IGES 2013 Abstracts

1

SMART-scan (Selection of Models for the Analysis of Risk-factor Trees): Leveraging biological knowledge to mine large sets of risk factors with application to microbiome data

Qunyuan Zhang (1) Felicia Gomez (1) Nita H Salzman (2) Jeffrey Schwimmer (3) Alan Templeton (4) Michael A Province (1) Ingrid I Borecki (1)

(1) Division of Statistical Genomics, Washington University School of Medicine
(2) Children's Research Institute, Medical College of Wisconsin
(3) Department of Pediatrics, University of California, San Diego
(4) Department of Biology, Washington University in St. Louis

The association between a microbiome signature and clinical outcomes is of growing interest because of the potential for yielding insights into mechanism and pathogenesis. Bacterial 16S RNA variable regions can be used to assign taxonomic membership, and relative abundances can be measured in a human sample, but it is usually not known which of these features are relevant for a disease. We propose a novel tree scanning method, SMARTscan, for identifying relevant branches that are associated with a disease or trait. SMARTscan is a model selection technique that can operate on any basic generalized linear model (applicable to quantitative, qualitative and survival/onset phenotypes), with/without covariates, complex interactions and/or random effects. Using a pre-defined taxonomy to organize the large pool of possible predictors, SMARTscan focuses on detecting associated evolutionary divergent events and exhaustively searches the tree for those units that best predict outcome. Permutation tests provide empirical p-values and adjust for the nested nature of the multiple comparisons. Further refinement is achieved via subsequent conditional splits of the tree. We investigate the statistical properties (type-1 error rate and power) through simulations and compare with other distance- and variable-selection-based methods. Our method thus incorporates more information by leveraging knowledge regarding the relationship among predictors and using a hierarchical approach, resulting in an improved understanding of complex predictive structures.

2

Polygenic Risk Score Associations may be improved with simple procedures

Kelly S Benke (1) Shan Andrews (1) M Daniele Fallin (1) Brion S Maher (1)

(1) Johns Hopkins Bloomberg School of Public Health

Polygenic risk scores are useful genetic tools, but their utility and interpretability are likely limited by heterogeneity in both local and global ancestry. It has become standard practice to adjust for global ancestry via principle components or similar ancestry summary statistics when assessing associations between risk scores and disease. Little work, however, has been dedicated to the residual influence of local ancestry. Through simulation we show that in situations of admixture, adjustment for global ancestry in genetic risk score analysis does not fully correct for confounding by ancestry and may also limit interpretation of the score itself. We further present and compare strategies for addressing this residual confounding by other methods of global ancestry adjustment and using estimates of local ancestry as adjustment factors in the score calculation.

3

Geographic genetic diversity in the United States and implications for genomewide association studies

Ronnie A Sebro (1) Nan M Laird (2) Neil J Risch (1)

(1) University of California, San Francisco
(2) Harvard School of Public Health

The United States of America (USA) is a multiethnic and multiracial population shaped by immigration and admixture. Previous studies have shown that self-reported ancestry is an excellent proxy for genetic ancestry. Here we utilize self-reported ancestry data from the diennial United States Census 2010 and utilize published allele frequencies from 45 genes available for 93 common ancestries. We use the relative proportion of individuals of each ancestry within each country and the allele frequency data to reconstruct the genetic demography of the USA, by generating maps showing the geographic distribution of the mean calculated Wright's FST. We use Shannon's diversity index (H) and show that most of the genetic variation was within ancestry; that there is population substructure in the USA and that this substructure varies significantly by geographic locale even between adjacent counties in the same state. We show that the coastal regions and South are more genetically diverse than the Midwest. When all races/ethnicities were considered 79.5% of the genetic variation was between ancestries within a county, 6.5% between counties within a state, 4.6% between states within a region and 9.4% between regions. 95% of the White genetic variation was between ancestries in the same county, compared to only 43.1% and 67.8% for the Hispanic and Native Hawaiian/Pacific Islander (NHPI) populations respectively, illustrating that the Hispanic and NHPI ancestries were more geographically clustered relative to Whites. Multi-center genome-wide association studies performed in the USA and meta-analyses are susceptible to population substructure and require both within center and between center corrections for population substructure.

IGES 2013 Abstracts

4

Investigating genetic and epigenetic variation in the chromosome 2q region linked to tissue factor pathway inhibitor plasma levels

Jessica Dennis (1) Alejandra Medina (2) Mathieu Lemire (3) Dylan Aïssï (4) Phil Wells (5)
Pierre Morange (6) Michael Wilson (2) David Trégouët (4) France Gagnon (1)

(1) Dalla Lana School of Public Health, University of Toronto
(2) Genetics and Genome Biology, University of Toronto
(3) Ontario Institute of Cancer Research
(4) UMRS 937
(5) Ottawa Hospital Research Institute
(6) University of the Mediteranean

BACKGROUND: Tissue factor pathway inhibitor (TFPI) is a key regulator of coagulation. Low levels of this protein promote fibrin clot formation, increasing the risk of venous thromboembolism (VTE), myocardial infarction, and ischemic stroke. TFPI plasma levels are highly heritable, yet the mechanisms underlying this heritability are poorly understood. In 2005, a genome-wide scan of 21 Spanish families found a significant linkage signal for TFPI plasma levels in the 2q region surrounding the TFPI gene, but follow-up analyses excluded known TFPI polymorphisms. This finding has yet to be followed up in an independent study. The objective of our study is to investigate genetic and epigenetic associations with TFPI plasma levels in this region. METHODS: In 254 individuals from 5 large French-Canadian families ascertained on VTE, using the variance component method in SOLAR, we tested for association with TFPI plasma levels 9759 SNPs and 6371 DNA methylation probes (measured in peripheral blood) in the 63 Mbp region spanning the 2q linkage interval. RESULTS: The most significant SNP (p=2.15e-06) lies in an enhancer 33 Mbp upstream of the TFPI gene. The second ranked SNP (p=1.15e-04) is within a gene 5 Mbp from TFPI, while the top-ranked DNA methylation probe (p=8.30e-05) lies in the transcription start site of this same gene that is 5 Mbp from TFPI. Neither SNPs nor methylation probes in the TFPI gene itself were statistically significantly associated with TFPI plasma levels. NEXT STEPS: Results are preliminary and are being replicated in an independent study sample. Integrating genetic and epigenetic information may help to identify novel genes associated with TFPI plasma levels.

5

Optimal Selection of Individuals for Genotyping in Genetic Association Studies with Related Individuals

Miaoyan Wang (1) Johanna Jakobsdottir (2) Albert V. Smith (3) Vilmundur Gudnason (3) Mary Sara McPeek (1)

(1) University of Chicago
(2) The Icelandic Heart Association
(3) The Icelandic Heart Association and The University of Iceland

Suppose a sample of phenotyped individuals is available, with some subset of them possibly already genotyped, and with others not genotyped. We consider the problem of choosing an additional subset of n individuals to genotype, so as to maximize power to detect association. The problem arises naturally in studies in which the genotyping budget is limited. We consider samples that include at least some related individuals, with the kinship assumed known. In this context, due to the dependence of genotypes and phenotypes among related individuals, power can be gained by including in the association analysis partial information, such as phenotype data on ungenotyped relatives. It is important to take this into account in assessing which individuals to genotype. We propose G-STRATEGY, an optimal method for selecting additional individuals to genotype from a sample that includes relatives, provided that the phenotypes and kinship are known. G-STRATEGY uses simulated annealing to maximize the noncentrality parameter based on the MQLS and MASTOR methods, both of which increase power in this context by incorporating phenotype information on ungenotyped relatives. In simulations, G-STRATEGY performs well for a range of complex disease models, providing a substantial power increase compared to strategies such as genotyping only founders or selection based on extreme phenotype enrichment. G-STRATEGY is computationally feasible even for large complex pedigrees. We apply G-SRATEGY to data on HDL from the AGES-Reykjavik and REFINE-Reykjavik studies; the data include over 7200 phenotyped individuals, with 3219 already genotyped..

6

Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches

Jae Hoon Sul (1) Buhm Han (2) Chun Ye (3) Ted Choi (4) Eleazar Eskin (1)

(1) Computer Science Department, University of California, Los Angeles, California, USA
(2) Division of Genetics, Brigham & Womenâ€™s Hospital, Harvard Medical School, Boston, MA, USA
(3) Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
(4) Predictive Biology, Inc., San Diego, California, USA

Gene expression data, in conjunction with information on genetic variants, have enabled studies to identify expression quantitative trait loci (eQTLs) or polymorphic locations in the genome that are associated with expression levels. Moreover, recent technological developments and cost decreases have further enabled studies to collect expression data in multiple tissues. One advantage of multiple tissue datasets is that studies can combine results from different tissues to identify eQTLs

more accurately than examining each tissue separately. The idea of aggregating results of multiple tissues is closely related to the idea of meta-analysis which aggregates results of multiple genome-wide association studies to improve the power to detect associations. In principle, meta-analysis methods can be used to combine results from multiple tissues. However, eQTLs may have effects in only a single tissue, in all tissues, or in a subset of tissues with possibly different effect sizes. This heterogeneity in terms of effects across multiple tissues presents a key challenge to detect eQTLs. In this paper, we develop a framework that leverages two popular meta-analysis methods that address effect size heterogeneity to detect eQTLs across multiple tissues. We show by using simulations and multiple tissue data from mouse that our approach detects many eQTLs undetected by traditional eQTL methods. Additionally, our method provides an interpretation framework that accurately predicts whether an eQTL has an effect in a particular tissue.

7
Accounting for cellular heterogeneity is critical in epigenome-wide association studies

Andrew E Jaffe (1) Rafael A Irizarry (2)

(1) Libeber Institute for Brain Development
(2) Johns Hopkins Bloomberg School of Public Health

Epigenome-wide association studies (EWAS) of human disease and other quantitative traits are becoming increasingly common. A series of papers reporting age-related changes in DNA methylation (DNAm) profiles in peripheral blood have already been published. However, blood is a heterogeneous collection of different cell types, each with a very different DNA methylation profile. Using a statistical method that permits estimating the relative proportion of cell type components from DNAm profiles, we examine data from five previously published, and find strong evidence of cell composition changes across age. We also demonstrate that, in these studies, cellular composition explains much of the observed variability in DNAm. Furthermore, we find high levels of confounding between age-related variability and cell composition at the CpG level. Our findings underscore the importance of accounting for cell composition variability in epigenetic studies based on whole blood and other heterogeneous tissue sources..

8
Reconstructing Pedigrees from Estimates of Genomic Sharing in Admixed Populations

Jennifer E Below (1) Jeffrey Staples (2) Alexander Reiner (2) Lynette Ekunwe (3) Ermeg L Akylbekova (4) Solomon K Musani (4) James G Wilson (4) Craig R Hanis (1) Deborah Nickerson (2)

(1) University of Texas, Health Science Center
(2) University of Washington
(3) Jackson State University
(4) University of Mississippi Medical Center

Understanding and correctly utilizing relatedness among samples is essential for all genetic analysis. However, records of sample relatedness are often incorrect, incomplete, or unavailable. PRIMUS is an algorithm that utilizes genome-wide estimates of identity by descent (IBD) to assign relationship categories and leverages these pairwise relationships to identify all possible pedigrees consistent with the observed genetic sharing. Reconstructing pedigrees in admixed populations is complicated by the fact that many algorithms for estimating IBD from genome-wide SNP data assume sampling from a single homogeneous ancestral population. The presence of ancestry informative markers (AIMs) in estimating IBD using a method of moments can strongly bias relationship classification, resulting in inaccurate or failed pedigree reconstruction. We have implemented a principal component analysis within PRIMUS to identify signatures of admixture, appropriate reference population minor allele frequencies, and identify and remove AIMs prior to IBD estimation. Controlling for multiple ancestral groups, we reconstructed pedigrees for all 1,985 Mexican Americans from the Starr County Health Study (SC) as well as 3030 African Americans from the Jackson Heart Study (JHS). PRIMUS unambiguously identified 197 previously undescribed pedigrees in SC and reconstructed pedigrees for 338 families within JHS. Our method provides accurate reconstruction in genetically heterogeneous samples. We present the resulting pedigrees, and show that PRIMUS is powerful for both identifying novel pedigrees in large admixed genetic cohorts and for validating known pedigrees.

