Large Panel of SNPs and Environmental Factors

Lyndsey H Taylor (1) Anthony G Wilson (1) M Dawn Teare (1)
(1) University of Sheffield

Candidate gene association and linkage studies have identified a role for genetics in predicting Rheumatoid Arthritis (RA) severity. Whilst many genetic variants have been individually identified, few have attempted to model many genes and environmental variables together. In this work, sparse partial least squares (SPLS) is used to create a linear model able to predict erosive joint damage using single nucleotide polymorphisms (SNPs) and environmental factors. The Larsen score (LS) is a quantitative validated measure of erosive joint damage. Initial modelling to predict the LS for 912 subjects with RA investigated 387 variables. These consisted of 19 environmental factors and 368 SNPs; chosen due to previous links with RA or other related autoimmune diseases. 10-fold cross-validation (repeated 50 times) was used, extracting the 95 most predictive variables for each fold. The variables which were selected in 8 out of the 10 folds, in all 50 runs, were retained for the final model. The correlation between actual and predicted LS, calculated from the final model containing 58 variables, was r=0.60. Binary classification (<10 versus >=10) resulted in 76.1% of subjects being correctly classified. SPLS is a promising approach able to identify key variables contributing to the amount of erosive joint damage in patients with RA. Work is currently continuing applying the methodology to a genome wide association study with 336076 SNPs and 396 subjects.

144

Comparing Power for Gene-based and Low-frequency SNP Tests in Quantitative Traits (a GoT2D Study)

Alisa K Manning (1) Xueling Sim (2) Adam E Locke (2) Vineeta Agarwala (3) Loukas Moutsianas (4) Robert Weyant (2) Mark I McCarthy (5) David Altshuler (6) Michael Boehnke (2)
(1) Broad Institute of Harvard and MIT, Cambridge, MA, USA; Massachusetts General Hospital, Boston, MA, USA (2) Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA (3) Broad Institute of MIT and Harvard, Cambridge, MA, USA; Harvard Medical School, Boston, MA, USA (4) Department of Statistics, University of Oxford, Oxford, UK (5) Churchill Hospital, Oxford, UK.; Wellcome Trust Centre for Human Genetic Research, Oxford, UK (6) Broad Institute, Cambridge, MA, USA; Harvard Medical School, Boston, MA, USA.; Massachusetts General Hospital, Boston, MA USA

There have been few published comparisons of gene-based association tests with quantitative traits (QTs). Among these tests are burden tests that pool low frequency alleles within a gene and variance component tests such as SKAT. Hypothesis-driven tests can also be used to analyze SNPs with functional annotation (i.e. loss of function variants.) Here, we provide comparisons of these gene-based tests as well as single variant tests such as the Wald, likelihood ratio and score tests under a range of underlying genetic models. Models being considered are those with 1–5 SNPs per gene explaining 0.25–5% of the variance of the simulated trait. Questions are: (1) what are the power and type 1 error rates for the burden tests under different functional hypotheses and, in order to determine the frequency range of SNPs entering into the gene-based tests, (2) what underlying rare variant causal alleles (under which frequency/effect size combinations) can be detected using single variant tests? Our simulations used an interim data freeze from the Genetics of Type 2 Diabetes (GoT2D) project of 918 individuals. Genotypes from low pass sequencing were used as input for Hapgen2, a genetic simulation program. Phenotypes were simulated under a variety of genetic models and power and type 1 error were estimated for each of the tests. These simulations help elucidate the type of genetic effects we have the power to detect for QTs in GoT2D.

145

A Powerful Gene Level Association Test for Genome-wide Association Studies

Jingyuan Zhao (1) Garrett Teoh Hor Keong (1) Anbupalam Thalamuthu (1)
(1) Genome Institute of Singapore

Genome-wide association studies (GWASs) aim to identify genetic variants associated with the disease risk. The statistically significant single nucleotide polymorphisms (SNPs) detected by single SNP analysis capture only a small to modest proportion of disease associated variants. Genes, as functional units of genetics, may carry more information than SNPs, so gene level association tests are efficient alternatives for the analysis of GWAS's of complex diseases. Here we propose a gene level test that combines the information of the most significant SNP subset to perform an overall test for the association within the gene. As a

computationally intensive permutation procedure is not feasible in GWAS, we derive an asymptotic distribution of the proposed test statistic to evaluate empirical significances. Using simulations data sets, we verify the asymptotic distribution of the proposed test statistic and demonstrate that the proposed method has improved the power over several alternative gene level tests under a wide range of gene sizes, LD structures and effect sizes. To illustrate its application in GWAS, we apply the proposed gene level test to analyze a published GWAS data set.

146

Canonical Correlation of Set Interactions (CASI): A Novel Method

Joshua Millstein (1)
(1) Division of Biostatistics of the Keck School of Medicine of USC

Differences in LD between case and control populations for two risk loci reflect departures from additivity of relative risks, indicating statistical epistasis. This concept applies to interactions between sets of risk factors, whether genetic or environmental. Disease susceptibility could be affected by epistatic interactions between multiple SNPs underlying one gene and multiple SNPs underlying a second gene, or interacting in *cis* with epigenetics. I will describe a novel test to detect interactions between sets of variables that can identify the largest contributors to a significant result. There are 7 steps: 1) canonical coefficients in cases only, relating two feature sets (e.g., SNPs to methylation probes); 2) compute canonical coefficients in controls with loadings from cases; 3) compute case-control differences between Fisher transformed canonical coefficients; 4) maximum absolute difference is the raw statistic; 5) permute case-control labels and repeat steps 1–4; 6) scale statistics by permutation standard deviations; 7) compute FDR. In simulations, I found type I error to be well controlled by the permutation procedure. In addition, the proposed approach was substantially more powerful than standard logistic regression likelihood ratio tests of pair-wise interaction parameters adjusted for multiplicity by FDR. I will also discuss applications to real data. The CASI approach represents a general method for identifying sets of interacting molecular features.

147

Integrating Over Multiple SNP Signals to Improve Power of SNP Set-, Gene-, and Pathway-based Association Tests

Haley J Abel (1) Ingrid B Borecki (1) Michael A Province (1)
(1) Washington University School of Medicine

As attention in genomic scans moves from single SNP to multiple SNP associations, there is a need for robust, meta-analysis compatible methods to combine information across multiple loci, across many cohorts. We have developed a method for SNP set association testing that maximizes use of information by incorporating all association signals from the set while controlling for LD. Our program, QQAUC, uses as a test statistic the area under the set-wide qq-plot and assesses significance via simulation informed by publicly available allele frequency and LD data. The method can be applied in the context of meta-analysis: it requires as input only the results of association tests for the SNP set. It is flexible and allows for testing whether the relationship between a SNP set and phenotype is different between groups (e.g., by sex, treatment group, or across ethnic groups).
We have shown in simulations that our method is well-powered under a variety of architectures. In particular, QQAUC compares favorably with MAGENTA in scenarios where causal SNPs occur in a minority of pathway genes or when multiple independent causal SNPs per gene affect phenotype. As an example, in simulated pathways where 5 of 14 genes each harbor 1–3 SNPs of modest effect, QQAUC detected the pathway association with 86% power, compared with MAGENTA's 38% power. Finally, we have illustrated the ability of our method to confirm the association of the PPAR pathway with HDL cholesterol (p=0.008).

148

Novel Kernel Function in the Logistic Kernel Machine Test for Pathways in GWA Studies

Saskia Freytag (1) Chris Amos (2) Heike Bickeboller (1) Thomas Kneib (3) Martin Schlather (4)
(1) Department of Genetic Epidemiology, University Medical Center Gottingen (2) Departments of Epidemiology and Biomathematics, University of Texas, MD Anderson Cancer Center (3) Department of Statistics and Econometrics, Georg-August University Gottingen (4) Institute for Mathematics, University of Mannheim

The logistic kernel machine test (LKMT) is a flexible testing procedure tailored towards high-dimensional genetic data. Its use in pathway analyses of GWA case-control studies results from its computational efficiency, flexibility, natural incorporation of covariates and power. Unfortunately, the choice of a kernel function strongly influences the power of the method. The kernel function, which determines the underlying disease-SNP model, can be any positive definite function. Like a prior it can contain additional information. Until now most authors recommended the use of the simple linear kernel function. We propose a novel kernel function containing information about gene membership of SNPs in the pathway. Even this basic genomic structure can improve the ability of the LKMT to identify meaningful associations. Moreover, unlike the linear kernel function, our novel kernel does not suffer from size bias due to appropriate standardization. Size bias is the tendency of larger pathways to reject the null hypothesis of no association with a greater probability just by chance. We apply the LKMT with the linear, our novel as well as 4 further kernel functions to data from the NARAC Rheumatoid Arthritis Consortium. We are not only able to confirm many susceptibility pathways, but also identify associations with a pathway for primary immunodeficiency. Furthermore, we empirically and via a simulation study demonstrate the existence of size bias for the linear kernel function.

149

An Empirical Bayes Method for Unified Association Analysis of a Gene, Region or Pathway Containing Multiple SNPs

Laura C. Lazzeroni (1) Amrita Ray (1)
(1) Stanford University

These days, candidate gene studies and genome-wide association studies include many genetic polymorphisms within the same gene, genomic region or biological pathway. The

resulting wealth of genetic information improves coverage and increases the chance of detecting genetic associations that might otherwise be missed due to a lack of relevant data. However, it can be difficult to sort through a large number of SNP results and decide what is meaningful, especially when hampered by the impact of multiple testing corrections on statistical significance. Furthermore, primary scientific interest is often on the larger unit, such as a gene, rather than the individual SNP. For this reason, various methods have been proposed for unified testing of a gene using data from multiple SNPs. We will discuss the desirable properties of multi-SNP methods such as power, maintenance of nominal significance levels, interpretability of results and accuracy of effect sizes at both the gene and SNP levels, and potential for use in gene-gene and gene-environment interaction analyses. We will propose an empirical Bayes method for unified association analysis of genes, regions or pathways using multiple SNPs and consider how well the method meets the stated criteria.

150

Venous Thromboembolism (VTE)-susceptibility Pathways by Gene Set Analyses

John A Heit (1) Sebastian M Armasu (1) Jason P Sinnwell (1) Daniel J Schaid (1) Mariza de Andrade (1)
(1) Mayo Clinic

Background: Gene set analysis is a promising approach for complementing genome-wide association studies.
Objective: To identify groups of functionally related genes associated with VTE.
Methods: Genome-wide scan genotypes from 1270 white adults with objectively-diagnosed VTE and 1302 controls were mapped to genes, and genes to the Gene Ontology (GO) and KEGG structures. Gene sets analyses were performed using two methods, a score-statistic method, and a p-value combination approach. The analyses were adjusted for age, sex, USA state of residence and stroke/MI status.
Results: After controlling for multiple comparisons, the following three gene sets were significantly associated with VTE: (1) Positive Regulation of Fibrinolysis (*KLKB1, F11, F12, PLG*) within the GO Biological Process structures by both analysis methods, either adjusting for covariates only, or for covariates and *F5* rs6025 (Factor V Leiden); relatively strong single SNP association in *F11* may drive this result; (2) Glycoprotein-fucosylgalactoside Alpha-N-acetylgalactosaminyltransferase Activity [*ABO*] within the GO Molecular Function structures; and (3) Granular Component (*CDKN2AIP, SURF6, FBL*) within the GO Cellular Component (nucleolus) structures. None of the gene set structures within KEGG achieved statistical significance.
Conclusion: Fibrinolysis, glycosyltransferase and nucleolus granular component gene sets are associated with VTE.