9
Design matters! A statistical framework to guide sequencing choices in pedigrees

Charles Yin Kiu Cheung (1) Ellen M Wijsman (1)

(1) University of Washington

The use of large pedigrees is a promising design to identify rare functional variants in heritable traits. Cost-effective studies using sequence data are achieved by using pedigree-based imputation in which some subjects are sequenced and missing genotypes are inferred. As these large-scale studies are very expensive, we must carefully prioritize who to sequence. Here, we introduce a novel statistical framework based on a coverage measure that reflects the fraction of alleles that can be accurately imputed in a pedigree. This framework enables comparison among designs prior to sequencing. Using this framework, we present a method to select subjects that can use inferred inheritance vectors to optimize design choices when candidate regions are known. This method also allows a-priori

preference for selecting subjects to account for realistic situations in which users may have a few subjects that they want to first select before deciding the remaining subjects to sequence. On a 52-member simulated pedigree with 46 selectable subjects, we compared our method with Primus, ExomePicks, and 100 replicates of random selection of subjects (RSS). For each method, we selected 10 subjects and imputed missing genotypes in a 4 cM region with 1000 markers using GIGI for pedigree-based imputation. Our method substantially outperformed Primus, ExomePicks, and median of RSS in imputation accuracy (88% vs. 78%, 83%, and 84%). Comparing accuracy with those from the distribution of RSS, our method yielded higher accuracy than all replicates (vs. Primus=0 and ExomePicks=39). This framework and program facilitate improved and informed sequencing decisions.

10
Poly-Omic Prediction of Complex Traits – OmicKriging

Heather E Wheeler (1) Keston Aquino-Michaels (1) Eric R Gamazon (1) Vasya Trubetskoy (1) M Eileen Dolan (1) R Steph Huang (1) Nancy J Cox (1) Hae Kyung Im (1)

(1) The University of Chicago

High-confidence prediction of complex traits such as disease risk or drug response is an ultimate goal of personalized medicine. Although genome-wide association studies have discovered thousands of well-replicated polymorphisms associated with a broad spectrum of complex traits, the combined predictive power of these associations for any given trait is generally too low to be of clinical relevance. We propose a novel systems approach to complex trait prediction, which leverages similarity in genetic, transcriptomic or other omics-level data. We translate the omic similarity into phenotypic similarity using a method called Kriging, commonly used in geostatistics and machine learning. Our method called OmicKriging emphasizes the use of a wide variety of systems-level data, such as those increasingly made available by comprehensive surveys of the genome, transcriptome and epigenome, for complex trait prediction. The approach facilitates exploration of the etiology of disease risk or drug response by quantifying specific omic contributions to prediction. Our method is a fast, simple and flexible approach to polygenic, and more generally, poly-omic, prediction. Using seven diseases from the Wellcome Trust Case Control Consortium (WTCCC), we show that our method yields performance similar to more computationally intensive methods when restricted to genotypic data. Using a cellular growth phenotype, we show that integrating mRNA and microRNA data with genotypic data substantially increases performance.

11
Genome-wide scan of inversions predisposing to secondary rearrangements using case-parent trio data

Jianzhong Ma (1) Christopher I Amos (2)

(1) Department of Genetics, University of Texas MD Anderson Cancer Center
(2) Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College

Although chromosomal inversions are usually believed to be phenotypically neutral, unless the breakpoints disrupt a putative gene or its corresponding regulatory element, increasing evidence has emerged that inversions carried by parents may cause a secondary rearrangement in the germline of their children that in turn is susceptible to diseases. Recent advances in mapping structural variants have shown that inversions may be much more abundant than previously expected, implying that inversions may significantly contribute to the missing inheritability that the genome-wide association studies failed to identify. Based on our recently developed approach to detecting and genotyping inversions using principal components analysis (PCA), we propose a method of genome-wide analysis of inversions susceptible to a detrimental rearrangement using case-parent trios. Using single nucleotide polymorphisms (SNPs) data, the inversion status of individuals can be inferred either from local PCA when there are sufficient markers inside the inversion region or from the genotype of the surrogate SNPs identified from HapMap data. For each of the known inversions and those predicted by our PCA-based approach, we compare the rate of inversions in parents of the cases to that in the general population. We are specifically interested in the rate of inversion heterozygote in the parents, because the presence of an inverted and a non-inverted segment originates an unpaired region at the pachytene stage, which may result in a misalignment and non-allelic homologous recombination. We illustrate our method using the Genetic Analysis Workshop data.

12
Sequence Data in Family Studies: the Framingham Heart Study

Adrienne Cupples (1) Ching-Ti Liu (2) Ming-Huei Chen (3) Qiong Yang (1) Kathryn L. Lunetta (1) Josée Dupuis (1)

(1) Department of Biostatistics, Boston University School of Public Health, Boston, MA.National Heart, Lung, and Blood Institute (NHLBI) Framingham Heart Study, Framingham, MA
(2) Department of Biostatistics, Boston University School of Public Health, Boston, MA
(3) Department of Neurology, Boston University School of Medicine, Boston

Genome-wide association studies have yielded many novel common genetic variants associated with a wide variety of complex traits. Recently, we have seen the rise of sequencing studies to delineate the genetic contributions of rare variants to complex traits. Most such studies have used samples of unrelated individuals for these efforts. Yet, when a rare variant is observed, family studies may provide more copies of shared rare variants than studies with unrelated individuals. Further, family studies will provide formal examination of transmission of genetic variants associated with traits. The Framingham Heart Study has participated in the CHARGE Sequencing Study for targeted, exome and whole genome sequencing. We will present a description of selection of individuals for sequencing, quality control issues in dealing with sequence data, imputation of rare variants in families and strategies for analysis of such data. Specifically, we will present approaches to conduct analyses of genomic regions using burden tests that aggregate data over a genomic region or use model selection. Finally, we will describe the advantages and disadvantages of using family studies versus samples of unrelated individuals in evaluating rare variants. These issues will be highlighted in the context of evaluating the contribution of rare variants to body mass index.

13
Sequence Kernel Association Test in Family Samples with Repeated Phenotype Measurements or Multiple Traits

Wei Gao (1) George T O'Connor (2) Josée Dupuis (2)

(1) Department of Biostatistics, Boston University School of Public Health, Boston, MA
(2) The NHLBI Framingham Heart Study, Framingham, MA; Pulmonary Center, Department of Medicine, Boston University School of Medicine, Boston MA.
(3) Department of Biostatistics, Boston University School of Public Health, Boston, MA; The NHLBI Framingham Heart Study, Framingham, MA

There are many statistical methods for rare variant association analysis. Among them, the sequence kernel association test (SKAT) has been shown to be a powerful method in various scenarios. It is a score-based variance component test that allows rare variants with different direction of effects, and is computationally efficient because it does not require permutation to evaluate P-values. The family-based SKAT (famSKAT) is an extension of SKAT to handle family data. In this paper, we developed a general SKAT framework that is applicable to family sample, repeated phenotype measurements and multiple traits. The general SKAT (genSKAT) takes correlation between multiple measurements into consideration in the test statistic formulation. When each participant has only one measurement, genSKAT is equivalent to SKAT when there is no familial correlation and to famSKAT in the presence of familial correlation. Our simulations show that SKAT and famSKAT have inflated type I error if correlations between repeated phenotype measurements or multiple traits are inappropriately ignored. In contrast, genSKAT has the correct type I error. We illustrate our approach to evaluate the association of rare genetic variants using pulmonary function traits from the Framingham Heart Study.

14
Sharing of rare variants by affected relatives: building evidence for causal variants based on exact sharing probabilities

Alexandre Bureau (1) Ingo Ruczinski (2) Margaret M Parker (2) Margaret Taub (2) Mary L Marazita (3) Jeffrey C Murray (4) Joan E Bailey-Wilson (5) Cheryl D Cropp (5) Alan F Scott (6) Terri H Beaty (2)

(1) Department of Social and Preventive Medicine, Université Laval, Canada
(2) Johns Hopkins Bloomberg School of Public Health, USA
(3) School of Dental Medicine, University of Pittsburgh, USA
(4) Department of Pediatrics, University of Iowa, USA
(5) Inherited Disease Research Branch National Human Genome Research Institute National Institutes of Health, USA
(6) School of Medicine, Johns Hopkins University, USA

Family based study designs are regaining popularity because large scale sequencing can help to interrogate the relationship between disease and variants too rare in the population to be detected through tests of association in a conventional case-control study, but may nonetheless co-segregate with disease within families. Where only a few affected subjects per family are sequenced, evidence that a rare variant may be causal can be quantified from the probability any variant would be shared by all affected relatives given it was seen in any one family member under the null hypothesis of complete absence of linkage and association. For variants seen in M families and shared by affected relatives in m of them, a p-value can be obtained as the sum of the probabilities of sharing events as (or more) extreme. We generalized the expression for the sharing probability to more than two subjects per family. We also examined the impact of unknown relationships and proposed approximation of sharing probability based on empirical estimates of kinship between family members obtained from genome-wide marker data. We applied this method to a study of 55 multiplex families with apparent non-syndromic forms of oral clefts from four distinct populations. Whole exome sequencing was performed by the Center for Inherited Disease Research (CIDR) on two or three affected members per family. The rare single nucleotide variant rs149253049 in the gene ADAMTS9 was shared by affected relatives in three Indian families (p=2x10-6), illustrating the power of this sharing approach.

15
Design of Sequence-based Follow-up to GWAS

Paul Marjoram (1) Duncan Thomas (1) David Conti (1) Matthew Salomon (1)

(1) U.S.C.

GWAS aims to find SNPs associated with phenotypes of interest. Since detected SNPs are likely to be in linkage disequilibrium with the truly causal SNP, rather than being causal themselves, a common follow-up study will be to sequence an area around one or more regions containing such SNPs. Multiple non-trivial questions arise: What regions should we chose? How wide should those regions be? How deep should the coverage be? How do we best analyze the resulting data? We will describe our efforts as part of NIH's "GWASeq" consortium, which aims to provide the empirical data that will allow the community to answer these questions. Our component of the consortium focuses on data from the Colon-Cancer Family Registry (C-CFR). We have sequenced around 4000 samples from the C-CFR, across 10 regions, at in excess of 50X coverage. Our data is a mixture of population-based and pedigree-based samples. We will describe the study, the data that results from it, and the conclusions that can be drawn regarding design of such studies in future and the implications regarding so-called "missing heritability".

16
Variants Affecting Exon Skipping in Very Important Pharmacogenes

Younghee Lee (1) Eric R Gamazon (1) Nancy J Cox (1) Daniel Levy (1)

(1) University of Chicago
(2) Boston University

15-50% of all human heritable diseases have been estimated to arise from variants that are implicated in alternative splicing (AS) machinery such as canonical splice sites or splicing regulatory elements. Variants that affect AS contribute to lung adenoma prognosis, prostate cancer risk, and retinoblastoma. There is clear value to exploring impact of genetic variations affecting splicing on pharmacogenomic phenotypes. To investigate the pharmacogenetic relevance of AS, we focus on an expanded list of Very Important Pharmacogenes (VIP). With this subset of VIP genes, we test for statistically significant correlations between SNPs and level of exon skipping by utilizing exon-level expression microarrays of 176 HapMap lymphoblastoid cell lines (LCLs) or RNA-seq data in GTEx. We are able to ascribe function to a number of variants in VIP genes that showed phenotype associations but for which mechanism for the association was unknown. For example, one of our findings in VIP genes, TNFRSF1A, intronic rs1800693 effectively neutralizes an intronic splicing enhancer and is also associated with a moderate decrease in exon expression (linear regression, $p< 0.000723$, $R2> 0.08$, minor allele frequency of CEU=0.48). Furthermore, the skipped exon is one of the exons encoding a domain of the TNFR/NGFR cysteine-rich region and rs1800693 SNP has already been associated with multiple sclerosis and primary biliary cirrhosis. We will summarize results for these studies in VIP genes in a variety of tissues and provide information on a database serving these results.

17
Imputation without doing imputation: a new method for the detection of non-genotyped causal variants

Richard Howey (1) Heather J. Cordell (1)

(1) Newcastle University

Genome-wide association studies (GWAS) allow the detection of non-genotyped disease causing variants through the testing of nearby genotyped SNPs that are in strong linkage disequilibrium (LD). Several genotyped SNPs in weak LD with the causal variant may provide equivalent information; imputation-based methods aim to exploit this but are complex and computationally intensive. Here we present a new straightforward GWAS method (and accompanying software package, SnipSnip) designed for this scenario. 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 SNP). These two SNPs are then used as predictors in a linear or logistic regression analysis, in order to generate a final significance test associated with the anchor SNP. As a demonstration of our method using data in which LD between causal and genotyped SNPs is weak, we considered a case/control data set of severe malaria in The Gambia genotyped using the Affymetrix 500K GeneChip. A previous analysis showed that fine-scale sequencing of a Gambian reference panel in the region of the HbS locus followed by multipoint imputation increases the signal of association to genome-wide significance levels. We show that our method can also increase the signal of association from $P\approx10\text{-}6$ to $P\approx10\text{-}10$. Our method thus, in some cases, potentially eliminates the need for more complex methods such as sequencing and imputation or haplotype analysis, and provides a useful additional test that may be used with existing GWAS data to identify genetic regions of interest.

18
Detecting genetic heterogeneity in complex diseases with a weighted U statistic

Changshuai Wei (1) Robert C Elston (2) Qing Lu (1)

(1) Department of Epidemiology and Biostatistics, College of Human Medicine, Michigan State University
(2) Department of Epidemiology and Biostatistics, Case Western Reserve University

For most complex diseases, a large proportion of the genetic variants remain undiscovered. While current research interests have shifted toward uncovering rare variants, gene-gene/gene-environment interactions, and structural variations, the impact of genetic heterogeneity in human diseases has been largely overlooked. Converging evidence suggests that diseases with the same or similar clinical manifestations could have different underlying genetic etiologies. Most of the existing analytical approaches assume the disease under investigation has a homogeneous genetic cause and could, therefore, have low power if the disease undergoes heterogeneous biological pathways. In this paper, we propose a statistical approach, a heterogeneity weighted U (HWU) approach, for high-dimensional association analysis taking genetic heterogeneity into account. HWU can be applied to various types of traits (e.g., binary and continuous), and is designed for detecting heterogeneous genetic effects. Through simulations, we compared HWU with a non-heterogeneity weighted U (NHWU) and the conventional generalized linear model (GLM). The results showed that HWU has substantial gain in power compared to NHWU and GLM in the presence of genetic heterogeneity, while retaining a performance similar to that of NHWU and GLM when the effects are homogeneous. Using HWU, we conducted a genome-wide analysis to study genetic heterogeneity in nicotine dependence. The genome-wide analysis of nearly one million SNPs from the Study of Addiction: Genetics and Environments (SAGE) took 5.8 hours, identifying heterogeneous effects of two new genes (i.e., CYP39A1 and VDAC3) on nicotine dependence.

19
Functional data analysis of blood-based DNA methylation profiles and ovarian cancer risk

Stacey J Winham (1) Sebastian M Armasu (1) Mine S Cicek (1) Melissa C Larson (1) Julie M Cunningham (2) Kimberly R Kalli (1) Brooke L Fridley (3) Ellen L Goode (1)

(1) Department of Health Sciences Research, Mayo Clinic, Rochester MN
(2) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester MN
(3) Department of Biostatistics, Kansas University Medical Center, Kansas City KS

In studies involving blood-based DNA methylation and cancer risk, white blood cell proportions correlate with methylation and cancer status and are potential confounders. Associations between complete blood count (CBC) measures and CpG methylation are known to vary by genomic region. DNA methylation can be viewed as a series of repeated measures for subjects, as closer CpGs tend to be more highly correlated. Functional Data Analysis (FDA) can be applied to such data, viewing the repeated measures as subject-specific functions. We developed an FDA framework to test associations between CBC measures and regional subject-specific CpG methylation (i.e., gene or CpG island). We applied our FDA set-based method to an epithelial ovarian cancer (EOC) study (N=170) to evaluate CBC-CpG associations and differences in case-control status; data used were the proportion of neutrophils among leukocytes and CpG beta values from the Illumina 450k array. We identified multiple CpG regions associated with EOC case-control status, including a 3 KB region of the transcription factor 'CCAAT/enhancer binding protein epsilon' (CEBPE; p=2.0E-15) that also associated with neutrophil proportion (p<1E-16). CEBPE remained associated with EOC status after adjustment for neutrophils (p=8.4E-5), with similar results in a joint regression model (p=1.7E-4) and gene-level principal components analysis (p=3.9E-6). Simulation studies are on-going to evaluate the performance of FDA compared to other methods.

20
Inferring Human Phenotype Networks from Pathway-based Analysis

Christian Darabos (1) Derek Leung (1) Jason H Moore (1)

(1) Institute for Quantitative Biomedical Sciences, Dartmouth College

Expanding volumes of genomic data and research results demand novel perspectives for understanding the pathobiology of common human diseases. Modeling complex biological systems using network analysis offers a promising approach to evaluate the macro-relationships between these biological components. In this project, we construct a Human Phenotype Network (HPN) of over 800 physical attributes, diseases and behavioral traits. We draw links or edges between traits using mapped genes from NHGRI GWAS catalog and gene-to-pathway associations from Reactome, a curated pathway database. The resulting unfiltered HPN is very dense, with over 40,000 edges representing trait to trait relationships defined by shared pathways. Based on its heavy-tailed degree distribution, the HPN appears to be scale-free indicating most traits have few connections and a few traits have many connections. To refine our analysis and discard weak links, we prune the HPN using a multi-scale filtering algorithm, and extract a backbone of statistically significant edges. Then, by applying a community detection algorithm to the filtered HPN, we classify traits by quantifying their shared biology. This classification help us identify non-intuitive relationships and elucidate the shared etiology for certain traits. The HPN provides a means of integrating the accumulating wealth of data on genetic interactions and computationally identifies significant links between traits, attributes and diseases. This model has

tremendous clinical potential to identify risk factors for certain diseases, and common drug targets.

21
Microsatellite Polymorphisms Create an Abundant Source of Expression Variablity

Melissa Gymrek (1) Stoyan Georgiev (2) Barak Markus (3) Jenny Chen (1) Perla I Villarreal (4) Jonathan Pritchard (2) Yaniv Erlich (3)

(1) Harvard-MIT Division of Health Sciences and Technology
(2) Department of Human Genetics, University of Chicago, Chicago, Illinois, USA
(3) Whitehead Institute for Biomedical Research
(4) MIT

A central goal in genomics is to elucidate the genetic architecture of complex traits. So far, efforts to discover eQTL (expression Quantitative Trait Loci) have been mainly focused on the contribution of SNPs and CNVs to gene expression. A few dozens of single gene studies in human and model organisms have suggested that microsatellite variations can modulate expression of nearby transcripts by the formation of secondary structures. Here, we report the first genome-wide survey to identify microsatellites that affect gene expression. We analyzed microsatellite variations across hundreds of 1000Genomes samples using a custom algorithm and performed association tests with the expression levels of nearby transcripts. This process identified more than 3,000 statistically significant expression microsatellites. These associations were replicated across populations and expression assays (RNA-seq and expression array). Moreover, fine-mapping techniques indicated that expression microsatellites association signals are unlikely to stem from SNPs or other variations in linkage disequilibrium, and implicated a large number of them as the causal sites. Our findings raise the possibility of a novel layer of regulatory variants in the human genome. Expression microsatellites might set an additional component to the genetic architecture of gene expression and contribute to the heritability of complex traits.

22
Meta-Analysis of Gene-Level Associations for Rare Variants Based on Single-Variant Statistics

Yijuan Hu (1) Sonja Berndt (2) Stefan Gustafsson (3) Andrea Ganna (3,4) Joel Hirschhorn (5) Kari North (6) Erik Ingelsson (3,7) Dan-Yu Lin (8)

(1) Department of Biostatistics and Bioinformatics, Emory University
(2) Division of Cancer Epidemiology and Genetics, National Cancer Institute
(3) Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University Hospital, Sweden
(4) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden
(5) Department of Genetics, Harvard Medical School
(6) Department of Epidemiology and Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill
(7) Wellcome Trust Centre for Human Genetics, University of Oxford
(8) Department of Biostatistics, University of North Carolina, Chapel Hill

Meta-analysis of genome-wide association studies (GWAS) has led to the discoveries of many common variants associated with complex human diseases. There is a growing recognition that identifying causal rare variants also requires large-scale meta-analysis. The fact that association tests with rare variants are performed at the gene level rather than at the variant level poses unprecedented challenges in the meta-analysis. First, different studies may adopt different gene-level tests, so the results are not compatible. Second, gene-level tests require multivariate statistics (i.e., components of the test statistic and their covariance matrix), which are difficult to obtain. To overcome these challenges, we propose to perform gene-level tests for rare variants by combining the results of single-variant analysis (i.e., p-values of association tests and effect estimates) from participating studies. This simple strategy is possible because multivariate statistics can be recovered from single-variant statistics, together with the correlation matrix of the single-variant test statistics, which can be estimated from one of the participating studies or from a publicly available database. We show both theoretically and numerically that the proposed meta-analysis approach provides accurate control of the type I error and is as powerful as joint analysis of individual participant data. This approach accommodates any disease phenotype and any study design and produces all commonly used gene-level tests. An application to the GWAS summary results of the Genetic Investigation of ANthropometric Traits (GIANT) consortium reveals low-frequency variants associated with human height. The relevant software MAGA is freely available.

23
Integrating Multiple Correlated Phenotypes for Genetic Association Analysis Through Heritability

Jin J.J.Z Zhou (1) Nan N. M. L Laird (1)

(1) Department of Biostatistics, Harvard University, Boston, MA 02115

In genetic studies of complex diseases, many correlated disease variables may be measured for the disorder. A common statistical approach to this problem involves assessing the relationship between each phenotype and each single nucleotide polymorphism (SNP) individually; and taking a

Bonferroni correction for the effective number of tests conducted. Alternatively, one can apply a dimension reduction technique, such as principal components analysis, and test for the association with the principal components of the phenotypes rather than the individual phenotypes. In this paper, by taking the advantage of the heritability and co-heritability, we constructed a more heritable phenotype which is a linear combination of the various phenotypes with the maximal heritability, i.e., MaxH phenotype. Our approach can be applied to both population and family studies; it only requires a method to estimate heritability and co-heritability of the phenotypes. Theoretically and through simulations we compared our approach with various standard methodologies and assessed both the heritability of the overall phenotype and the power. Moreover we provided a guideline of how to choose the phenotypes for combination. Applications of our approach to a COPD genome-wide association study showed the practical relevance of our approach. Finally, we explored the possibility of a new disease classification based on the MaxH phenotype.