151

Bayesian Collapsing Model for Rare Variant Detection

Liang He (1) Samuli Ripatti (2) Janne Pitkaniemi (3)
(1) Department of Public Health, Hjelt Institute, University of Helsinki (2) Institute for Molecular Medicine Finland FIMM, University of Helsinki (3) Department of Public Health, Hjelt Institute, University of Helsinki

For rare variant association analysis, current statistical methodologies suffer from power loss or other serious limitations and weaknesses in certain situations. In this study, we propose a Bayesian collapsing method (BCM), which circumvents the obstacle by introducing random effects to model the signals of rare variants. BCM manages to deal with non-associated variants, allow both protective and deleterious effects, capture SNP-SNP interactions, provide estimates for global and individual contributions and can be applied to both complex traits and case-control design.
We compare the performance of BCM with the Collapsing method with regression, Granvil and KBAC using simulated and GAW 17 data. For complex traits, BCM evidently outperforms the Collapsing method and Granvil in both global power and the accuracy of effect size estimation, particularly when non-associated variants are present. For case-control design, in general cases our model is 5% more powerful than KBAC. In the scenarios where the SNP-SNP interactions are involved, BCM achieves more power by more than 50% compared to KBAC. For GAW 17 data, our model is able to detect the overall association between the trait and the gene, and in the meantime successfully identifies the underlying variants which contribute most and estimates their individual effects. The simulation results show robustness with respect to the proportion of non-associated variants and flexibility in reflecting various association types.

152

Using Bayes Factors to Analyse Fine-mapped Genotype Data

Amy V Baddeley (1) Kevin Walters (1) Angela Cox (2) Wei-Yu Lin (2)
(1) School of Mathematics and Statistics, University of Sheffield (2) Department of Oncology, University of Sheffield

Identifying the causal variant in a disease association region is now a high priority. Methods utilised in published studies to identify causal SNPs include the likelihood ratio (LR) and other frequentist methods. However, high levels of short range LD, rare causal variants and those with small effect sizes mean such analyses may not work in all situations. The restrictive effects of these may be partially countered by incorporating functional biological information into an analysis.
The main method that will be presented uses the Bayes Factor (BF), the ratio of the probability of the data under alternative and null hypotheses, with a larger value indicating more evidence in favour of the alternative hypothesis. Using datasets simulated using HAPGEN, the results of BF analyses is explored, initially using uninformative priors to compare the results to likelihood analysis results, and finally using priors based on functional data. SNP-specific functional data used includes variant location and function, openness of chromatin and conservation across species. It can be seen from the simulations that ranking by BF increases the probability of ranking the causal SNP in the top 1% or 2% compared to ranking by LR, even when using

uninformative priors. When the causal SNP has MAF≥0.05 and per-allele OR≥1.12, these probabilities are well over 0.9. Our results indicate that BFs are a promising tool for incorporating functional information into fine mapping studies.

153

Genotype-based Bayesian Analysis of Gene-Environment Interactions with Multiple Genetic Markers and Misclassifi cation in Environmental Factors

Iryna Lobach (1) Ruzong Fan (2)
(1) New York University School of Medicine (2) NIH/NICHD

Despite the recent yield of candidate gene and genome-wide association studies, the identified genetic variants explain only a small proportion of heritability of complex diseases, such as cancer, hypertension, diabetes, and alcoholism. Part of the unexplained heritability may be described by gene-environment interactions. This work is motivated by the following concerns in the analysis of gene-environment interactions. First, multiple genetic markers in moderate linkage disequilibrium may be involved in susceptibility to a complex disease. Second, many environmental factors are subject to misclassification. We develop a genotype-based Bayesian pseudo-likelihood approach that accommodates linkage disequilibrium in genetic markers and misclassification in environmental factors. Since our approach is genotype-based, it allows the observed genetic information to enter the model directly thus eliminating the need to infer haplotype phase and simplifying computations. This method is based on pseudo-likelihood and hence conventional Bayesian techniques may not be applied directly. We propose a MCMC sampling as well as computationally simple method based on an asymptotic posterior distribution. Simulation experiments demonstrated that our method produced parameter estimates that are nearly unbiased even for small sample sizes. An application of our method is illustrated using a case-control study of interaction between early onset of drinking and genes involved in dopamine pathway.

154

A Multi-marker Genome-wide Association Study: The Story of Bayes C?

Nora Bohossian (1) Mohamad Saad (1) Andres Legarra (2) Maria Martinez (1)
(1) Inserm U1043 – CPTP, Toulouse, France (2) INRA UR631-SAGA, Castanet-Tolosan, France

While the single-marker analysis is still considered as the classical approach for association studies at the genome-wide level, the use of multi-marker approaches is increasingly gaining momentum. Two main classes of such approaches are penalized estimation methods and Bayesian estimation methods. In the former class, all methods use a penalty function to shrink the marker effect estimates towards zero relative to their maximum likelihood estimates. Some methods apply the same large penalty to all estimates whereas others apply large penalties to some estimates and small penalties to the remaining estimates. Thus, the defining factor for these methods is the choice of the penalty function. Bayesian estimation methods, on the other hand, require the specification of a prior distribution on the marker effects. There is a close relationship between these two classes of approaches. In particular, certain penalty functions lead to equivalent estimates under certain priors. In this study, we focus on the Bayesian estimation methods and, specifically, on the Bayes Cπ method (Habier et al., BMC Bioinformatics, 2011 May 23;12:186), which requires the specification of a prior distribution on the marker effects as well as the specification of π, the proportion of markers expected to be associated with the complex trait. We investigate the association results as a function of π in real genome-wide association datasets and present some computational considerations from our experience.

155

Imputation Across Genotyping Arrays for Genome-wide Association Studies: Assessment of Bias and a Correction Strategy

Eric O Johnson (1) Dana B Hancock (1) Joshua L Levy (1) Nathan C Gaddis (1) Nancy L Saccone (2) Laura J Bierut (2) Grier P Page (1)
(1) RTI International (2) Washington University in St. Louis

A great promise of publicly sharing GWAS data through dbGaP is the potential to create composite sets of controls, but studies often use different genotyping arrays. Imputation to a common set of SNPs from different arrays has shown substantial bias, but there remains no broadly applicable solution. Based on the idea that using differing sets of genotyped SNPs across arrays as inputs creates differential imputation error and thus bias in the composite set of controls, we examined the degree to which 1) imputation across arrays based on the union of all genotyped SNPs results in bias as evidenced by spurious associations between imputed genotypes and arbitrarily assigned case/control status; 2) imputation based on the intersection of genotyped SNPs does not evidence such bias; and 3) imputation quality varies by the size of the intersection of genotyped SNPs. Imputations were conducted in studies of European and African descent with reference to HapMap phase III. Imputation based on all genotyped SNPs across the Illumina 1M and 550v3 arrays showed spurious associations in 0.2% of SNPs: ∼2000 false positives per million SNPs imputed. Biases remained problematic for very similar arrays (550v1 vs 550v3) and were substantial for more dissimilar arrays (Illumnia 1M vs Affymetrix 6.0). In all cases, imputing based on the intersection of genotyped SNPs (down to as few as 30% of the total SNPs genotyped) eliminated such bias but still achieved good imputation quality.

156

Comparison of Genotype Imputation Strategies Using 1000 Genomes for African American Studies

Dana B Hancock (1) Joshua L Levy (1) Nathan C Gaddis (1) Laura J Bierut (2) Nancy L Saccone (2) Grier P Page (1) Eric O Johnson (1)
(1) Research Triangle Institute International (2) Washington University in St. Louis

SNP genotype imputation in African Americans using the diverse 1000 Genomes reference panels has had limited evaluation. We compared imputation results for 595 African

Americans genotyped on Illumina HumanHap550v3, using 4 software programs (MaCH, MaCH-Admix, BEAGLE, and IMPUTE2) and 3 reference panels from different combinations of 1000 Genomes populations (February 2012 release): 3 specific populations (YRI+CEU+ASW), 8 European- or African-derived populations (EUR+AFR), and all 14 available populations (ALL). Based on chromosome 22 imputations, we calculated two performance metrics: concordance (percentage of masked genotyped SNPs with imputed and true genotype agreement) and average r2hat (estimated squared correlation between the imputed and true genotypes, for all imputed SNPs). IMPUTE2 and MaCH had the highest concordance for all reference panels (91–93%). IMPUTE2 had the highest quality for all panels: YRI+CEU+ASW average r2hat=0.68, AFR+EUR average r2hat=0.62, and ALL average r2hat=0.55. Imputation quality dropped with the addition of more distantly related panels for all software. Reduced quality was entirely driven by low minor allele frequency (MAF<2%) SNPs. While we achieved optimal imputation performance for common SNPs in African Americans using IMPUTE2 with reference to the 1000 Genomes ALL panel (average r2hat=0.86 for SNPs with MAF≥2%), our findings show that imputation quality of rarer SNPs may benefit from using more closely related reference panels.

157

Assessment of Haplotype Estimation on Two-step Strategies for Large-scale Imputation Projects

xiangjun xiao (1) Jouke Jan Hottenga (2) Maria M Groen-Blokhuis (2) Erik A Ehli (3) Abdel Abdellaoui (2) Eco de Geus (2) James J Hudziak (4) Gareth E Davies (3) Dorret L Boomsma (2) Paul Scheet (1)
(1) MD Anderson Cancer Center (2) VU University (3) Avera Institute for Human Behavioral Genetics (4) UVM College of Medicine

Two-step procedures for imputation have recently been suggested for large-scale reference panels. In these, haplotypes are first estimated. Then, subsequent to this step, imputation is performed assuming these haplotypes are known. This procedure is much faster than one in which haplotypes are jointly estimated with genotypes and is necessary for imputation from extremely large reference panels. Such an approach will result in no loss of information when the precision of haplotype estimates is high. We wished to assess the impact on imputation accuracy in realistic settings. To do so, we used genotype data from a large family-based GWA study, using individuals from the Netherlands Twin Register. We first masked the genotypes at a portion of the observed SNP markers. Next, we leveraged the family structure to obtain highly accurate haplotype reconstructions at the remaining SNP markers using BEAGLE. To mimic less-accurate reconstructions, we manually degraded these haplotypes to varying degrees by forcing switch errors. Finally, we applied minimac to the various sets of estimated and degraded haplotypes, imputing genotypes from both HapMap2 and 1000 Genomes Project reference panels. Preliminary results indicate that performance of a 2-step procedure is highly robust to imperfections in the estimated haplotypes across allelic frequencies. We also characterize the information gain from using a higher density, though potentially less accurate, reference panel.