24

Association of plasma uric acid with ischemic heart disease and blood pressure: Mendelian randomization analysis of two large cohorts

Nicholas J Timpson (1) Tom M Palmer (2) Børge G Nordestgaard (3) Marianne Benn (3) Anne Tybjærg-Hansen (3) George Davey Smith (1) Debbie A Lawlor (1)

(1) University of Bristol
(2) University of Warwick
(3) Copenhagen University Hospital

We aimed to assess and explain known observational relationships between uric acid/hyperuricemia and ischemic heart disease(IHD), diastolic and systolic blood pressure(DBP and SBP) using variation at SLC2A9(rs7442295) as an instrument for uric acid and FTO(rs9939609), MC4R(rs17782313), and TMEM18(rs6548238) for body mass index (BMI). In 58,072 participants from the Copenhagen General Population Study(CGPS) and 10,602 from the Copenhagen City Heart Study(CCHS) comprising 4,890 IHD cases and 2,282 IHD cases respectively we measured uric acid and related covariables and had access to blood pressure measurements and prospectively assessed IHD. Estimates confirmed known observational associations between plasma uric acid/hyperuricemia and IHD risk, DBP and SBP. However, when using genotypic instruments for uric acid/hyperuricemia in a Mendelian randomisation design there was no evidence for causal relationships between uric acid and primary outcome. Using genetic instruments for the potentially confounding feature BMI, we provided evidence of causal effect of BMI on uric acid/hyperuricemia: 0.03(95%CI 0.02,0.04)mmol/L increase in uric acid and 7.5%(95%CI 3.9%,11.1%) increased risk of hyperuricemia per 4kg/m2. In contrast to observational findings, there is no strong evidence for causal relationships between uric acid and IHD or blood pressure. There is, however, evidence supporting a causal effect between BMI and uric acid level/hyperuricemia risk driven by BMI. This strongly suggests BMI as a confounder in observational associations and suggest a role for elevated BMI/ obesity in the aetiology of uric acid related conditions.

25

Enhancing case-control genetic studies using sample surveys

Parichoy Pal Choughury (1) Daniel Scharfstein (1) Joshua Galanter (2) Chris Gignoux (2) Lindsey Roth (2) Sam Oh (2) Esteban Burchard (2) Saunak Sen (2)

(1) Johns Hopkins University
(2) UCSF

Case control studies are widely used for studying the etiology of diseases with a genetic component. Cases (exhibiting the primary phenotype) are oversampled relative to controls (not exhibiting the primary phenotype). In these studies, it is common for investigators to collect a battery of secondary phenotypes and reuse such data for genetic association. For example we may wish to study obesity or gene expression in a asthma case-control study. Since case-control data are not a random sample from the target population, the observed association between a gene and a secondary phenotype may be biased; to correct for this bias, external information is required. We propose an inferential framework using information from sample surveys that provide representative information about the target population. By way of illustration, we study the relationship between a candidate gene (associated with asthma) and obesity and how this relationship differs by ethnicity. We use data from the GALA II study (an asthma case control study in Latino American children) and the NHIS study (a national sample survey of children). The GALA II study provides information about the conditional distribution of the gene, obesity, and key confounders given asthma status and ethnicity; the NHIS study provides information about the probability of asthma given ethnicity and the key confounders. Information about these distributions from these two distinct data sources are combined together to estimate standardized associations between the gene and obesity within ethnicity strata, which are then compared across the different ethnicities.

26

Testing Association without Calling Genotypes Allows for Systematic Differences in Read Depth between Cases and Controls

Glen A Satten (1) H. Richard Johnston (2) Andrew S Allen (3) Yijuan Hu (2)

(1) Centers for Disease Control and Prevention
(2) Emory University
(3) Duke University

The quality of genotype calling for next-generation sequence data depends on read depth. Loci with high coverage can typically be called reliably, while those with low coverage may be difficult to call. In an association study, if case participants are sequenced to a greater depth than controls, the difference in genotype quality can introduce a systematic bias. This can easily occur when historical controls (e.g., data from The 1000 Genomes Project) are used as controls. We show how to address this bias by directly comparing the proportion of calls for the minor allele between cases and controls, rather than comparing genotypes. We show that tests based on comparing proportions of calls are valid even in the presence of systematic differences in coverage rate between cases and controls in situations where tests based on genotype have inflated size. We also demonstrate that power gains are possible using designs where we increase the number of controls while decreasing the read depth (while keeping total reads constant).

27
Population Stratification Detection and Correction in Rare Variant Collapsing Methods Using Principal Component Analysis

John R Wallace (1) Carrie B Moore (2) Alex T Frase (1) Marylyn D Ritchie (1)

(1) The Pennsylvania State University
(2) Vanderbilt University

Principal Component Analysis (PCA) has often been used in genome-wide association studies (GWAS) correct for population stratification and prevent increased type I errors. Applying PCA to rare variant collapsing methods has not been well studied, the effectiveness of adjusting for population stratification using principal components (PCs) on rare variants is largely unknown. To explore population stratification correction in rare variant data, we collapsed the low frequency (< 5% MAF) variants in the 1000 Genomes Project Phase 1 data based on Entrez gene boundaries using BioBin, a tool for binning rare variants into biologically relevant bins (genes, pathways, regulatory regions, etc.). These bins can then be evaluated with any number of popular rare variant collapsing statistical methods. We analyzed the data for each pairwise combination of the 14 populations available and found dramatic stratification and clustering of populations into continental groups. We then examined multiple approaches to construct PCs using different subsets of the genetic data. To compare approaches, we defined a normalized distance metric between sets of PCs as well as notions of correctness for a given stratification and predictive power without given stratification. We found that for populations close in ancestral history, rare variants should be excluded from PCA, and we developed a method for increasing the sensitivity of PCA, though this method may disguise other hidden stratification. Identifying and properly correcting for stratification remains an important issue; using the techniques described, we demonstrate solutions that identify and correct the ancestry stratification as well as stratification along sequencing technology.

28
Maximizing the power in Principal Components Analysis of Correlated Phenotypes

Hugues Aschard (1) Bjarni Vilhjalmsson (1) Nicolas Greliche (2) Pierre-Emmanuel Morange (3) David-Alexandre Tregouet (2) Peter Kraft (1)

(1) Harvard School of Public Health, department of Epidemiology, Boston, USA
(2) INSERM UMR_S 937, ICAN Institute for Cardiometabolism And Nutrition, Pierre et Marie Curie University, Paris 6, France
(3) INSER UMR_S 1062, Aix-Marseille University, Marseille F-13385, France

Principal Component analysis (PCA) is a useful tool that has been widely used for the multivariate analysis of correlated variables. It is usually applied as a dimension reduction method: the few top principal components (PCs) explaining most of total variance are tested for association with a predictor of interest, and the remaining PCs are ignored. This strategy has been widely applied in genetic epidemiology, however some aspects of this analytical technique are not well appreciated in the context of single nucleotide polymorphisms (SNPs) testing. In this study we review some of the theoretical basis of PCA and describe the behavior of PCA when testing for association between a SNP and two correlated traits under various scenarios. We then evaluate through simulations the power of a few different PCA-based strategies when analyzing up to 100 traits. We show that contrary to widespread practice, testing the top PCs only can be dramatically underpowered since PCs explaining a low amount of the total phenotypic variance can harbor a substantial part of the total genetic association. We also demonstrate that PC-based strategies can only achieve a moderate gain in power in the presence of positive pleiotropy, but have great potential to detect negative pleiotropy (e.g. positive correlation and opposite genetic effects) or genetic variants that are associated with a single trait highly correlated to others. Finally we show that a joint test of all PCs is the optimal approach in most scenarios. To illustrate these phenomena, we present an analysis of five venous thrombosis related traits in 685 subjects from the MARTHA study. Joint analysis of all five PCs identified two new potential candidates SNPs, which had strongest associations with the 5th PC.

29
Imputation of Case/Control Study Samples Genotyped On Different SNP Chips

Kristin L Ayers (1) Heather J Cordell (1)

(1) Institute of Genetic Medicine, Newcastle University

Imputation has become an important tool for association testing of untyped markers between cohorts. As large sets of controls have become available, it is not uncommon for case and control samples to be typed on different genotype chips and/or platforms. Naively performing imputation and association on these data sets can result in large numbers of false associations. The common practice of performing imputation on the set of overlapping SNPs is not ideal as information from a number of genotyped markers may be dismissed. We have developed a pipeline that uses information from the imputation reference panel to eliminate or recalibrate the problem SNPs amongst those genotyped in one study sample and not the other, which are responsible for the majority of the false positives. Problems occur at markers where either: (1) a study sample has an allele frequency that differs between the reference and study sample, and (2) the imputation is more accurate in one study sample than the other. As an illustration, we use 2 sets of WTTCC data for our case/control samples, one typed on the Affy6.0 chip, and one typed on the Illumina2.1M chip. On chromosome 22, there are approximately 8K Affy SNPs and 14K Illumina SNPs, of which only 4K overlap. Using all 500K 1000 Genomes SNPS to perform imputation, there are over 600 genome-wide significant SNPs after standard imputation QC. However, using our pipeline, only 10-15% of the genotyped SNPs need to be removed prior to imputation to eliminate the majority of the false positives.

30
Prevention and control of β-thalassemia and other genetic diseases in consanguineous Pakistani population

Shahid SMB Baig (1) Uzma UA Abdullah (1) Syeda SSW Waseem (1) Abubakar AM Moawia (1) Maria MA Asif (1) Hafsa H Hafsa (1) Muhammad MS Sher (1) Ambrin AF Fatima (1)

(1) National Institute for Biotechnology and Genetic Engineering (NIBGE)

The tradition of consanguineous marriages in Pakistan (>60%) is resulting in high incidence of various genetic diseases. There is no national program for prevention of any genetic disorder for this world's sixth largest population. β-thalassemia is the most common genetic disorder in this country with an allele frequency of 6.2% leading to more than 7,000 transfusion dependent births every year. Similarly, incidence of other genetic disorders inherited including deafness, primary microcephaly, skin, skeletal, neurological and eye disorders is also higher than other populations of the world. The objective of this study was to carry out molecular characterization of various common and rare diseases in the Pakistani population to establish genetic counseling, carrier screening and prenatal diagnosis for prevention and control of genetic diseases. Under this program we identified more than 900 β-thalassemia families through blood transfusion dependent index children attending transfusion centers in various cities and offered the couples options to prevent the affected births. In the last four years, more than 400 retrospective prenatal diagnoses of β-thalassemia through first trimester chorionic villus sampling (CVS) in at risk pregnancies have been carried out. For other genetic diseases, 980 large consanguineous families with various disease phenotypes having at least three affected births were identified from hospitals and mainly by conducting field sampling trips of remote rural areas. We have already reported a number of disease loci, genes and mutations for various genetic disorders and started to provide genetic counseling, carrier screening and prenatal diagnosis. We are also providing prenatal diagnosis service to Punjab Thalassemia Prevention Program (PTPP) to control this most common serious genetic disorder in the province of Punjab having more than 100 million people. In addition, a large consanguineous family with history of childhood Glioblastoma (Brain tumor) and six childhood deaths due to brain tumor was also identified and characterized under this program. In this family, using the candidate gene strategy we identified a point mutation (-T) in exon 6 of PMS2 gene involved in DNA repair. Subsequently, three prenatal diagnoses have been provided to this extended family for prevention f affected births. These are the first ever prenatal diagnoses of brain tumor (glioblastoma) in the first tirmester of pregnancy by DNA analysis. The molecular analysis for various genetic diseases is underway and we are optimistic to significantly control the incidence of genetic disorders in this highly inbred population of an alarming incidence of genetic diseases.

31
Analyses of WES data in multiplex Syrian oral clefts families

Joan E Bailey-Wilson (1) Margaret M Parker (2) Silke Szymczak (1) Qing Li (1) Cheryl D Cropp (1) Marcus M. Nöthen (3) Jacqueline B Hetmanski (2) Hua Ling (4) Elizabeth W Pugh (4) Priya Duggal (2) Margaret A Taub (5) Ingo Ruczinski (5) Alan F. Scott (4) Mary L. Marazita (6) Hasan Albacha-Hejazi (7) Elisabeth Mangold (3) Terri H. Beaty (2)

(1) Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD USA
(2) Department of Epidemiology, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD, USA
(3) Institute of Human Genetics, University of Bonn, Bonn, Germany
(4) Center for Inherited Disease Research, Johns Hopkins

University, Baltimore, MD, USA
(5) Department of Biostatistics, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD, USA
(6) Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA, USA
(7) Ibn Al-Nafees Hospital, Damascus, Syrian Arab Republic

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 genome wide association and linkage studies. This whole exome sequencing (WES) study used 22 affected 2nd degree or more distant relatives drawn from 10 multiplex inbred families (2 families with 3 relatives and 8 families with 2 relatives) initially ascertained in the Syrian Arab Republic for linkage studies. WES was done by the Center for Inherited Disease Research using the Agilent SureSelect v.4 capture reagents & Illumina HiSeq 2000 sequencers. Variants were called for all samples together within this project using Unified Genotyper (2.3-9). Variants were flagged by VQSR annotation using a Gaussian Mixture model for both SNVs and INDELs. Random Forests was used to estimate probability of high versus low quality calls. Results of analyses using the Ingenuity software on these newly recalled data will be presented. 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.

32
Computationally efficient inference for family-based genomewide association studies with random effects and missing data

M Fazil Baksh (1) Pianpool Kirdwichai (1)

(1) Reading University

Family studies of disease-gene association are well known to be robust against spurious findings due to potential genetic confounders and, in studies of rare alleles, are often preferred to studies based on the use of unrelated cases and controls. However, unlike studies of unrelated individuals and due to the effect of both genetic and environmental factors, data on families tend to be correlated. Standard solutions to this problem use a generalised linear mixed model with analyses that are computationally complex and time-consuming, being based on either methods that require high dimensional integration of the random components, or on resampling techniques. In this talk alternative, computationally efficient methods based on application of hierarchical generalised linear models are developed and evaluated for data from family studies with random effects and missing parental data. Use of a h-likelihood makes it possible to make inference directly without resorting to either computationally intensive methods, or the use of prior probabilities. The models will be assessed via simulations and applied to data from a study of human systemic lupus erythematosus (SLE) and polymorphisms in the c-reactive protein (CRP) gene.

33
Leveraging Auxiliary Information for SNP Selection in Genetic Association Studies

Adrian Coles(1), Veera Baladandayuthapani (1) Robert Yu (1) Sanjay Shete (1)

(1) UT MD Anderson Cancer Center

Genetic association studies aim to find marginal or joint effects of multiple SNPs on outcome(s) of interest and several approaches have been proposed with varying degree of success. Most of these approaches assume that (a priori) each SNP has equal chance of being associated with an outcome. However, in some cases there exists substantial auxiliary information from different studies in the same disease area that can be incorporated into the analyses for more refined inference. Examples of such auxiliary information can be disease-specific alternate domain knowledge such as those obtained from transcriptomic (expression-based studies), epigenetic and integrative studies. Our aim is to leverage this information and refine selection of disease-associated SNPs. We do so in a Bayesian variable selection framework and incorporate the auxiliary information as structural priors on the probabilities of selection of the SNPs – thus allowing simultaneous selection and sparse modeling. We illustrate our methods using a Glioblastoma GWAS dataset where we leverage the gene signatures obtained by The Cancer Genomic Atlas (TCGA) based genomic studies.

34
Powerful testing via hierarchical linkage disequilibrium in haplotype association studies

Brunilda B. Balliu (1) Jeanine J Houwing-Duistermaat (1) Stefan S Boehringer (1)

(1) Leiden University Medical Centre

Haplotypes play key roles in the study of the genetic basis of disease. Studies have shown that haplotype-based methods (HBM) may provide more power and accuracy in disease gene mapping than those based on single markers. A limitation of HBM is that the number of parameters increases exponentially with the number of loci l, incurring many degrees of freedom (df) and weakening the power to detect associations. Moreover, the success of these approaches depends on getting the "right size" haplotypes. If the haplotypes are too long to cover relevant correlations (e.g. more than 10 loci), such approaches are not feasible. These situations can occur when relatively recent mutations have introduced long range correlations in low

linkage disequilibrium (LD) regions. To address these limitations, we propose hierarchical modeling of LD for disease mapping. We develop a new parametrization of the haplotype distribution where every parameter corresponds to the cumulant of each possible subset of a set of loci. That is, the new parametrization consists of the allele frequencies at each locus, the pairwise, and higher-order (3,...,l) LD parameters. This introduces a hierarchy in our parameters and enables us to selectively test differences that are described in terms of a certain number of loci (for example 2-way interactions), ignoring higher order parameters and sparing df to test for the full haplotype. We perform a simulation study and show that our approach maintains the type I error at nominal level and has increased power under many realistic scenarios, as compared to traditional HBM. To evaluate the performance of our proposed methodology in real data we analyze data from the WTCCC study, a Genome Wide Association Study on Rheumatoid Arthritis.

35
Reclassification in genetic risk prediction over time

Richard T Barfield (1) Joel Krier (2) Robert Green (3) Peter Kraft (1,4)

(1) Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts
(2) Harvard Medical School Genetics Training Program, Boston, Massachusetts
(3) Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts
(4) Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts

Genome wide association studies have identified thousands of common alleles with modest effects on risk of complex disease. The clinical utility of genetic risk profiles based on these variants depends crucially on the number and effect size of identified loci, and how stable the predicted risks are as additional loci are discovered. Individuals flagged as high risk at one time may be reclassified as low risk or vice versa as more loci are identified. Using breast cancer (BrCa) and heart disease (CHD) as examples, we quantify this reclassification over the last six years and use the distribution of identified genetic effects to project likely changes in risk prediction. The range of the predicted risks increased from 2007 to 2013: the 95th risk percentiles rose from 1.34 and 1.33 times the population average to 2.11 and 2.22 for CHD and BrCa, respectively. This caused 6% of the population to be reclassified from lower than 2x average CHD risk in 2007 to higher in 2013. The reclassification proportion for BrCa was 7%. The future reclassification projected from doubling available GWAS samples was notably smaller. This suggests that risk estimates in the 0.5-2.0 times average range are stable for diseases that have been as extensively studied as CHD and BrCa. Moreover, our projections place upper bounds on the contribution of rare variants to total heritability, giving insight into clinical utility of rare variants.

36
Population Structure in the Cincinnati area

Michelle B. Baric (1) Mehdi A. Keddache (2) Lisa J. Martin (2) Cynthia A. Prows (2)

(1) University of Cincinnati
(2) Cincinnati Children's Hospital Medical Center

Background: Population studies from a single study site may result in unexpected substructure, especially in the United States, where admixture is common. It is important to examine substructure of US cities to better understand potential biases in genetic studies. This study examines continental structure and substructure of a metropolitan city, Cincinnati, OH. Methods: Study subjects included self-reported white and black participants in the Cincinnati Genomic Control Cohort (GCC). Principal component analyses (PCA) and Eigensoft were used to analyze the data, by chromosome, to determine population structure and the level of agreement between self-reported and genetic race. Within the self-reported white population, substructure was evaluated using markers from several chromosomes at varying allele frequencies and with three ancestry informative marker (AIMs) sets. Results: Continental structure was observed between self-reported white and black populations. The overall rate of agreement between self-reported and genetic race was 99%. AIMS failed to identify substructure in the self-reported white population however clusters were present when using minor allele frequencies. Moreover, different chromosomes produced different clustering patterns, when examined at the same allele frequency. Conclusions: Self-reported race is a good tool for determining continental ancestry; however it may not be sufficient when trying to achieve homogeneous cases and controls. If identified substructure differs based on the variants selected, it will be important to ensure that a robust set of variants are selected, that reflect variation across the genome, to detect potential population stratification.

37
Robust methods in three common statistical genetics applications

Justo Lorenzo Bermejo (1) Miriam Kesselmeier (1) Carine Legrand (1) Maria Kabisch (2) Christine Fischer (3) Ute Hamann (2) Barbara Peil (1)

(1) Statistical Genetics Group, Institute of Medical Biometry and Informatics, University of Heidelberg, Germany
(2) Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany

(3) Institute of Human Genetics, University of Heidelberg, Germany

Departing observations (outliers) have a strong impact on standard statistics. Robust statistics aim to reduce the influence of outliers whilst maximizing statistical efficiency in case of normality. Although outliers are common and often indiscernible in high-dimensional genomewide association studies, robust techniques are hardly applied in this field. We have examined the potential of a subset of robust methods to investigate population structure, to build a reference panel for genotype imputation, and to predict genetic risk. Regarding population structure, the investigated techniques for principal component analysis (PCA) include spherical PCA and projection pursuit. Simulations relying on the simuPOP environment and HapMap suggest that robust PCA reflects better evolution-related population structure than standard PCA. Regarding genotype imputation, we consider the sequencing of a subset of the study population in addition to external genotypes. HapMap-based results indicate that additional sequences from study individuals improve imputation accuracy, especially when individuals are selected relying on the univariate depth after standard PCA. Regarding genetic risk prediction, our results show a lower impact of genotyping errors on relative risks estimated by robust logistic regression than on standard estimates. Our findings on three common applications suggest some advantage of robust methods in genetic association studies.

38
Detecting Rare Haplotype-Environment Interaction with Logistic Bayesian LASSO

Swati Biswas (1) Shuang Xia (2) Shili Lin (2)

(1) University of Texas at Dallas
(2) The Ohio State University

Two important contributors to missing heritability are believed to be rare variants and gene-environment interaction (GxE). Thus, detecting GxE where G is a rare haplotype is a pressing problem. Haplotype analysis is usually the natural second step to follow up on a genomic region that has been implicated to harbor associated variants through single nucleotide variants (SNV) analysis. Further, rare haplotype can tag associated rare SNV and provide greater power to detect them than popular collapsing methods. Recently we proposed Logistic Bayesian LASSO (LBL) for detecting rare haplotype association with case-control data. LBL shrinks the unassociated (especially common) haplotypes towards zero so that an associated rare haplotype can be identified with greater power and its effect size estimated more precisely. Here we incorporate environmental factors and their interactions with haplotypes in LBL. As LBL is based on retrospective likelihood, this extension is not trivial. We model the joint distribution of haplotypes and covariates given the affection status. We carry out simulations to investigate the properties of LBL for detecting interactions. We find that it is easier to detect an interaction if the corresponding main effects of haplotype or covariate are not strong. We also apply the proposed approach to the Michigan, Mayo, AREDS, Pennsylvania Cohort Study on Age-related Macular Degeneration (AMD). LBL is able to detect interaction of a specific rare haplotype in CFH gene with smoking. To the best of our knowledge, this is the first time in the AMD literature that interaction of smoking with a specific (rather than pooled) rare haplotype has been implicated.