158

Fast and Accurate 1000 Genomes Imputation Using Summary Statistics or Low-coverage Sequencing Data

Bogdan Pasaniuc (1) Noah Zaitlen (1) Gaurav Bhatia (1) Alexander Gusev (1) Nick Patterson (2) Alkes L Price (1)
(1) Harvard School of Public Health, Epidemiology Dept (2) Broad Institute

Imputation of untyped genotypes using external reference panels is a widely used approach for increasing power in GWAS and meta-analysis. Current HMM-based imputation requires individual level genotypes and is computationally intensive (e.g. Beagle requires more than 2 weeks for imputing 1000 Genomes data into 3,000 samples on 50 nodes). An alternative is to use Gaussian models that employ linear predictors to infer missing data, an approach that has been previously proposed (Wen&Stephens, AnApStat 2010) but has not been tested on empirical 1000 Genomes data. We find that Gaussian imputation recovers 84%(56%) of the effective sample size for common (>5%) and low-frequency (1–5%) variants, vs. 88%(60%) effective sample size recovered by Beagle imputation. Interestingly, even when only summary level data is available, Gaussian imputation of association statistics recovers 83%(53%) of the effective sample size.
In our recent work (Pasaniuc et al. Nat Genet 2012) we showed that GWAS based on extremely low-coverage sequencing can attain several times the effective sample size of array-based GWAS per unit cost. Here we extend Gaussian models to call genotypes from extremely low-coverage sequencing, again using 1000 Genomes reference panels. We show that our approach attains similar accuracy to Beagle at common SNPs with vast reduction in runtime (e.g. ∼1 day on 1 node for 1000 Genomes imputation into 3,000 samples).

159

Imputing Genotypes in Large Pedigrees: A Comparison between GIGI and BEAGLE

Charles Yin Kiu Cheung (1) Ellen M Wijsman (1)
(1) University of Washington

Imputation of dense genotypes facilitates identification of rare risk variants. GIGI is a new genotype imputation approach for large pedigrees. While GIGI uses correlation from identity-by-descent in pedigrees, BEAGLE uses correlation from linkage disequilibrium in the population. Here we used a 95-member pedigree to compare the imputation quality from GIGI and BEAGLE when BEAGLE was used while ignoring the pedigree structure. In a 50 cM region, 60 subjects were observed for SNPs. In an initial analysis, we kept complete data on only 13 subjects, masked all but 35 evenly spaced SNPs on 47 other subjects, and imputed missing genotypes on 288 other SNPs. In BEAGLE, we also evaluated the effects of adding 202 outside reference subjects, as well as providing denser SNPs in the reference panel in a "leave-one-out" analysis. In calling overall genotypes using the most probable genotype configuration, GIGI was more accurate than BEAGLE under the initial setup (79.7% vs. 70.2%) but was less accurate when the outside reference subjects and denser SNPs were included (79.7% vs. 95.4%). However, in calling rare genotypes, GIGI was more accurate under all settings (62.9% vs. 4.5–28.1%) and also detected 46.2% of the rare alleles in the 35 completely untyped relatives, which BEAGLE could not do. These results

demonstrate advantages in use of large pedigrees for identifying rare risk variants through co-transmission of alleles to multiple affected individuals.
Supported by NIH GM046255

160
Improved Criteria for Identifying SNPs that Do Not Impute Well
Shelina Ramnarine (1) Cynthia A. Helms (1) Weimin Duan (1) Juan Zhang (1) Tae-Hwi Schwantes-An (1) Nancy L. Saccone (1)
(1) Washington University School of Medicine

Imputation is used to obtain association results for untyped single nucleotide polymorphisms (SNPs). We address two barriers to assessing true imputation accuracy: actual genotypes are needed to compare with imputed data but are typically unavailable in study data without new genotyping, and commonly used measures (e.g. allelic R^2) overestimate accuracy for rare variants due to chance agreement. In our approach 1000 Genomes reference populations (AFR, ASN, EUR) are masked to match the SNP coverage of several commercial arrays (Illumina and Affymetrix), creating sample data. This optimized reference-sample relationship allows us to obtain upper bounds for accuracy to help filter study data. Imputability is evaluated by several measures including the imputation quality score (IQS), which adjusts for chance agreement. We apply the method to genomic regions associated with smoking. Even at moderate to high minor allele frequency and correlation with typed SNPs, several SNPs have low imputability. Although allelic R^2 often agrees with IQS, in some cases it misclassifies a substantial proportion of SNPs as accurately imputed when compared to recommended IQS accuracy thresholds; e.g. on the CHRNA5/A3/B4 region of chromosome 15 with Affy 500K, 79% of SNPs have allelic $R^2 \geq 0.4$ but IQS < 0.9, and 8% of SNPs have allelic $R^2 \geq 0.9$ but IQS < 0.9. Our findings provide criteria to filter poorly imputed SNPs in association studies and meta-analyses. *Supported by NIDA R01DA026911 and NHLBI T32HL083822*.

161
The Genotype Imputation Endgame: Improving Accuracy in a World of Massive Sequencing
Bryan Howie (1) Matthew Stephens (1)
(1) University of Chicago

Genotype imputation methods are often used to increase power and resolution in association studies of complex traits. The most widely used methods are based on hidden Markov models (HMMs) that approximate the coalescent-with-recombination process. These approximations are effective at capturing patterns of common variation in human populations, but it remains to be seen whether they can also capture the many low-frequency (MAF<5%) and rare (MAF<0.5%) variants that are being discovered in large-scale sequencing studies. Work on isolated populations has shown that when enough haplotypes from a population have been sequenced, unobserved alleles can be imputed with high accuracy using conceptually simple methods that look for long tracts of recent shared ancestry. The same methods should eventually work in large outbred populations, but they may not gain traction until millions of individuals have been sequenced. In the meantime, it is useful to know whether the imputation of rare variants might be improved by using more sophisticated population-genetic models. We address this question by testing a rich collection of models that more flexibly capture patterns of genetic variation. We also consider a range of practical issues that may affect the success of long-range phasing in high-throughput sequence data. Our results from real and simulated data show the possibilities and limitations of imputation methodology as we move toward a world with millions of sequenced genomes.

162
Second Generation DCEG Reference Set Improves Performance of Genotype Imputation
Zhaoming Wang (1) Kevin B Jacobs (1) Meredith Yeager (1) Amy Hutchinson (1) Joshua Sampson (2) Margaret Tucker (2) Stephen J Chanock (2)
(1) SAIC-Frederick/DCEG, NCI (2) DCEG, NCI

Genotype imputation has become an essential analytic approach for fine-mapping and meta analysis of GWAS data sets genotyped and analyzed on different arrays. The NCI Division of Cancer Epidemiology of Genetics (DCEG) recently released a first version of an imputation reference set through dbGaP webportal (Wang et al. Nature Genetics 44:6–7, 2012 available at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000396.v1.p1). Earlier, we showed that performance is better for genotyped samples compared to low-pass sequence data. To improve the practical utility of the DCEG Reference Set, we have sustantially increased the number of genotyped SNPs in the same set of 1,249 cancer-free individuals, primarily of European ancestry. New content includes SNP microarray data from both Illumina (omni2.5S, Exome, Immunochip, Metabochip and the iCOGs custom cancer microarrays) and Affymetrix (6.0 and two Axiom microarrays), which will furhter facilitate larger, international meta-analyses and imputation projects. Ongoing work includes comparison of version 2 to recent builds of the 1000 Genomes Project. Later this year, version 2 of the DCEG Imputation Set will be available on the dbGaP webportal.

163
Genomewide CNV Association Study for Dengue Shock Syndrome
Rongli Zhang (1) Martin L Hibberd (2) Cameron P Simmons (3) Chiea Chuen Khor (4) Anbupalam Thalamuthu (1)
(1) Human Genetics, Genome Institute of Singapore (2) Genome Institute of Singapore; Department of Epidemiology and Public Health, National University of Singapore (3) Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam; Centre for Tropical Medicine,Oxford University, Oxford, UK (4) Genome Institute of Singapore; Centre for Molecular Epidemiology, National University of Singapore

Dengue is globally the most common mosquito-borne infection after malaria, with an estimated 100 million infections occurring annually. The genetic contribution to severe form of dengue remains largely unexplored. The goal of this study is to identify the copy number variants (CNVs) associated with the dengue shock syndrome. Genotypes

were obtained for 2,079 Vietnam pediatric dengue patients and 2,038 Vietnam control subjects using Illumina 660W SNP chip. Two popular CNV typing algorithms QuantiSNP (http://www.well.ox.ac.uk/QuantiSNP/) and the PennCNV (http://www.neurogenome.org/cnv/penncnv/) were used to call CNVs.
Two different approaches were used to identify CNV regions based on the filtered high confident CNV calls from both the CNV typing algorithms. Rare and common CNV regions were tested for association using standard and novel statistical methods.
In total, 128,421 CNVs were identified with an average number of 32 CNVs/subject for both case and control subjects. We observed significantly more deletions than duplications. The association study showed 145 significant CNV regions (p-value <1E-5). We are currently validating some of the rare CNV events overlapping with some interesting candidate genes.

164
Testing Copy Number Variant/Trait Associations Detected Using Manhattan Plots
Glen A Satten (1) Dhanya Ramachandran (2) Jennifer G Mulle (2) Andrew S Allen (3) Lora JH Bean (2) Cheryl Maslen (4) Stephanie L Sherman (2) Roger H Reeves (5) Michael E Zwick (2)
(1) Centers for Disease Control and Prevention (2) Emory University (3) Duke University (4) Oregon Health Sciences University (5) Johns Hopkins University

Finding association between copy number variants (CNVs) and disease is hampered by poor detection of CNVs using log-intensity ratio (LIR) data. Breheny et al. (Plos One 2012; 7(4):e34262) suggested comparing LIR values between cases and controls, rather than copy number states, obtaining a p-value at each locus. CNV-trait association is then found by using CNV-detection software on the Manhattan plot of $-\log(p)$ values. Using this approach on data from children with Down syndrome (DS) with atrioventricular septal defects (cases, n=238) and with DS but without heart defects (controls, n=267), we found several candidate associations. We tested significance with a bootstrap hypothesis test. We fit a regression model for LIR at each locus, including case status and genotypic principal components to control for population stratification. We resampled residuals by individual to preserve correlations, generating 1000 bootstrap replicate datasets that preserve population structure but have no other difference between cases and controls. We used the CNV BEAST (http://www.duke.edu/$\sim$asallen/Software.html) to find runs of large $-\log(p)$ values. Because the CNV BEAST gives a score to each CNV, we compared the largest scores from each replicate to that observed in the actual data. Although our largest signal seemed robust (>100 adjacent loci on Chr 1), we found a 7.2% chance of seeing such a large signal under the null hypothesis.

165
Rare Copy Number Variation in Neuropsychiatric Disorders: Exploring the Phenotype
Alison K Merikangas (1) Aiden P Corvin (1) Louise Gallagher (1)
(1) Trinity College Dublin

Background: There is emerging evidence that copy number variants (CNVs) provide a new vista on understanding unique and pleiotropic susceptibility to neuropsychiatric disorders such as Autism Spectrum Disorders (ASD) and schizophrenia.
Methods: Rare CNV and detailed phenotype data were derived from the Autism Genome Project and Irish schizophrenia cases. Patients were classified by whether a rare CNV impacted any genes previously implicated in ASD or Intellectual Disability (ID) or not (0/1), or any genes that are differentially brain expressed (BE) or not (0/1), and association with candidate neurodevelopmental phenotypes were examined. Random forests and mixture models were used to explore whether phenomic features identify CNV-associated sub-groups.
Results: No statistically significant univariate associations between CNVs and selected phenotypes were identified for either ASD or schizophrenia. Exploratory analyses suggest sub-phenotypes that might provide good targets for association analyses in future studies, and indicate that distinguishing deletions and duplications is important. Inconsistency of measurements by site in large collaborative studies is a major impediment to assessment of genotype-phenotype associations.
Discussion: Sophisticated modeling suggests that CNV-associated subgroups may exist, however the clinical applicability of these remain to be demonstrated.

166
Genome-wide Disease Association Studies of Inversion Variants
Jianzhong Ma (1) Christopher I Amos (1)
(1) Department of Genetics, U.T. MD Anderson Cancer Center

Although chromosomal inversions are a ubiquitous class of structural variation, understanding of the prevalence of inversions and their role in human disease has been lagged behind other types of structure variants. Standard cytogenetic approaches, such as fluorescence in situ hybridization (FISH)-based assays, and recently developed paired-end sequencing approaches do not efficiently apply to large number of samples that are needed to characterize inversions in a population and detect their association with diseases. We have recently proposed a cost-efficient approach of detecting and characterizing inversions using widely available high-density genotype data of single nucleotide polymorphisms (SNPs) for genome-wide association studies (GWAS).
Our approach, based on principal components analysis (PCA), applies to non-recurrent inversions for which recombination between the inverted and non-inverted segments in inversion heterozygotes is suppressed due to the loss of unbalanced gametes. Locally performing PCA in the inversion region, one can classify a large number of samples into three clusters corresponding to the three inversion genotypes, making it possible to perform a disease-inversion association test.
Using inversion polymorphisms reported in the literature and our predicted inversions, we applied our approach to available case-control data for GWAS and identified a few inversions associated to melanoma and/or lung cancer.

167

Epigenomic Indicators of Age

Alicia L. Lazarus (1) Jennifer A. Smith (1) Thomas H. Mosley Jr. (2) Stephen T. Turner (3) Yan V. Sun (4) Sharon L.R. Kardia (1)
(1) Department of Epidemiology, University of Michigan (2) Department of Medicine (Geriatrics), University of Mississippi Medical Center, Jackson, MS (3) Division of Nephrology and Hypertension, Mayo Clinic (4) Department of Epidemiology, Emory University

Age is a well-established risk factor for chronic disease, though functional pathways to link age to cellular disease processes are not well understood. Epigenetic mechanisms may elucidate a molecular relationship between age and chronic disease. DNA methylation data (M-Values) were collected on 1,008 African-Americans (age 39–94 years) from the Genetic Epidemiology Network of Arteriopathy study using the Illumina HumanMethylation27 platform. As a predictor for methylation, age was significantly associated with M-Values of 7,157 sites at $p<1.89\times10^{-6}$ (Bonferroni-corrected $\alpha<0.05$). Because the relationship between age and epigenetic variation was so strong for the top 25 markers ($p<10^{-23}$), we hypothesized that DNA methylation could be a molecular indicator of age. Consequently, we assessed how well epigenetic markers predict age. There were 1,848 sites significantly associated with age as the outcome at $p<1.89\times10^{-6}$. To evaluate the combined predictive capacity of the epigenome, we estimated the principal components (PCs) using the 1,848 sites. The top 5 PCs explained 66% of the variance in epigenetic marker M-Values, and the top 5 epigenetic PCs explained 29% of variation in chronological age ($p=2.36\times10^{-76}$). We also performed this analysis using Beta Values and results were similar. This study indicates that DNA methylation and cellular aging processes are significantly related, and could underlie a range of processes associated with age-related chronic diseases.

168

Age-Associated Methylation Profiles in a Hypertensive African-American Population

Kristina Jordahl (1) Siying Huang (2) Travis Hughes (3) Caren Weinhouse (2) Alicia Lazarus (3) Jennifer Smith (3) Thomas Mosley Jr (4) Stephen Turner (5) Sharon Kardia (3)
(1) Department of Biostatistics, University of Michigan (2) Department of Environmental Health Sciences, University of Michigan (3) Department of Epidemiology, University of Michigan (4) Department of Medicine (Geriatrics), University of Mississippi Medical Center (5) Division of Nephrology and Hypertension, Mayo Clinic

Variation in DNA methylation (DNAm) associated with chronological age may provide information about the molecular nature of the aging process. To better understand the correlation structure among age-associated DNAm markers and its relation to biological pathways, we measured DNAm from blood leukocyte samples of 972 African-Americans within hypertensive sibships on the Illumina HumanMethylation27K BeadChip platform. Using a linear mixed effects model with a random intercept to control for relatedness, we identified 1385 DNAm sites that were significantly associated with age at the Bonferroni adjusted significance level of $1.89*10^{-6}$. Among the age-associated sites, 92% showed a positive association and 8% showed a negative association with age. The distribution of pairwise correlations between age-associated DNAm sites was 71% with $R^2 < 0.15$, 28% with R^2 between 0.15 and 0.75, and <1% with $R^2 > 0.75$. K-means clustering was used to group sites based on the correlation between DNAm beta values across individuals sorted by increasing age. Using consensus clustering, the most stable grouping had k=4 clusters. Cluster 1 had significantly enriched gene ontology (GO) terms related to stimulus response, receptor and transducer activity, and the extracellular region, and cluster 2 showed significant enrichment of cellular component GO terms related to intracellular parts, cytoplasm, and organelles.

169

Patterns of SNP-based Genome-wide Heritability of Methylation in Four Brain Regions

Gerald Quon (1) Christopher Lippert (1) David Heckerman (1) Jennifer Listgarten (1)
(1) Microsoft Research

We used a mixed model to investigate the extent to which genetic variation in DNA sequence explains variation in CpG dinucleotide methylation in data from 150 individuals for four brain regions. In particular, our goals were to investigate what role cis-DNA sequence plays in influencing methylation, what an optimum definition of cis (*i.e.*, locality) is in this context, and whether CpG dinucleotides with heritable methylation were more likely to be associated with particular classes of genes.
We found that 5–7% of CpG sites assayed were heritable, with a median narrow-sense heritability of 21% (and a mean of 3.5% over all sites) when using an optimal cis window of 50kb. We also found good concordance across brain regions. Our percentage of significant sites is similar to work by Bell *et al.*, who found that 6.3% of CpG sites tested had significant QTLs, although they found a mean genome-wide heritability of 18%, higher than ours, unsurprisingly given that they used a twin-based correlation estimate. Finally, we show that the set of genes potentially targeted by these methylation loci are enriched in human leukocyte antigen (HLA) genes, which are important in a variety of developmental and immune processes. Our estimates of heritability are conservative, and we suspect that the number of heritable loci will increase in the near future as the methylome is assayed across a broader range of tissue and cell types and the density of the tested loci is increased.

170

Comparing Count Regression Models to Investigate the Relationship Between Genomic Instability and Gene Methylation in Human Hepatocellular Carcinoma

Miriam Kesselmeier (1) Thomas Longerich (2) Olaf Neumann (2) Robert Geffers (3) Justo Lorenzo Bermejo (1)
(1) University Hospital Heidelberg, Institute of Medical Biometry and Informatics (2) University Hospital Heidelberg, Institute of Pathology (3) Helmholtz Center for Infection Research, RG Genome Analytics

A better understanding of the relationship between chromosomal alterations and gene methylation may facilitate the identification of relevant steps in the development of human hepatocellular carcinoma (HCC). Accumulation of genetic alterations can be explored using information gained

by array-based genomic hybridization (aCGH). A collection of patients with aCGH, gene expression and methylation has been generated by the SFB/TRR77 "Liver Cancer – From Molecular Pathogenesis to Targeted Therapies" [http://www.livercancer.de].
We have investigated the relationship between genomic instability and gene methylation for about 600 genes, previously proposed to define expression-based subclasses of HCC (Hoshida et al., 2009, Cancer Res 69: 7385–7392). We used linear and Poisson (robust) regression models. Genomic instability was represented by the rate of chromosome arms with aCGH-based gain or loss of genetic material.
Complete information was available for 54 tumor samples with 40 to 43 chromosome arms. The median number of instable arms was 15 (range 4–37). Robust Poisson regression showed the highest prediction accuracy but also the highest variability and computational time. According to deviance-based forward variable selection, all four regression models included the methylation of a single gene. Genomic instability and the methylation of this gene were negatively correlated (Spearman's rank rho=-0.61). Technical details and additional results will be discussed during the meeting.

171
An Epigenetic Intersection of Age, Inflammation, and Kidney Function
Jennifer A Smith (1) Alicia L Lazarus (1) Thomas H Mosley (2) Steven T Turner (3) Sharon LR Kardia (1)
(1) Department of Epidemiology, University of Michigan (2) Department of Medicine (Geriatrics), University of Mississippi Medical Center (3) Division of Nephrology and Hypertension, Mayo Clinic

Kidney function declines with age, leading to chronic kidney disease (CKD). However, the molecular mechanisms that mediate this process are not well known. Since DNA methylation (DNAm) patterns demonstrate ubiquitous and strong associations with age, they may be an important measure of cellular aging underlying the relationship between kidney function and age. To identify the DNAm markers associated with age and estimated glomerular filtration rate (eGFR), a primary indicator of kidney function, we measured 26,428 DNAm markers in the blood leukocytes of 972 African Americans (ages 39–94) from the Genetic Epidemiology Network of Arteriopathy (GENOA) study using the Illumina HumanMethylation27 BeadChip. We identified 146 DNAm markers associated with eGFR at $p<10^{-4}$, and 131 (90%) of these were also associated with age. To better understand the mechanisms through which age influences eGFR, we assessed the relationship between the 131 significant markers and demographic, traditional, and novel risk factors for CKD, including markers of inflammation. We found, for example, that 89 (67%) of the 131 markers were associated with interleukin-6 (IL-6) levels, suggesting that age and inflammation may have joint effects on methylation patterns related to declines in kidney function. Since the majority of DNAm markers associated with eGFR were also associated with age and IL-6, this research provides an innovative step in mapping the pleiotropic epigenetic mediators of CKD development.

172
Epigenetic Association with Plasma Homocysteine Levels in African Americans
Qiuzhi Chang (1) Stephen T Turner (2) Sharon LR Kardia (3) Yan V Sun (1)
(1) Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia (2) Division of Nephrology and Hypertension, Mayo Clinic, Rochester, Minnesota (3) Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan

Homocysteine is a downstream product of S-adenosyl methionine, the methyl donor of DNA methylation. Increased plasma homocysteine levels have been shown to be an important risk factor for endothelial dysfunction and atherosclerosis, which can be mediated by the modification of DNA methylation (DNAm). To identify genes and pathways that are associated with plasma homocysteine levels through DNAm, we conducted a methylome-wide association study of 22,927 autosomal DNAm sites using peripheral blood leukocytes from 972 African Americans. DNAm profile was measured using the Illumina Infinium HumanMethylation27 BeadChip. Controlling for age, sex, BMI and cigarette smoking, hypermethylation of the two DNAm sites were significantly associated with high plasma levels of homocysteine corrected for multiple testing. The two DNAm sites located in *IL7R* (Interleukin-7 receptor) and *FGD2* (FYVE, RhoGEF and PH domain-containing protein 2), which both play important roles in the immune response. While the underlying pathogenic mechanism of homocysteine is not complete, these findings suggest that the epigenetics of specific genes, especially those that regulate the inflammatory response, may be modulated by elevated plasma homocysteine levels. Understanding the molecular and epigenetic mechanisms triggered by increased plasma homocysteine levels will progress the understanding of its role in disease development and prevention.