39
Ensemble testing of multivariate phenotypes in genetic association studies

Stefan Boehringer (1) Dagmar Wieczorek (2) Andreas Wollstein (3)

(1) Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands
(2) Institut fuer Humangenetik, Universitaet Duisburg-Essen, Germany
(3) Section of Evolutionary Biology, Department of Biology II, University of Munich LMU, Germany

Multivariate phenotypes can be challenging to analyse, especially in the high-dimensional setting. For example, in the analysis of facial shapes, distances derived from landmarks can range in the thousands. Often, a first research question is hypothesis-free, namely to identify any genetic association with any transformation of the multivariate phenotype. Such associations can be found by global testing and we propose a procedure working on ensembles of global tests. Given a fixed number of testing procedures, we develop a cross-validation/permutation procedure that selects only non-overfitting members of the ensemble. This is guaranteed by an inherent permutation step resulting in maintenance of type-I error. Simultaneously, we can select the most powerful of the testing procedures which do maintain type-I error. We apply this framework to genetic associations of facial traits and choose our ensemble from penalized linear regression models on a grid of penalty parameters. SNPs are regressed on distances between landmarks of facial data thereby forming linear combinations of distances associated with SNPs by means of inverse regression. The power of our procedure is optimal whenever the true model is part of the ensemble. We demonstrate in a data set of 2D images, that our testing procedure compares favorably to marginal analysis, i.e. testing one SNP against one component of the multivariate trait and some other global testing procedures. Our testing procedure can be applied in the high-dimensional setting and can potentially also be used in the analysis of rare variants with respect to univariate phenotypes. We discuss a challenging extension considering both multivariate phenotypes and multiple markers simultaneously.

40

Comparison of permutations strategies to assess gene-set significance in gene-set-enrichment analysis

Myriam Brossard (1) Amaury Vaysse (1) Eve Corda (2) Hamida Mohamdi (1) Diana Zelenika (3) Brigitte Bressac-de Paillerets (4) Marie-Françoise Avril (5) Mark Lathrop (6) Florence Demenais (1)

(1) INSERM U946, Paris, France; Université Paris Diderot, Paris, France
(2) INSERM U946, Paris, France
(3) CEA, Institut de Génomique, CNG, Evry, France
(4) INSERM U946, Paris, France; Université Paris Diderot, Paris, France; Institut de Cancérologie Gustave Roussy, Villejuif, France
(5) Hôpital Cochin, Paris, France
(6) CEA, Institut de Génomique, CNG, Evry, France; Genome Quebec Innovation Centre, McGill University, Montreal, Canada

A popular approach in pathway analysis is the gene-set enrichment analysis (GSEA) that tests for a pathway enriched in genes more strongly associated with the phenotype than genes outside the pathway. Significance for enrichment is classically obtained through permutations of phenotypes (PP), which retain the linkage disequilibrium (LD) pattern among SNPs but are time consuming and require access to raw SNP data. More efficient permutation strategies that only need single-SNP statistics have been proposed: (1) an approach based on randomization of SNP statistics (R-SNP); (2) a new circular genomic permutation approach which permutes SNP statistics ordered with respect to their genomic locations by rotation to account for LD (CG-SNP). Our goal was to compare R-SNP and CG-SNP permutations to the reference PP strategy in GSEA. These comparisons were made on a dataset of 1,179 French melanoma cases (MELARISK study) and 2,797 controls analyzed for association with 1,032,745 Hapmap3-imputed SNPs. SNPs were assigned to 21,810 genes (NCBI Build 37.1) and genes to 316 Level 4-Gene Ontology (GO) classes. False-Discovery Rates (FDR) of GOs were estimated using 1,000 phenotype permutations, 10,000 R-SNP and 3,000 CG- SNP permutations. FDRs estimated from PP were more strongly correlated with CG-SNP FDRs than with R-SNP FDRs (Spearman correlations of 0.90 ($p<10-5$) and 0.80 ($p<10-5$) respectively). At FDR ≤10%, 3 GO classes (response to light stimulus, regulation of mitotic cell cycle, induction of programmed cell death) were detected by all 3 strategies while there were 7 additional GOs identified by R-SNP permutations. CG-SNP permutations appear to keep similar FDRs to the reference strategy while R-SNP permutations may be too liberal.

41

Genome-Wide Association Analysis of Density Gradient Ultracentrifugation Data

Angelo J Canty (1) Nabin M Shrestha (1) Marie-Pierre Sylvestre (2) John D Brunzell (3) Andrew P Boright (4) Shelley B Bull (5) Andrew D Paterson (6)

(1) McMaster University
(2) Université de Montréal
(3) University of Washington
(4) University of Toronto
(5) Samuel Lunenfeld Research Institute
(6) Hospital for Sick Children, Toronto

Density gradient ultracentrifugation (DGUC) is a method which separates and measures the abundance of various fractions of lipoproteins, characterized by size and density, in plasma. These measurements can be considered an ordered vector giving a lipid profile for each study subject. We present a method to model the lipid profile using penalized splines fitted by standard linear mixed models. For any SNP, the effect of genotype on the fitted profile can be incorporated using fixed effects for the genotype and genotype*fraction interactions where both genotype and fraction are coded as categorical covariates. To test for a difference between the profiles by genotype we compare likelihoods with and without these effects. Simulations, based on observed DGUC data collected for participants in the Diabetes Control and Complications Trial (DCCT), show that the resulting test statistic has a non-standard null distribution with a large degree of excess variability relative to the expected asymptotic chi-squared distribution. We propose to account for this over-dispersion using a new three parameter distribution related to the chi-squared. The parameters of the null distribution can be estimated from observed GWAS test statistics and hence the test p-values can be evaluated without the need for permutations. Applying our method to the DCCT data we find significant association at known HDL and LDL loci. We also find a new locus not previously reported in studies using standard measures of HDL or LDL. We perform simulations based on the observed DCCT data to evaluate the type I error control of our method and compare its power to alternative methods for DGUC data analysis as well as methods using standard lipoprotein measures.

42

Classifying rare variants from sequenced data

Marinela Capanu (1) Venkatraman Seshan (1) Colin B Begg (1)

(1) Memorial Sloan-Kettering Cancer Center

The detection of rare deleterious 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. 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 harmful versus harmless is a particularly important but challenging goal because of the sparseness of the evidence for each individual variant. Using simulations we derive a strategy for classifying rare variants as deleterious versus neutral with well understood properties. We illustrate the methods with an application to a real study of breast cancer.

43
Integrative analysis of genetic variation and DNA methylation in an ovarian cancer etiology study

Prabhakar Chalise (1) Sebastian M. Armasu (2) Mine S. Cicek (2) Julie M. Cunningham (2) Kimberly R. Kalli (2) Melissa Larson (2) Robert A. Vierkant (2) Ya-Yu Tsai (3) Zhihua Chen (3) Devin Koestler (4) Thomas A. Sellers (3) Ellen L. Goode(2) Brooke L. Fridley (1)

(1) University of Kansas Medical Center, Kansas City, KS USA
(2) Mayo Clinic, Rochester, MN USA
(3) Moffitt Cancer Center, Tampa, FL USA
(4) Dartmouth, Hanover, NH USA

We propose a novel approach for the integration of genome-wide genetic and epigenetic data to determine methylation related genetic loci associated with complex diseases. The method determines whether the difference in methylation for a given CpG probe between the age-matched case-control pairs can be explained by the difference in the genotype (additive genetic model) at the gene-level. For each SNP marker, we determined the difference in the genotype between the matched case-control pairs, followed by a gene-level summary of the difference using the principal components which explain 80% of the variation. This analysis approach was applied to two epithelial ovarian cancer (EOC) studies involving 66 (Study 1) and 148 (Study 2) EOC case and age-matched control pairs, for which both blood-based genome-wide SNP and methylation (Illumina 27K) data were available. To reduce false positives, CpG probes known to be associated with leukocyte cell type and non-specific probes were removed from the analysis, leaving 15,944 CpG probes. A total of 300,878 SNPs were mapped to +/- 20 KB of 26,546 genes. Integrative analysis found the top associations in both studies to involve the methylation probe cg10237469 near CEACAM4 and the genetic variation within: CEACAMP8 ($p = 1.4\times10\text{-}21$ and $1.4\times10\text{-}61$); CEACAMP2 ($p = 1.4\times10\text{-}21$ and $6.9\times10\text{-}59$); and CEACAMP1 ($p = 1.7\times10\text{-}20$ and $4.7\times10\text{-}55$). Further studies are needed to determine the functional relevance of these genes in EOC etiology.

44
Genome-wide pattern of informative missingness using HapMap data

Yu-Ping Chang (1) Chao-Yu Guo (1)

(1) National Yang Ming University

In addition to the popular case-control design, the family-based gnome-wide association study is another crucial design in identifying SNPs associated with traits of interest without the confounding of population stratifications. In general, quality controls for such type of study include Hardy-Weinberg Equilibrium and the allele frequencies. However, missing pattern of parental genotypes is usually ignored. Based on the freely available phase-I HapMap data consists of trios of CEPH and YRI, the genome-wide pattern of informative missingness due to genotyping quality is examined using the new test of informative missingness (TIMBD, 2012). The results revealed that SNPs with significant signals do not have replicated studies as well as the other SNPs in the neighborhood. The phenomenon may explain part of the SNPs that could not be replicated in multiple studies.

45
Application of Weighted Quantile Regression to Pathway Analysis

Zhu Chen (1)

(1) University of Southern California

Genome-wide scan is widely used to identify pathogenesis of complex diseases. However, genes must have very strong effects to survive through multiple testing correction, while complex diseases are often caused by multiple moderate effect genes. Therefore, pathway-based approaches - which summarize cumulative effects of genes - have been proposed and exploded in use. Most of those methods could be categorized into two types. In the overrepresentation analysis, genes are classified into significant/non-significant; and then disparity for the proportion of significant genes between the pathway and the rest are tested. Another type is to rank genes by p-values and test rank difference. Those methods could be implemented in regression framework with pathway as independent variable and p-values (significance/rank) as dependent, and also with the incorporation of other covariates. We proposed to apply quantile regression framework to pathway analysis, based on the assumption that the disparity for the p-value mean/median are vague. Instead, the p-values (decreasing order) in large percentile (80th – 98th) are different. Generalized linear square are utilized to summarize the results from correlated quantiles. To evaluate the performance of our method, we conduct simulations to compare the power among quantile regression, logistic

regression, rank transformation regression, Mann-Whitney test, weighted Kolmogrov-Smirnov, etc. Based on our results, the regression framework outperforms others under all scenarios. Logistic regression have the largest power, while type I error is slightly larger than 0.05. The quantile regression has good performance with large pathway (>40 genes) and the best range of quantiles is 0.70-0.98..

46
Rare Variant Tests for Time-to-event Outcomes

Han Chen (1) Thomas Lumley (2) Jennifer Brody (3) Nancy L Heard-Costa (4) Caroline S Fox (5) Adrienne Cupples (1) Josée Dupuis (1)

(1) Department of Biostatistics, Boston University School of Public Health, Boston, MA
(2) Department of Statistics, University of Auckland, Auckland, New Zealand
(3) Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA
(4) Department of Neurology, Boston University School of Medicine, Boston, MA
(5) Division of Endocrinology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

Rare variant tests have been of great interest in testing genetic associations with diseases and disease-related quantitative traits in recent years. Burden tests (BT) are one class of rare variant tests, which collapse genotypes from multiple variants into a summary genetic burden score and test the association between the trait and the burden score. Burden tests are most powerful when the proportion of causal variants is high and all causal variants have the same direction of effects. Alternatively, other tests, including the Sequence Kernel Association Test (SKAT) [Wu et al. 2011, Am J Hum Genet], are performed without collapsing genotypes and preferred when variants with both protective and detrimental effects are expected. It is also known that SKAT statistic can be expressed as a weighted sum of single marker test statistics, when the linear kernel is used. However, little attention has been paid to rare variant tests for time-to-event outcomes. Here we present both BT and SKAT based on a Cox proportional hazard model, using both the score test and likelihood ratio test (LRT). All approaches can be easily performed in a meta-analysis context. We show in simulation studies that score tests of BT and SKAT have inflated type I errors at low significance levels, when the proportion of censoring is high. In contrast, LRT provides accurate control of type I error. We also present results of a real data example from the Framingham Heart Study.