173
Differential Methylation of the Arsenic (III) Methyltransferase Promoter According to Arsenic Exposure: Preliminary Evidence from the Strong Heart Study
Matthew O Gribble (1) Wan Y Tang (2) Yan Shang (2) Jonathan Pollak (2) Jason G Umans (3) Kevin A Francesconi (4) Walter Goessler (4) Ellen K Silbergeld (2) Eliseo Guallar (1) Shelley A Cole (5) M. Daniele Fallin (1) Ana Navas-Acien (2)
(1) Johns Hopkins Bloomberg School of Public Heath, Department of Epidemiology (2) Johns Hopkins Bloomberg School of Public Heath, Department of Environmental Health Sciences (3) Georgetown-Howard Universities Center for Clinical and Translational Science (4) Karl-Franzens University, Institute of Chemistry – Analytical Chemistry (5) Texas Biomedical Research Institute

Inorganic arsenic is methylated in the body by arsenic (III) methyltransferase, an important enzyme for arsenic metabolism. Arsenic methylation is thought to play a role in arsenic-related epigenetic phenomena including abnormal DNA and histone methylation. However, it is unknown whether the promoter of the AS3MT gene, which codes for arsenic (III) methyltransferase, is differentially methylated as a function of arsenic exposure. In this pilot study we evaluated AS3MT promoter methylation according to exposure, assessed by urinary arsenic excretion in a stratified random sample of participants from the Strong Heart Family Study who had DNA available from 1989–1999 and 1998–1999.

We selected 16 participants within each study region (Arizona, Oklahoma, and North/South Dakota), including 8 with high and 8 with low inorganic arsenic exposure per study region (total N=48). We measured promoter methylation at 48 CpG loci by bisulfite sequencing of clones. We compared mean% methylation at each CpG locus by arsenic exposure group using linear regression adjusted for region and sex, with bootstrap resampling of observations within study regions. A differentially hypomethylated region in the AS3MT gene was associated with higher arsenic exposure. These findings are preliminary but may suggest that arsenic exposure could affect the epigenetic regulation of a major arsenic metabolism gene.

174

Integrating Genome-wide Gene Expression and Genotype Data to Predict HDL Cholesterol Levels in the Cholesterol and Pharmacogenetic Study (CAP)

Emily R Holzinger (1) Scott Dudek (1) Alex Frase (2) Marisa Medina (3) Ronald Krauss (3) Marylyn Ritchie (2)
(1) Vanderbilt University (2) Pennsylvania State University (3) Children's Hospital Oakland Research Institute

Technology is driving the field of human genetics research with advances in the ability to generate high-throughput data that interrogate various levels of biological regulation (genomic, transcriptomic, proteomic, etc). An effective study design for this data may be a systems biology approach that integrates different types of data, or a meta-dimensional analysis.
For this study, we integrated ~3 million SNPs and ~25,000 expression variables (EVs) from 480 individuals that participated in the CAP simvastatin clinical trial to predict HDL cholesterol (HDL-C) levels. First, we used a variable selection method that allows for main and interaction effects to filter the data set. Next, we used the filtered variables to find significantly enriched pathways that cluster into non-redundant biological functions. The most significant cluster involved cellular membrane pathways (p-value 1.8E-4). Cellular membrane processes play a major role in HDL-C synthesis and metabolism. Finally, to generate more parsimonious meta-dimensional models, we integrated the filtered SNPs and EVs using the Analysis Tool for Heritable and Environmental Network Associations (ATHENA). Our best model consisted of 3 EVs and one SNP and explains 9.4% of the variation in HDL-C. This systems biology approach is proof of concept that with the right tools, different types of high-throughput data can be integrated to generate models that predict complex human traits.

175

Integrative Analysis of Single Nucleotide Polymorphisms and Gene Expression Efficiently Distinguishes Samples from Closely Related Ethnic Populations

Hsin-Chou Yang (1) Pei-Li Wang (1) Chien-Wei Lin (1) Chien-Hsiun Chen (2) Chun-Houh Chen (1)
(1) Institute of Statistical Science, Academia Sinica (2) Institute of Biomedical Sciences, Academia Sinica

Single nucleotide polymorphisms (SNPs) are useful for classifying individuals from distinct continental origins but cannot discriminate individuals with subtle genetic differences from closely related ancestral lineages. Proof-of-principle studies have shown that gene expression (GE) also is a heritable human variation that exhibits differential intensity distributions among ethnic groups. This motivated us to integrate SNPs and GE to construct ancestry informative marker panels with a reduced number of required markers and provide high accuracy in ancestry inference. We integrated a forward selection procedure into flexible discriminant analysis to identify key SNPs and/or GE with the highest cross-validation prediction accuracy. User-friendly BIASLESS software was developed. By analyzing genome-wide SNPs and/or GE in 210 independent samples from four ethnic groups in the HapMap II Project, we found that average testing accuracies for a majority of classification analyses were quite high, except for SNP-only analyses that were performed to discern samples from two close Asian populations. The average testing accuracies ranged from 0.53 to 0.79 for SNP-only analyses and increased to around 0.90 when GE markers were integrated together with SNPs for the classification of samples from closely related Asian populations. In conclusion, integrative analysis of SNPs and GE provides high-accuracy and cost-effective classification and prediction results for ancestry inference.

176

Directed Causal Network Construction Using Linkage Analysis with Metabolic Syndrome-Related Expression Quantitative Traits

Sung-il Cho (1) Kyee-Zu Kim (1) Jin-Young Min (2) Geun Yong Kwon (3) Joohon Sung (1)
(1) Seoul National University Graduate School of Public Health (2) Seoul National University Institute of Health and Environment (3) Division of Epidemic Intelligence Service, Korea Center for Disease Control and Prevention

In this study, we propose a novel, intuitive method of constructing an expression quantitative trait (eQT) network that is related to the metabolic syndrome using LOD scores and peak loci for selected eQTs, based on the concept of gene-gene interactions. We selected 49
eQTs that were related to insulin resistance. A variance component linkage analysis was performed to explore the expression loci of each of the eQTs. The linkage peak loci were investigated, and the "support zone" was defined within boundaries of an LOD score of 0.5 from the peak. If one gene was located within the "support zone" of the peak loci for the eQT of another gene, the relationship was considered as a potential "directed causal pathway" from the former to the latter gene. SNP markers under the linkage peaks or within the support zone were searched for in the database to identify the genes at the loci. Two groups of gene networks were formed separately around the genes IRS2 and UGCGL2.
The findings indicated evidence of networks between genes that were related to the metabolic syndrome. The use of linkage analysis enabled the construction of directed causal networks. This methodology showed that characterizing and locating eQTs can provide an effective means of constructing a genetic network.

177

Evidence for HDL-associated Variation in T-cell Receptor Gene Expression

Ellen E Quillen (1) Harald HH Göring (1) Eugene Drigalenko (1) Matthew P Johnson (1) Jean W MacCluer (1) Thomas D Dyer (1) Eric K Moses (1) Michael C Mahaney (1) Joanne E Curran (1) John Blangero (1) Laura Almasy (1)
(1) Texas Biomedical Research Insititute

High-density lipoprotein (HDL) levels, a common predictor of heart disease, are determined by genetic and environmental factors. Previous research on the genetic basis of this variation has had limited results. Examination of gene expression may bridge the gap and aid in identifying novel risk loci. Expression levels of 47,289 transcripts were characterized for lymphocytes in 1243 Mexican American participants in the San Antonio Family Heart Study. 117 transcripts from 99 genes show significant genetic correlation ($p<5E-5$) with HDL levels. Pathway analysis was used to identify biological domains enriched for these 99 genes. Enrichment was determined by an empirical p-value derived from comparing the test set to 100 randomly selected sets of 99 genes represented on the array. Of 723 pathways ascertained from Ariadne Pathway Studio, HDL-associated genes were overrepresented ($p<0.05$) in four. Three are signaling pathways for T-cell receptors which recognize MHC antigens and trigger transformation into cytotoxic or helper T-cells. The inhibitory natural killer cell receptor pathway, which produces immunoglobulin-like receptors responsible for the recognition of cell-surface MHC I antigens, also appears involved. While the anti-atherosclerotic role of HDL is well documented and focuses predominantly on inflammation, the genetic correlations between variation in expression of these genes and HDL levels suggest an additional role for HDL in immune cell function.

178

Building and Assessing Protein-Protein Interaction Networks from Genome Wide Association Results in Cancer

Linda T Hiraki (1) Amit Joshi (1) Sara Lindstrom (1) Andrew T Chan (2) Stephen Chanock (3) Peter Kraft (1)
(1) Harvard School of Public Health (2) Massachusetts General Hospital (3) National Cancer Institute

Genome wide association studies (GWAS) have identified hundreds of single nucleotide polymorphisms (SNPs) associated with cancer risk yet, identifying the functional genes remains a challenge. Mendelian disease studies have shown causal gene protein products tend to physically interact. We hypothesized that genes with protein-protein interactions (PPIs) with GWAS-identified and Mendelian cancer loci are likely involved in carcinogenesis. We used the Disease Association Protein-Protein Link Evaluator (DAPPLE) software to explore PPI networks for genes within loci identified by GWAS including 26 breast (BRCA), 25 prostate (PCA) and 16 colorectal (CRC) cancer SNPs. We found significantly more direct connections among GWAS-identified CRC loci than expected by chance, but no significant enrichment for BRCA ($p=0.33$) or PCA ($p=0.77$) loci. DAPPLE identified several genes more connected to other risk loci than expected by chance ($p<0.05$): *MAP3K1* for BRCA and *C11orf53, EIF3H*, and *GREM1* for CRC. Several genes outside of the GWAS-identified loci also demonstrated higher interaction with seed proteins. This approach leverages GWAS and PPI information to highlight potential mechanistic pathways and therapeutic targets for cancer. We will expand our investigation to other cancers and assess whether candidate genes nominated by DAPPLE are enriched for novel mutations using the large, agnostic GWAS conducted by the Collaborative Oncology Gene-environment Study (COGS).

179

Assessing Heterogeneity of cis eQTLs Across Eight Ancestry Groups

Christopher P Grace (1) John C Whittaker (2) Julie Huxley Jones (2) Emmanouil T Dermitzakis (3) Mark I McCarthy (1) Andrew P Morris (1)
(1) Wellcome Trust Centre for Human Genetics, University of Oxford (2) GlaxoSmithKline, Medicines Research Centre (3) Department of Genetic Medicine and Development, University of Geneva Medical School

Trans-ethnic meta-analysis can be used to increase power to detect complex trait loci and improve fine-mapping resolution. Fixed-effect meta-analysis has been performed to locate *cis* eQTLs in eight phase III HapMap samples: 107 CEPH, 79 Chinese, 75 Gujahati Indians, 81 Japanese, 83 Luhya, 41 Mexican, 135 Maasai and 108 Yoruba.
5765 eQTLs were identified (FDR: 5%, $p<2.26\times10^{-4}$), and evidence of heterogeneity (Cochran's Q $p<1\times10^{-3}$) was detected at 246 eQTLs (binomial test $p<2.2\times10^{-16}$). Of these, 49 with signals ($p<0.05$) in all populations, and 52 of the remaining 194 had signals with opposite effect size direction. Signal overlap across populations was assessed at FDR 5%. For probe matches only the range was 0.157 to 0.598, for exact SNP probe matches the range was 0.115 to 0.604. Specific examples of heterogeneity were selected for analysis.
Probe *ILMN_22465_2350368* (HGNC: *PTER*) has peak signal SNP rs7909832 ($p=1.90\times10^{-253}$, Cochran's Q $p=1.22\times10^{-15}$) with signals in all populations. The difference in effect-sizes can be explained by differing LD (Linkage Disequilibrium) across the populations. Probe *ILMN_13285_4210136* (HGNC: *FANCA*) has peak signal SNP rs2239360 ($p=9.23\times10^{-89}$, Cochran's Q $p=9.23\times10^{-89}$) with signals in six populations, but not in Chinese ($p=0.50$) and Japanese ($p=0.12$) groups. In this case differences are not obviously due to LD, to investigate whether this is due to genotyping density 1000 genomes imputation of this dataset is planned.

180

Fast Genome-Wide QTL Association Mapping with Pedigrees

Hua Zhou (1) Eric Sobel (2) Kenneth Lange (2)
(1) North Carolina State University (2) University of California, Los Angeles

Genome-wide association studies (GWAS) have successfully identified many common genetic variants associated with some complex diseases and traits. However, it is widely accepted that these common variants explain at most a small fraction of the population variation of most complex traits. This results in a renewed interest in linkage-type analysis to detect rare variants. Family designs allow for control of population stratification and study of parent-of-origin effects. The burgeoning next-generation sequencing (NGS) tools are promising to increase the power of family-based approaches. However, pedigree likelihoods are notoriously hard to compute and current softwares for association mapping in pedigrees are prohibitively slow for dense

marker maps. In this paper we re-examine the computation bottlenecks and propose ultra fast score tests for association mapping with pedigree-based GWAS or NGS data. In general, our strategy works for random sample data, pedigree data, or a mix of both, allows for covariate adjustment and correction for population stratification, and accommodates both univariate and multivariate quantitative traits. The proposed method is implemented in our comprehensive genetic analysis software MENDEL for easy use by the genetics community.

181

Sequence Kernel Association Test in Family Data

Han Chen (1) James B Meigs (2) Josee Dupuis (3)
(1) Department of Biostatistics, Boston University School of Public Health (2) General Medicine Division, Massachusetts General Hospital, and Department of Medicine, Harvard Medical School (3) Department of Biostatistics, Boston University School of Public Health, and NHLBI's Framingham Heart Study

Rare variants have become of great interest in genetic association studies due to technology development in exome sequencing and whole genome sequencing. Recently, the Sequence Kernel Association Test (SKAT) has been shown to perform well in the association study of rare genetic variants. SKAT tests only one parameter, regardless of the number of variants, thus reduces the degrees of freedom in the test and greatly increases the power. Unlike many other rare variants analysis methods, SKAT does not make any assumption about the effect size or the direction of effect of each variant in the test. However, the original SKAT can only be applied to unrelated individuals and not to analysis of rare variants in families. Here we extend SKAT to related individuals and develop familial SKAT (famSKAT) to be used in family data. In a simulation study without associated variants, we show that the original SKAT has inflated type I error when used in family data without accounting for the familial correlation, while famSKAT does not have inflated type I error. In a simulation study with associated variants, we show that famSKAT has more power when related individuals are present, compared to applying the original SKAT to a subset of unrelated individuals. As the proportion of related individuals increases, the power difference between famSKAT and the original SKAT also increases. To illustrate, we apply our approach to glycemic traits from related subjects of the Framingham Heart Study.

182

PRIMUS: Pedigree Reconstruction and Identification of the Maximum Unrelated Set

Jeffrey Staples (1) Deborah A. Nickerson (1)
Piper E. Below (1)
(1) The University of Washington

Recently, researchers have successfully leveraged familial relationships to attain the necessary power in analyses to identify rare causes of disease (e.g. Kabuki and Freeman-Sheldon syndromes), leading to a renewed interest in family-based analysis of genetic diseases, in which obtaining accurate pedigree information is crucial. Reconstruction of pedigrees is useful both to verify these clinically ascertained pedigrees as well as to reconstruct cryptic pedigrees. Previous methods for reconstructing pedigrees are incapable of handling missing samples (gaps) in the family, large and/or multigenerational pedigrees, or non-monogamous relationships.
We have developed a program (PRIMUS) that uses genome-wide Identity By Descent (IBD) estimates to quickly reconstruct large, arbitrary, human and non-human pedigrees that may include gaps as distant as first cousins. PRIMUS uses genome-wide estimates of IBD to predict the type of familial relationship between each pair of individuals in the dataset, reconstructs the most likely pedigree given the data. PRIMUS reconstructs highly informative pedigrees in seconds, but slows as information content drops. PRIMUS can use affection status and reconstructed pedigrees to select the optimal samples for exome or whole-genome sequencing, and PRIMUS will aid researchers in verifying family data and generating previously unknown pedigrees from DNA, allowing them to utilize these familial relationships to increase power in disease studies.

183

Incorporating Parental Information into Family-based Association Tests

Zhaoxia Yu (1) Daniel Gillen (1) Carey F Li (1) Michael Demetriou (1)
(1) University of California, Irvine

Information regarding the true underlying genetic model, i.e., the mode of inheritance, is crucial in detecting association. For family-based association studies, we show that the underlying genetic model can be learned from parental data. Specifically, for parental mating type data, we propose a novel statistic to test whether the underlying true genetic model is additive, dominant, or recessive; for parental genotype-phenotype data, we propose several strategies to learn the underlying true genetic model. We then illustrate how to incorporate the learned information into family-based association tests. Because family-based association tests are conducted conditional on parental genotypes, their type I error rates are not inflated by the information learned from parental data, even if such information is weak or learned when the assumption of Hardy-Weinberg equilibrium is violated. Our simulations demonstrate that incorporating parental data into family-based association tests improves the power of association tests. The application of our proposed methods to a candidate-gene study of type 1 diabetes successfully detects a recessive effect that would otherwise be missed using the conventional family-based association tests.

184

Accounting for Relatedness in Genomewide Association Studies: An Empirical Methods Comparison

Jakris Eu-Ahsunthornwattana (1) Nancy Miller (2) Matti Pirinen (3) Michaela Fakiola (2) The WTCCC2 (4) Selma MB Jeronimo (5) Jenefer M Blackwell (6) Heather J Cordell (1)
(1) Newcastle University (2) University of Cambridge (3) University of Oxford (4) The Wellcome Trust (5) Universidade Federal do Rio Grande do Norte (6) The University of Western Australia

Linear mixed models (LMMs) have been proposed for modelling population substructure and relatedness in

genomewide association studies. Here we compare the performance of several different LMM approaches (and software implementations, including EMMAX, GenABEL and FastLMM) via their application to a genomewide association study of visceral leishmaniasis in 348 Brazilian families comprising 3626 individuals (1970 genotyped and phenotyped individuals). The programs compared differ in the precise details of the methodology implemented and through various user-chosen options such as the method and number of SNPs used to estimate the kinship matrix. We investigate sensitivity to these choices and the success (or otherwise) in controlling the overall genomewide error-rate in both real and simulated data. We also compare our results to those obtained using alternative approaches based on extensions of standard case/control methods (and thus arguably more natural for dichotomous traits such as disease status) implemented in the software packages MQLS and ROADTRIPS. Overall we find strong concordance between the results from different LMM approaches and high correlation between the results from the LMM approaches and alternative approaches, although precise localisation of top SNPs in observed association signals can vary between methods. The software implementations also vary in speed, with analysis of the same data set taking anywhere between a few hours and several weeks.

185

A Clustered Mann-Whitney Approach for High-dimensional Family-based Genetic Association Studies

Qing Lu (1)
(1) Michigan State University

Although family studies were the basis for genetic research before the advent of modern molecular markers, they have been much less developed for association studies of complex diseases using high-dimensional data. Family studies offer many ideal features for large-scale genetic research. For instance, they provide robust protection against confounding bias when dealing with samples from multiple ethnic groups (i.e., population stratification). In this study, we developed a clustered Mann-Whitney approach for high-dimensional family-based association studies. The new approach offers robust protection against population stratification, as well as improved accuracy by considering both with-family and between-family information. Through simulation studies, we evaluated the performance of the new approach under various disease models and the presence of population stratification. The new approach was then applied to a large-scale genome-wide dataset from an international cleft lip with or without cleft palate (CL/P) study, revealing a potential new locus associated with CL/P.

186

On the Analysis of Imputed Genotypes in Family-based Association Studies

Aurelie Cobat (1) Alexandre Alcais (2) Erwin Schurr (1)
(1) McGill University (2) INSERM

Genotype imputation is an efficient statistical approach for inferring genotypes at ungenotyped variants. However, because of the probabilistic nature of imputed SNPs specific techniques are needed for their analysis to account for genotypic uncertainty. One of the most popular approach relies on the expected allele count $E_1=2P_{11}+P_{12}$, also called allele dosage, to test for association. In population-based designs, testing of association between the phenotype and expected allele count is straightforward using standard linear or logistic regression frameworks. By contrast, equivalent methods for family-based studies are missing. To close this gap, we developed a new analytical approach which is based on the methodology proposed for the analysis of copy-number variants[1] to perform association testing of imputed SNPs in the family-based design (FBAT-dosage). Simulation studies under the null hypothesis showed that the FBAT-dosage method provides very consistent type I errors, whatever the level of certainty with which the marker under study was imputed. Under the alternative hypothesis, FBAT-dosage provides substantial increase in power compared to the Best-guess genotype approach for most scenarios.The FBAT-dosage method has been implemented in C++ language and is suitable for genome-wide imputation analysis.

1. Ionita-Laza I et al. Genet Epidemiol. 2008 Apr;32(3):273–84.

187

A Novel Method to Identify Highly Penetrant Pedigree-specific Facial Morphological Phenotypes for Genetic Mapping

Craig C Teerlink (1) Yen-Yun Yu (1) Preston T Fletcher (1) Lisa A Cannon-Albright (1) Alun W Thomas (1)
(1) University of Utah

The heritable nature of normal human facial morphological variation is well established and suggests the presence of highly penetrant Mendelian components that could be mapped to the genome via a pedigree-based design. We propose a novel method to identify the linear combination of facial features that best distinguishes a given family from other sampled people which represents a multivariate combination of features that could best be mapped in that family. For a set of pedigrees with face image data, we first assign a set of points to each image corresponding to prominent facial features. Then, taking one pedigree at a time, we assign the selected pedigree to one group and all other pedigrees to a second group. We then use linear discriminant analysis to identify the linear combination of points that best distinguishes the two groups. Each person in the selected pedigree is scored according to their level of agreement with the linear combination of points. The resulting score can then be used as a quantitative trait in a pedigree-specific linkage scan. We have successfully demonstrated the feasibility of our method on a set of 300 individuals from 50 multigenerational pedigrees with photographic images. We will apply the method to 45 3-generation CEPH pedigrees with face image data and high-density genotypes from the Utah Genetics Reference Project which can be used in an attempt to map the resulting highly penetrant pedigree-specific phenotypes to the genome.