47
Knowing Your NGS Downstream: Functional Predictions

G. Bryce G. B. Christensen

(1) Golden Helix

Next-Generation Sequencing analysis workflows typically lead to a list of candidate variants that may or may not be associated with the phenotype of interest. Any given analysis may result in tens, hundreds, or even thousands of genetic variants which must be screened and prioritized for experimental validation before a causal variant may be identified. To assist with this screening process, the field of bioinformatics has developed numerous algorithms to predict the functional consequences of genetic variants. Algorithms like SIFT and PolyPhen-2 are firmly established in the field and are cited frequently. Other tools, like MutationAssessor and FATHMM are newer and perhaps not known as well. This presentation will review several of the functional prediction tools that are currently available to help researchers determine the functional consequences of genetic alterations. The biological principals underlying functional predictions will be discussed together with an overview of the methodology used by each of the predictive algorithms. Finally, we will discuss how these predictions can be accessed and used within the Golden Helix SNP & Variation Suite (SVS) software. The content of this proposed presentation is based on a webcast that drew over 350 viewers with many positive reviews.

48
8q24 Risk Alleles and Prostate Cancer in African-Barbadian Men

Cheryl D. Cropp (1) Christiane M. Robbins (2) Anselm J.M. Hennis (3) John D. Carpten (2) Lyndon Waterman (3) Jeffrey M. Trent (2) M. Cristina Leske (4) Suh-Yuh Wu (4) Joan E. Bailey-Wilson (1) Barbara Nemesure (4)

(1) Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland
(2) Integrated Cancer Genomics Division, Translational Genomics Research Institute (TGen), 445 N. Fifth Street, Phoenix, AZ
(3) Department of Biological & Chemical Sciences, University of the West Indies, Bridgetown, Barbados
(4) Department of Preventive Medicine, Stony Brook University Medical Center, Stony Brook, NY

African American men (AA) exhibit a disproportionate share of prostate cancer (PC)incidence, morbidity and mortality compared to other groups. Several genetic association studies have implicated select loci in the 8q24 region as increasing PC risk in AA. We evaluated the association between previously reported 8q24 risk alleles and PC in African-Barbadian (AB) men, also known to have high rates of PC. Ten previously reported tag SNPs were genotyped in 447 AB men with PC and 385 AB controls from the Prostate Cancer in a Black Population (PCBP) study. Only rs2124036 was nominally significant in AB men, (OR = 2.0, 95% CI (1.0-4.3), P=0.06)

for the homozygous C/C genotype after correction for multiple testing. We also conducted a meta-analysis including our AB population along with two additional African-Caribbean populations from Tobago and Jamaica for SNPs rs16901979 and rs1447295. A significant association resulted for the rs16901979 A allele (Z score 2.75; p=0.006; summary OR= 1.21 (95% CI: 1.01-1.46)). Our findings may indicate: i) the presence of a founder effect; ii) the selected SNPs not being tagged to an ancestral haplotype bearing the 8q24 risk allele(s) in this population; or iii) inadequate power to detect a true association. Additional GWAS and sequencing studies are underway to further interrogate any potential contribution of the 8q24 region to PC in this West African derived population.

49
Characterization of Polygenic Signatures for Tourette Syndrome and Obsessive Compulsive Disorder

Lea K Davis (1) Dongmei Yu (2) Hae Kyung Im (1) Benjamin M Neale (3) Eske M Derks (4) Evelyn S Stewart (5) James Knowles (6) Carol Mathews (6) James M Scharf (2) Nancy J Cox (1)

(1) University of Chicago
(2) Massachusetts General Hospital
(3) Broad Institute
(4) University of Amsterdam
(5) University of British Columbia
(6) University of Southern California

Genetic analysis of psychiatric disorders is quickly moving from single variant tests to cumulative testing of many thousands of variants in global analyses designed to elucidate the genetic architecture of complex phenotypes rather than to hone in on a single risk variant. Recent work from our group has shown that Tourette Syndrome (TS) and Obsessive-Compulsive Disorder (OCD) have differing, yet overlapping genetic architectures, suggest that each phenotype may harbor its own "polygenic signature". We have applied three methods to the development of polygenic signatures capable of predicting TS and/or OCD in independent data sets. These methods include prediction of phenotype based on a polygenic signature derived from 1) top association signals, 2) best linear unbiased predictors (blups), and 3) genomic kriging. We show that the TS and OCD polygenic signatures can predict TS (p=0.018) and OCD (p=0.0007) respectively in an independent sample. Additionally, we find that by using the kriging approach, the OCD polygenic signature can predict ~1% of the case/control variance in an independent TS (p=1.04e-08) sample, and the TS polygenic signature can predict ~1% of the variance in an independent OCD (p=6.76e-06) sample. We also present results from additional weighting schemes intended to improve predictive power.

50
Next generation sequencing and its application in clinical practice

Mariza de Andrade (1) Jie Na (1) Paul A Decker (1) Shannon K McDonnell (1) Stephen N Thibodeau (1) Eric W Klee (1)

(1) Mayo Clinic, Rochester, MN, USA

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
Should we account for the random effect of relatedness when using principal components analysis in GWAS?

Mariza de Andrade (1) Julia P Soler (2)

(1) Mayo Clinic, Rochester,MN, USA
(2) University of Sao Paulo, SP, Brazil

Studies of human complex diseases and traits associated with candidate genes are potentially vulnerable to bias (confounding) due to population stratification and inbreeding, especially in admixture population. In genome-wide association studies (GWAS) the Principal Components (PCs) method provides a global ancestry value per subject, allowing corrections for population stratification. However, these coefficients are typically estimated assuming unrelated individuals and if family structure is present and is ignored, such sub-structure may induce artifactual PCs. Extensions of the PCs method have been proposed by Konishi and Rao

(1992) taking into account only sibship relatedness and by Oualkacha et al. (2012) which can be applied to general pedigrees and high dimensional data. In this work we apply such analysis for estimation of global individual ancestry but admitting PCs extracted from different variance components matrix estimators. For the application we use the GENOA sibship data consisting of European and African American subjects and the Baependi Heart Study consisting of 80 extended families collected from the highly admixture Brazilian population, both with SNPs data from Affymetrix 6.0 chip. All the implementation are done using R package.

52
Genetic Loci on 8q23 and 15q21 Influence the Effect of Smoking on Carotid Plaque Burden: Results from a Multiethnic Cohort

Chuanhui Dong (1) Liyong Wang (2) Ashley Beecham (2) Digna Cabral (1) Susan H Blanton (2) Hongyu Zhao (3) Ralph L Sacco (1) Tatjana Rundek (1)

(1) Department of Neurology, University of Miami
(2) John P. Hussman Institute for Human Genomics, University of Miami
(3) Yale University

Smoking greatly increases the risk of atherosclerotic plaque and the effect may vary from individual to individual. A genome-wide scan was performed for smoking × SNP interactions on carotid plaque burden (CPB) to identify the potential genetic moderators. Carotid B-mode ultrasonography and genotyping with the Affymetrix 6.0 chip were performed in 1,010 subjects (15% white, 17% black, 66% Hispanic). CPB was expressed as the sum of plaque areas over the segments in common and internal carotid arteries and bifurcation. Smoking was classified as 0, < 20, and ≥20 cigarette pack-years. Assuming an additive genetic model, regression analysis was conducted to test for smoking × SNP interaction on the cube root transformed CPB while controlling for age, sex, and the top 3 principal components of ancestry. Two SNPs showed an interaction with smoking on CPB with a p value <5.0x10-6 and the effects were similar in whites, blacks and Hispanics. Specifically, for rs1436719 on 8q23, the adjusted mean CPB was greatly increased with cigarette pack-years among T allele carriers (0.9, 1.6, 2.4 for 0, <20, and ≥20 cigarette pack-years, respectively; p=1.4x10-10), but similar among non-T allele carriers (P=0.10). For rs971982 on 15q21, the adjusted mean CPB was substantially increased with cigarette pack-years among T allele carriers (1.0, 1.6, 2.0 for 0, <20, and ≥20 cigarette pack-years, respectively; p=5.0x10-10), but very similar among non-T allele carriers (P=0.50). Consistent evidence across race-ethnic groups suggested that the genetic loci on 8q23 and 15q21 may modulate the effect of smoking on CPB and highlighted these regions for further investigation of the functional genetic variants.

53
Improved detection of genetic exposures with unspecified effect modifiers

Todd L Edwards (1) Chun Li (2)

(1) Center for Human Genetics Research, Division of Epidemiology, Department of Medicine, Vanderbilt University
(2) Center for Human Genetics Research, Department of Biostatistics, Vanderbilt University

Complex phenotypes often result from interplay of multiple genetic and environmental factors. Association analyses can gain power by modeling interaction effects when they exist. However, existing methods require explicit specification of effect modifiers, and exhaustive pairwise scans of SNPs introduce well known computational, statistical, and logistical challenges. For continuous phenotypes, we propose a single-locus association method that accounts for interaction through modeling both marginal mean and variance as functions of genotype. We derive marginal mean and variance as functions of genotype and show that the marginal variance is a quadratic function of both genotype and strength of effect modification under commonly used interaction models. Using simulations we compare our method with a test of marginal genotypic effect with constant variance and with a method recently proposed by Aschard et al. (Genetic Epidemiology, 2013). The results show our method controls type I error rate and significantly improves power over both alternatives. For example, under one scenario, our method had 68% power at the GWAS significance level p<5e-8, while linear model had 39% and that of Aschard et al. had 45%; under another scenario, our method had 91% power versus 78% and 83% for the alternatives. This new method will help detect genes that affect phenotypes mostly through interactions with effect modifiers and may help explain some missing heritability for many complex phenotypes.

54
Meta-analysis of correlated traits using summary statistics from GWAS

Tao TF Feng (1) Xiaofeng XZ Zhu (1)

(1) Department of Epidemiology and Biostatistics, Case Western Reserve University

Genome-wide association study (GWAS) is one of the important approaches to detect genetic variants underlying complex traits. Meta-analysis is often conducted to summarize the association evidences from multiple studies. When multiple correlated traits are available, analyses are often performed for each trait separately. For example, researchers often perform analysis for systolic blood pressure, diastolic blood pressure and hypertensive status separately in searching genetic variants

underlying hypertension. The Bonferroni correction is used to account for multiple tests. Such analysis procedure may reduce power when the same genetic variants contribute to the variation of correlated traits. Here we propose a novel statistical approach to perform meta-analysis of multiple correlated traits. This method is robust to population structure and correlated samples between different cohorts. Both simulation and real data analysis will be present in this report.

55
Derivation of a Genome-Wide Significance Threshold for African Populations

Mary D Fortune (1) Ioanna Tachmazidou (1) Eleftheria Zeggini (1)

(1) Wellcome Trust Sanger Institute, Hinxton, UK

Genome-wide association studies examine common variation across the genome for association with complex traits of interest. Significance is declared at the widely-accepted threshold of p<5.0E-08. This has been derived from the total number of effective common variant (minor allele frequency [MAF]>0.05) tests in European populations and has been based on HapMap data. As the GWAS field is shifting to the study of more structured and heterogeneous populations, for example of African descent, a new statistical significance level has to be defined. Lower levels of linkage disequilibrium between common variants may necessitate a more stringent threshold. In addition, the availability of sequence data further empowers the assessment of the effective number of independent tests, as common variation has been comprehensively assayed . Many methods exist which exploit the correlation structure, either haplotypic or genotypic, between the variants. We have implemented several of these on two African datasets, Luhya (LWK) and Yoruba (YRI) , from the 1000 Genomes Project (sequence data, phase I integrated public data release), in order to estimate the effective number of tests for common genetic variation (MAF over 1 or 5%) in African populations. For comparison we also used the European, CEU, dataset from the same source. Using the haplotypic correlation coefficients, as proposed by Moskvina and Schmidt resulted in an estimate of 3.0E-8 for the European dataset, and 1.15E-8 for the African datasets at MAF over 5%, and an estimate of 1.5E-8 for the European dataset, and 7.0E-9 for the African datasets at MAF over 1%. This reflects the greater genetic variation in present day sub-Saharan African populations.