188

Randomized Family-based Association Tests Conditional on Latent Inheritance Vectors

Yanming Di (1)
(1) Oregon State University

I propose a new test of genetic association in the presence of linkage that uses information in all observed trait values and genomic marker data in a collection of pedigrees. The proposed test can be used for fine mapping under a linkage peak. The test is robust to population admixture and marker model misspecification, and can be applied to both qualitative and quantitative trait values and to general pedigree structures. The test can be viewed as an extension of the well-known TDT test of Spielman, McGinnis and Ewens (1993, AJHG, 52:506–516). The test also shares features with the FBAT test of Rabinowitz and Laird (2000, Human Heredity 50:211–223), but it differs from the FBAT test in the handling of missing founder allelic types and latent IBD sharing status.

189

Intermountain Genealogy Registry a Powerful Pedigree-Based Genetic and Familial Tool

Stacey Knight (1) Tim Maness (2) Sue Dintelman (2) Jeffrey L Anderson (1) Benjamin D Horne (1)
(1) Intermountain Medical Center; Intermountain Heart Institute (2) Pleiades Software Development

The purpose of this project is to describe the creation of the Intermountain Genealogical Registry (IGR). The IGR was created using medical records and a genealogy database. The genealogy database contains information from over 23 million individuals who or whose blood relative had lived in the Intermountain region of the US. This database was constructed from publically available records that were cleaned to eliminate duplicates and record processed to make links between pedigrees. There are ~300,000 pedigrees and 81% of the individuals have ≥4 generations of ancestors. The medical records are from 2.8 million adult patients (pts) treated in the Intermountain Healthcare System. A total of 703,732 (25%) of the pts linked to the genealogy data. Most linked pts (92%) have an average of ≥4 generations of ancestors. The majority of the linked pts are white (88%) and 52.1% are male. The average age of the linked pts at last medical encounter is 53 years. Using the Charlson comorbidity index, a total of 19% of the pts have a diagnosis of chronic pulmonary disease (e.g., asthma, chronic bronchitis, emphysema), 12% of diabetes, 9% of heart failure, 8% of cancer, and 7% of a stroke. A DNA sample, as part of the Intermountain Heart Collaborative Study, is available for 8525 of the linked pts. The IGR is a valuable tool that will help researchers and clinicians better understand the genetic and familial disease risk and identify high risk pedigrees for genetic studies.

190

A Novel Method, the Variant Impact on Linkage Effect Test (VIOLET), Leads to Improved Identification of Causal Variants in Linkage Regions

Lisa J Martin (1) Lili Ding (1) Xue Zhang (1) Michael Olivier (2) D Woodrow Benson (1)
(1) Cincinnati Children's Hospital Medical Center (2) Medical College of Wisconsin

Completion of the Human Genome Project was expected to individualize medicine by rapidly advancing knowledge of common complex disease. However, this has proved challenging. Although linkage analysis has identified highly replicated chromosomal regions, success in identifying causal variants for complex traits has been limited. One explanation for this difficulty is that using association to follow up linkage is problematic given that linkage and association are independent. To overcome this problem, we propose a novel method, the Variant Impact On Linkage Effect Test (VIOLET), which differs from other quantitative methods in that it is designed to follow up linkage by identifying variants influencing the variance explained by a quantitative trait locus. VIOLET's performance was compared to measured genotype and combined linkage association in 2 datasets. Using simulated data, VIOLET had similar power to detect the causal variant compared to standard methods, but reduced false positive rates. Using real data, VIOLET identified a single variant ($p<0.001$), which explained 24% of linkage; this variant exhibited only nominal association ($p=0.04$) using measured genotype and was not identified by combined linkage association. These results demonstrate that VIOLET is highly specific and reduces false positives. In summary, VIOLET overcomes a barrier to gene discovery, and thus may be broadly applicable to identify underlying genetic etiology for traits exhibiting linkage.

191

Fine-Mapping in a Covariate-based Genomewide Linkage Scan of Lung Cancer Susceptibility

Claire L Simpson (1) Tiffany Green (1) Betty Doan (1) Christopher I Amos (2) Susan M Pinney (3) Elena Y Kupert (3) Mariza de Andrade (4) Ping Yang (4) Ann G Schwartz (5) Pam R Fain (6) Adi Gazdar (7) John Minna (7) Jonathan S Wiest (8) Henry Rothschild (9) Diptasri Mandal (9) Ming You (3) Theresa A Coons (10) Colette Gaba (11) Marshall W Anderson (3) Joan E Bailey-Wilson (1)
(1) National Human Genome Research Institute, National Institutes of Health (2) Department of Epidemiology, University of Texas, M.D. Anderson Cancer Center (3) Medical College of Wisconsin (4) Department of Health Sciences Research, Mayo Clinic (5) Karmanos Cancer Institute, Wayne State University (6) University of Colorado (7) University of Texas Southwestern Medical Center (8) National Cancer Institute, National Institutes of Health (9) Louisiana State University Health Sciences Center (10) Saccomanno Research Institute and John McConnell Math & Science Center of Western Colorado (11) University of Toledo

Lung cancer (LC) is a leading cause of death in the developed world, with over 160,000 deaths expected in the US in 2012. Environmental risk factors such as smoking and asbestos exposures are well known. However, only 15% of smokers develop LC, suggesting genetic effects or gene-environment (GxE) interactions.

We previously mapped a major LC susceptibility locus to 6q23-q25, and discovered a rare risk haplotype in linked families that exhibits a GxE interaction between the 6q susceptibility locus and smoking. Genome-wide association studies have suggested other loci with common alleles of small effect on LC risk. However, these loci do not explain all familial risk of LC, suggesting that additional risk alleles exist.

We have also found additional susceptibility loci using linkage analysis including environmental covariates on 6p (LOD=5.75, 74cM) and 6q (LOD=3.25, 173cM), with novel evidence of linkage on 12q24 (LOD=5.46, 150cM) and 22q11 (LOD=5.19,10cM). Linkage to lung and throat cancer was

observed on 9p21 (LOD=4.97, 64cM). All analyses were on microsatellite data.
Here we present the results of a fine-mapping linkage analysis, with data from the microsatellite study combined with a dense SNP map. The data were checked for Mendelian inconsistencies and low call rate and the marker allele frequencies were estimated from the data. Linkage analyses of LC (adjusting for personal smoking) to the combined microsatellite/SNP dataset using LODPAL will be presented.

192
Ascertainment Correction for Model-based Linkage Analysis of Multiplex Families
Xiangqing Sun (1) Robert Elston (1)
(1) Case Western Reserve University

Model-based linkage analysis is more powerful than model-free linkage analysis when a good estimate of the segregation model is used. The only current implementation for combined segregation and linkage analysis is Loki, which has been extended to allow for single ascertainment. However, multiplex families are much more informative for linkage and are usually sampled because they are multiplex, rather than singly ascertained. We therefore investigate by simulation the effect on model-based linkage analysis of assuming an incorrect ascertainment model when obtaining a segregation model from multiplex families. As expected, we find increased type I and type II errors.
We further investigate how the problem can be overcome by a two-step procedure. The first step is to use a prevalence constraint in order to estimate the segregation model that would be obtained from randomly sampled pedigrees. The second step is to fix at the estimates thus obtained all the model parameters other than the allele frequencies, and then estimate new allele frequencies to reflect the trait allele frequencies in the founders of the multiplex families. We show that the type I and type II errors for model-based linkage analysis are then much better controlled.

193
An Adapted MCMC Linkage Analysis Approach – An Approximate Answer to the Right Question?
Kaanan P. Shah (1) Julie A. Douglas (1)
(1) Department of Human Genetics, University of Michigan

Dense arrays of SNPs are routinely genotyped and used to conduct association analyses of complex traits. In the context of family-based studies, these arrays also afford the opportunity to carry out genome-wide linkage analyses. However, exact calculation becomes computationally intractable when dense genetic maps are used with large pedigrees. Here we apply and evaluate the accuracy of Markov chain Monte Carlo (MCMC) methods for large pedigrees with considerable missing data and dense genetic maps ($\sim$1 cM inter-marker spacing). To our knowledge, the performance of MCMC-based methods has not been tested in this setting. Our goal is to analyze dozens of quantitative traits from our genetic study of mammographic density in the Old Order Amish. Although the women from this study (n=1,472) can be connected into a single 14-generation pedigree, we analyze them as a set of 177 trimmed pedigrees ($2n - f \leq 100$). Still, preliminary results suggest substantial variation in LOD scores between MCMC runs, suggestive of poor mixing performance. For example, based on 12 independent MCMC runs, the maximum LOD score for mammographic density ranged from 1.3 to 3 (mean of 1.9). To improve the precision of our LOD score estimates while remaining computationally feasible, we adopt a combined strategy of windowing and averaging across the genome. We evaluate the accuracy of this adapted MCMC approach via simulations with a known trait locus, conditional on our Amish pedigrees and marker data.

194
Obtaining Accurate P-values from a Dense SNP Linkage Scan
William C Stewart (1) Ryan L Subaran (2)
(1) Nationwide Children's Hospital and Ohio State Univeristy (2) Nationwide Children's Hospital

A major concern of resequencing studies is that the pathogenicity of most mutations is difficult to predict. To address this concern, linkage (i.e. co-segregation) is often used to exclude as many mutations as possible, as well as, to better predict pathogenicity among the candidate mutations that remain. However, when linkage disequilibrium (LD) is present in the population but ignored in the analysis, unlinked regions of high LD can provide false evidence for linkage. As a result, the type 1 error of most linkage tests can be inflated, and thousands of neutral mutations may be included mistakenly in a follow-up resequencing study. To illustrate the need for concern, we simulated data on a sparsely spaced panel of SNPs (average spacing 1.27 cM) using an LD pattern estimated from real data. In our simulations, we find that the type 1 error of the maximum LOD can be as high as 14%. Therefore, to control the type 1 error of linkage tests in general, we created Haplodrop–a simulation program that generates the haplotypes of founders with LD and then 'drops' these haplotypes with recombination to all non-founders in the pedigree. Haplodrop accommodates arbitrary pedigree structures and a wide range of LD patterns; and to the extent that comparisons are possible, it agrees well with existing software. Overall, Haplodrop should aid in the identification of pathogenic mutations by correctly excluding thousands of neutral mutations that happen to lie in unlinked regions of high LD.

195
Linkage and Exome Sequencing Analyses Identified the Causative Gene for Dyschromatosis Universalis Hereditaria
Yi Li (1) Hong Liu (2) Furen Zhang (2) Jianjun Liu (1)
(1) Genome Institute of Singapore (2) Shandong Provincial Institute of Dermatology and Venereology

Dyschromatosis universalis hereditaria (DUH) is a pigmentary dermatosis that is most commonly seen in Japanese and Chinese. DUH skin lesions appear within the first month of life and predominantly on the trunk. It shows autosomal dominant or recessive inheritance modes: chromosome 6q24.2-q25.2 with autosomal dominant inheritance in two Chinese families [Xing et al. 2003] and chromosome 12q21-q23 with autosomal recessive inheritance in an Arabian consanguineous family [Stuhrmann et al. 2008] were reported to be implicated in DUH. However, specific causative genes have not been reported so far. Next generation sequencing

(NGS) technology has changed the way we go about basic, applied and clinical researches in biology. In this study, we applied linkage analysis and exome sequencing technology to analyze a pedigree with DUH. One SNP was identified to be putative causative by a principled procedure, which was verified by Sanger sequencing to be present in all patients and absent in all healthy controls of the family. This SNP is also absent in 1000 ethnicity-matched population controls. Sanger sequencing of the same gene in another DUH family showed that another SNP in the same domain of the same gene is present in all patients of the second family, and absent in 1120 ethnicity-matched population controls. The functional annotations of the identified gene in skin pigmentation will also be discussed.

196

Improving Power of Meta-analyes by Using Sex-mismatch Correction in Genome wide Association Studies

Stefan Boehringer (1) Jeroen J Pijpe (1)
(1) Leiden University Medical Center

Genome wide genetic data offer the possibility to estimate genetic sex. Comparing this sex to reported sex will in most instances indicate sample mixup on disagreement. Such mixups can be interpreted as measurement error of genetic data with the special error mechanism of permutations of data points. We develop formulas for estimating error rates in balanced as well as unbalanced samples with respect to male/female ratios and/or case/control ratios. We examine the exact likelihood modeling all possible permutations given the estimated error frequencies that is approximated by stochastic integration and show that unbiased estimates for effect sizes can be obtained. Using calibration ideas from measurement error theory, we also develop approximate, closed-form formulas. We point out that further knowledge about the mix up-mechanism is needed, if mixup cannot be considered to be completely at random, which can be incorporated into the models.
We demonstrate the methods in an African data set with a high sex-mismatch rate of approx. 10%. In such a case odds ratios can be underestimated by as much as 10%. Meta-analyses can be improved by using sex-mismatch corrected effect-size estimates. We show by simulations that power can be increased especially in fixed-effects models when mismatch rates are heterogeneous across studies. Finally, biological interpretation benefits from accurate effect sizes, especially when comparisons across studies or species are made.

197

A Meta-analysis for Identifying Pleiotropic Loci Influencing Adiposity and Cardiometabolic Traits

Ingrid B Borecki (1) Michael A Province (1) Aldi T Kraja (1)
(1) Washington University School of Medicine

We sought to investigate traits associated with Metabolic Syndrome to explore the genetic basis of their correlation, and identify pleiotropic and novel trait loci. We conducted a genomewide search using a correlated meta-analysis of meta-results for 8 traits in 4 domains: BMI and waist circumference from GIANT (N∼123,000), HDL and triglycerides (TG) from GLGC (N∼99,000), fasting glucose and insulin from MAGIC (N∼38–46,000), and systolic and diastolic blood pressure from GBPG (N∼34,000). Analysis of all possible combinations of variables was carried out revealing loci with P<5e-08; the association evidence across phenotypes must be reinforcing, with a minimum P<e-03 for any single trait. Novel loci as well as loci previously identified as having a marginal effect on one of the phenotypes were found to have pleiotropic effects. These mapped to GO terms indicating roles within the cell, membrane and binding proteins, transcription factors and metal ion binding. Metabolic and signaling pathways were represented as expected, but also inflammatory pathways. Many of the loci mapped to multiple pathways, consistent with their presumed pleiotropic effect. There was also evidence of protein interactions among many of the identified genes, suggesting another mechanism of pleiotropic effect. Such bioinformatic investigation using large, informative GWAS datasets can provide valuable clues to the genetic architecture of correlated traits.

198

Meta-analysis of Genome-wide Association Studies (GWAS) has Become a Useful Tool to Identify Genetic Variants that are Associated with Complex Human Diseases. To Control Spurious Associations Between Genetic Variants and Disease that are Caused by Pop

Shudong Wang (1) Wenan Chen (1) Guimin Gao (1)
(1) Virginia Commonwealth University

Meta-analysis of genome-wide association studies has become a useful tool to identify genetic variants that are associated with complex human diseases. To control spurious associations between genetic variants and disease that are caused by population stratification, double genomic control (GC) correction for stratification in meta-analysis have been implemented in the software GWAMA and widely used by investigators. In this research, we conducted extensive simulation studies to evaluate the double GC correction method in meta-analysis and compared the performance of the double GC correction with the single GC correction and principal components analysis (PCA) correction in meta-analysis. Results show that using double GC correction and using single GC correction in meta-analysis generated similar type I error rates and power. The single and double GC correction methods generated inflated type I error rates in meta-analysis when the data consist of subpopulations with different allele frequencies. On the other hand, the PCA correction method could control type I error rates well and generated much higher power in meta-analysis compared to both the single and double GC correction methods. We applied the above three correction methods to meta-analysis of three real data sets. The results also suggest that PCA correction is more effective than both the single GC and double GC correction in meta-analysis.

199

Large-scale Genome-wide Association Meta-analysis Using Imputation from 2188-haplotype 1000 Genomes Reference Panel Identifies Novel Susceptibility Loci for Anthropometric and Glycemic Traits

Reedik Magi (1) Momoko Horikoshi (2) Ida Surakka (3) Steven Wiltshire (2) Antti-Pekka Sarin (3) Anubha Mahajan (4) Letizia Marullo (5) Teresa Ferreira (4) Sara Hagg (6) Janina S Ried (7) Thomas Winkler (8) Gudmar Thorleifsson (9) Natalia Tsernikova (1) Tonu Esko (1) Christina Willenborg

(10) Christopher P Nelson (11) Marian Beekman (12) Sara M Willems (13) Mark I McCarthy (2) Andrew P Morris (4) Cecilia M Lindgren (4) Samuli Ripatti (3) Inga Prokopenko (2) for ENGAGE Consortium (14)
(1) Estonian Genome Center, University of Tartu (2) WTCHG, University of Oxford, Oxford, UK; OCDEM, University of Oxford, Oxford, UK (3) FIMM, University of Helsinki, Helsinki, Finland; National Institute for Health and Welfare, Helsinki, Finland (4) WTCHG, University of Oxford, Oxford, UK (5) Department of Evolutionary Biology, University of Ferrara, Ferrara, Italy; WTCHG, University of Oxford, Oxford, UK (6) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden (7) Helmholtz Zentrum Munchen – German Research Center for Environmental Health, Neuherberg, Germany (8) Institute of Epidemiology and Preventive Medicine, Regensburg University Medical Center, Regensburg, Germany (9) deCODE Genetics, Reykjavik, Iceland (10) G Kardiovaskulare Genomik, Medizinische Klinik II, Universitat zu Lubeck, Lubeck, Germany (11) Department of Cardiovascular Sciences, University of Leicester, Leicester, UK (12) Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands (13) Department of Genetic Epidemiology, Erasmus MC, Rotterdam, The Netherlands (14) ENGAGE Consortium

Genome-wide association studies (GWAS) were successful in detecting numerous associations with quantitative glycemic and obesity traits. In order to find more loci and to approach potential causal variants by fine-mapping in already known loci, we performed imputation using 1000 Genomes panel with 2188 haplotypes (June 2011 release) in up to 18 European cohorts with GWA data within the ENGAGE consortium, followed by fixed-effects inverse variance meta-analysis for BMI (n=78,131), waist-to-hip ratio (WHR, n=44,301), fasting glucose (FG, n=38,326) and fasting insulin (n=16,831) traits. Analysis of these four phenotypes showed multiple genome-wide significant (p-value$<5\times10^{-8}$) loci previously unreported for BMI and FG. For BMI, we observed associations at *GALNT10* (minor allele frequency (MAF)=0.42, p=7.8×10^{-10}) linked to oligosaccharide biosynthesis; *ZHX2* (MAF=0.16, p=4.5×10^{-8}), zinc fingers and homeoboxes 2 gene; *PAX2* (MAF=0.19, p=2.0×10^{-8}), a tumor suppressor gene; and at *AKAP6* (MAF=0.49, p=7.7×10^{-9}). For FG, *RMST* (MAF=0.10, p=2.0×10^{-10}) locus was significantly associated. We found low MAF variants within common variant loci: at *BDNF* (p=9.0×10^{-4}, MAF=0.018) for BMI and at *G6PC2* (p=9.4×10^{-17}, MAF=0.011) for FG. Our results highlight the potential for the identification of associations using existing GWAS genotyping data, supplemented with imputation from high-density reference panel of 1000 Genomes project.

200

Estimating Phenotypic Variance Explained by Genetic Factors in Meta-Analysis Studies of Quantitative Traits and Linear Regression Models

Serkalem Demissie (1) L Adrienne Cupples (1)
(1) Boston University School of Public Health

Meta-analysis (MA) has become an increasingly valued strategy for detection of small genetic effects in single nucleotide polymorphisms (SNP) and complex trait association studies. It is well recognized that the types of analyses that can be performed in MA studies are generally limited. However, *it is not well known* that estimation of a variety of parameters, including the *proportion of trait variance explained* by a SNP ($R^2_{(SNP)}$), is in fact possible to do in both inverse-variance and weighted Z-score MA studies. In this study, we conducted a brief review of literature on a simple formula for $R^2_{(SNP)}$ computation and, using theoretical and simulation analyses, illustrated that estimates from this formula (calculated based only on a test-statistic and sample size from MA) are essentially identical to those obtained directly from individual-level data analysis. We also demonstrated that the formula can be used more generally as a simple alternative for squared partial-correlation coefficient estimation in a linear regression analysis. $R^2_{(SNP)}$, which provides an estimate for a quantitative trait locus-heritability, is an important effect measure deeply rooted and extensively used in genetic studies. We thus recommend adding $R^2_{(SNP)}$ to MA software programs to encourage reporting $R^2_{(SNP)}$ in genetic meta-analysis studies as routinely done in studies that use individual-level data.

Withdrawn abstract 201

202

Integrating Multiple Glycans and Genetic Data Using Joint Modeling Techniques: An Application in Leiden Longevity Study

Hae-Won Uh (1) Roula Tsonaka (1) Manfred Wuhrer (2) Eline Slagboom (3) Jeanine Houwing-Duistermaat (4)
(1) Dep. Medical statistics and Bioinformatics, Leiden University Medical Center (2) Dep. Parasitology, Leiden University Medical Center (3) Dep. Molecular Epidemiology, Leiden University Medical Center (4) Dep. Medical statistics and Bioinformatics, Leiden University Medical Cente

Identification of markers that reflect the biological age of individuals is of paramount importance in aging research. In the Leiden Longevity Study (LLS) several biomarkers and genetic data have been collected on 1671 offspring of nonagenarian sibling pairs and their 744 partners, who are treated as controls. From the recorded biomarkers, we focus on IgG glycosylation, and investigate their joint association with familial longevity. So far association testing between glycosylation and healthy aging is done separately for each glycan. However, joint analysis of multiple glycans and genetic data could enhance our understanding of the biological mechanisms and improve predictions. Therefore we have developed a novel joint modelling framework to test associations between multiple omics data, genetic data and healthy aging. In particular, we use hierarchical random-effects models to model jointly the multiple glycosylation data and build associations with gene-sets and phenotypes. The advantages of this approach are that it can accommodate familial relationships, the sampling design and the between glycans and SNPs correlations.
By applying the novel method for multiple glycans we found the significantly associated SNPs located in two genes using 300K genotyped data, whereas no SNP showed a genome-wide significant association with each glycan. The same two genes were identified, when GWAS was performed for each glycan separately using ca. 2.5 million imputed SNPs.