56
The impact of stochastic variation in genotype imputation on genome-wide association studies

Nathan C Gaddis (1) Joshua L Levy (1) Dana B Hancock (1) Grier P Page (1) Eric O Johnson (1)

(1) RTI International

Genotype imputation is used widely in GWAS to expand the SNP coverage, integrate results from different genotyping platforms, and fine map regions of interest. There is a stochastic nature to the imputation process that causes there to be a degree of randomness in imputed genotypes. To investigate the impact of this uncertainty on the precision of downstream association testing, we conducted 100 pre-phasing/imputation replicates on chromosome 15 in African Americans (AA) (N=709) and European Americans (EA) (N=1952) using the following analysis pipeline: (1) pre-phasing observed genotypes using SHAPEIT2; (2) genotype imputation using IMPUTE2; and (3) association testing using ProbABEL. The overall mean pairwise correlation of association test p-values between replicates was high for well-imputed SNPs with info score ≥ 0.9 (r=0.97 for AAs, r=0.96 for EAs). However, for the subset of well-imputed SNPs with low p-values (median p ≤ 0.001 among replicates), the mean pairwise correlation between replicates was dramatically lower (r=0.76 for AAs, r=0.81 for EAs); p-values for these SNPs were found to vary by as much as 100 fold between replicates. The majority of this variation is introduced in the pre-phasing step. A strategy of performing repeated pre-phasing/imputation may be necessary to establish accurate association test p-value estimates for imputed SNPs, particularly those with p-values in the range typically targeted for follow-up studies.

57
The link between hepcidin, iron and atherosclerosis: a Mendelian randomization approach

Tessel E. Galesloot (1) Luc L. Janss (2) Dorine W. Swinkels (1) Sita H. Vermeulen (1)

(1) Radboud University Medical Centre, Nijmegen, The Netherlands
(2) Aarhus University, Aarhus, Denmark

Background The ‘iron hypothesis’ states that people with elevated serum iron levels face a greater risk of cardiovascular disease, but epidemiologic studies on associations between iron depletion and cardiovascular risk remain inconclusive. Hepcidin, central regulatory molecule of systemic iron homeostasis, might play a role in the progression of atherosclerosis. In this study, we will investigate the link between hepcidin, iron and atherosclerosis using a Mendelian randomization approach. Methods We will include participants of the Nijmegen Biomedical Study (NBS) aged 46-67 years for whom measurements of hepcidin, iron parameters and non-invasive measurements of atherosclerosis as well as GWAS data are available (N±800). We will derive heritability estimates for hepcidin, ferritin and iron and estimate the genetic correlation between these traits using GWAS data by application of Bayesian modeling techniques. Subsequently, we

will apply a multivariate GWAS analysis for hepcidin, ferritin and iron using a multivariate test of association (MQFAM) implemented in PLINK. We will identify both shared (pleiotropic) and non-shared (independent) genetic regulators of hepcidin, ferritin and iron. The independent genetic variants and a combination thereof will be used as instrumental variables in a Mendelian randomization approach to study their association with atherosclerosis. Expected results Our study will apply up-to-date techniques to identify independent genetic determinants of correlated phenotypes. The identified differences in genetic etiology will be exploited to create insight in the causal roles of hepcidin, ferritin and iron in the development of atherosclerosis.

58
GxEscan: Software to detect GxE interaction in a GWAS

James Gauderman (1) Pingye Zhang (1) Victor Moreno (2) John Morrison (1)

(1) University of Southern California
(2) Institut Català d'Oncologia

GxEscan performs a genomewide scan for gene-environment (GxE) interaction in a case-control sample. The program implements several efficient 2-step methods as well as more traditional (but less powerful) case-control and case-only analyses. The 'environment' factor E may be either binary or continuous and can be an exogenous exposure variable (e.g., sunlight, air pollution), personal exposure (e.g., smoking, dietary fat), or other personal characteristic (e.g., sex, age, candidate gene). GxEscan will test GxE interaction with measured and/or imputed SNPs and will control the family-wise error rate at a user-defined level (e.g. 5%). Output includes QQ-plots, Manhattan plots, and tables of top GxE interaction hits. GxEscan is computationally efficient, can be executed on a Windows, Mac, or Unix/Linux operating system, and is freely available at http://biostats.usc.edu/software.

59
Importance of controlling for prognosis when assessing differential effects of treatment

Katrina A.B. Goddard (1) Elizabeth M. Webber (1) Elizabeth O'Conner (1) Tia L. Kauffman (1)

(1) Center for Health Research, Kaiser Permanente Northwest

Pharmacogenomic or other biomarkers can be used to direct patient care towards more effective therapies and avoid harmful adverse effects or costs from ineffective treatments. Pharmacogenomic effects are some of the largest effect sizes from GWAS analyses, suggesting potential for utility in clinical application. However, interpretation can be complex if the biomarker not only predicts response to treatment, but is also a prognostic factor. In some cases this relationship is further complicated by conflicting directions in these two factors. We illustrate this complexity using an example in colorectal cancer. Microsatellite instability (MSI) may predict response to treatment with a common chemotherapy drug, 5-fluorouracil (5FU). We conducted a systematic literature review and identified over 50 published articles to assess the relationship between MSI status and response to 5FU. Three previous meta-analyses have summarized this data with apparently conflicting results. We untangle these effects to demonstrate a consistent relationship between MSI status and response to 5FU by accounting for both the prognostic and predictive effects in the meta-analysis. This is done by jointly considering the effect of treatment and MSI status on survival. We show that individuals with tumors that exhibit MSI have better prognosis, but do not respond to treatment with 5FU. In contrast, individuals with microsatellite stable (MSS) tumors do respond to treatment with 5FU. Analyses to predict response to treatment based on genotype or biomarkers should include untreated individuals to distinguish between prognostic and predictive effects..

60
Performance of Rare Variant Association Tests Under Different Simulated Scenarios

Felicia Gomez (1) Qunyuan Zhang (1) Ingrid Borecki (1)

(1) Washington University School of Medicine in St. Louis

Much of human genetic variation is low frequency or rare variants. There is great interest in assessing the role of rare variants in the architecture of complex traits. Many methods have been proposed to address the statistical challenges of rare variant analyses; however, it is unclear which tests have optimal operational characteristics under different models. We used simulated data to assess the type I error rate and power of five rare variant association tests. We included several burden tests: a fixed threshold collapsing test (T1), PWST (Zhang et al. 2011), WSS (Madsen and Browning 2009), and aSUM (Han and Pan 2010); and one non-burden test: SKAT (Wu et al. 2010). We simulated genotypes for 30-60 variants with a maf of ~ 0.002 within a locus for 1,000 subjects under the null and several alternatives models. We varied the proportion of variants with functional effects, the effect size of the variants, the effect of including a common (maf >5%) functional variant, and contrasted situations where all variants were trait raising vs. a mixture of trait raising, lowering, and neutral variants. The results of our analyses suggest that all tests have similar type I error rates corresponding to the nominal rate of 0.05. We find variation in power of each test depending on whether all markers affect the trait in the same direction, the number of neutral sites, and the inclusion of a common functional variant. Because genes affecting quantitative traits are likely to have variants that are trait raising and trait lowering, methods that account for allelic heterogeneity have advantages over methods that are not robust to this variation.

61
Estimating genome-wide significance for whole genome sequencing studies

Celia M.T. Greenwood (1) ChangJiang Xu (2) Ioanna Tachmazidou (3) Antonio Ciampi (4) Eleftheria Zeggini (3) UK10K Consortium

(1) Lady Davis Institute, Jewish General Hospital, Montreal, QC, Departments of Oncology, Epidemiology, Biostatistics and Occupational Health, and Human Genetics, McGill University, Montreal, QC
(2) Lady Davis Institute, Jewish General Hospital, Montreal, QC, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC
(3) Wellcome Trust Sanger Institute, Hinxton, UK
(4) Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC

For common genetic variants, a genome-wide significance level of 5x10-8 is in common use, derived from a Bonferroni correction for the number of uncorrelated SNPs. However, in the new era of whole genome sequencing (WGS), analysis of rare genetic variation is usually undertaken by jointly examining all genetic variability in a series of pre-defined small genomic regions or windows. In nearby or overlapping windows, these test statistics will be correlated, and the degree of correlation is likely to depend on the choice of window size, overlap, and on the test statistic. Here we propose an empirical approach for estimating genome-wide significance thresholds for data arising from WGS studies, and we contrast this with theoretical calculations based on correlations between SKAT test statistics. Since calculations may need to be repeated with different choices of test statistics or windows, we show that genome-wide significance thresholds can be estimated by extrapolating from a small portion of the genome to the whole genome, thereby reducing the amount of computation required. We recommend a genome-wide significance threshold of 1x10-08 for European populations for the combined analytic strategy of using single-SNP tests for common variants (minor allele frequency (MAF)>1%) together with rare variants (MAF under 1%) using sliding windows.

62
Quantifying missing heritability from known GWAS loci and rare coding variants

Alexander Gusev (1) Benjamin M Neale (2) Gaurav Bhatia (3) Noah Zaitlen (4) Bjarni Vilhjalmsson (1) Dorothee Diogo (5) Eli A Stahl (5) Peter K. Gregersen (6) Jane Worthington (7) Lars Klareskog (8)

(1) Harvard School of Public Health
(2) Broad Institute
(3) Harvard-MIT Division of Health, Science, and Technology
(4) UCSF
(5) Harvard Medical School
(6) North Shore-Long Island Jewish Health System
(7) University of Manchester
(8) Karolinska Institutet and Karolinska University Hospital; Soumya Raychaudhuri (Harvard Medical School); Robert M. Plenge (Harvard Medical School); Bogdan Pasaniuc (UCLA); Patrick F. Sullivan (UNC School of Medicine); Alkes L. Price (Harvard School of Public Health)

GWAS currently explain only a small fraction of disease heritability. Two possible sources of missing heritability include rare coding SNPs throughout the genome, or additional causal SNPs at known GWAS loci, each of which may be poorly tagged by top GWAS associations. We describe methods to estimate these components of heritability while adjusting for LD both within and between variant classes, which can bias variance component estimates. Our simulations on real genotypes show that our methods produce unbiased estimates. We applied the methods to two data sets: a 23,000-sample study of rheumatoid arthritis (RA) with ImmunoChip data from 10 known RA loci (excluding HLA), and a 6,400-sample study of schizophrenia typed on both exome and GWAS chips. For RA known associations at these loci (including those identified in these data) explained 0.006 of the variance in liability of RA, while including all SNPs at the loci explained 0.014±0.002, increasing further to 0.032±0.003 when including 17 known autoimmune loci not associated with RA, indicative of additional causal variants at these loci. In the schizophrenia analysis, we partitioned total heritability explained by all typed SNPs of 0.38±0.04 into 0.09±0.03 from coding SNPs and 0.30±0.03 from noncoding GWAS-chip SNPs, demonstrating significant exonic heritability. However, the contribution of 0.04±0.03 from rare coding SNPs (after adjusting for LD between variants) was non-significant, and remained so when collapsing rare variants to reduce statistical noise; the remaining contribution from common coding SNPs was largely tagged by GWAS-chip SNPs. Our results shed light on components of missing heritability for these traits and provide insights into their genetic architecture.

63
Rare Variant Extension of the Transmission Disequilibrium Test Detects Associations with Autism Exome Sequence Data

Zongxiao He (1) Brian J O'Roak (2) Joshua D Smith (2) Gao Wang (1) Stanley Hooker (1) Regie L Santos-Cortez (1) Biao Li (1) Mengyuan Kan (1) Nik Krumm (2) Deborah A Nickerson (2) Evan E. Eichler (2) Suzanne M. Leal (1)

(1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
(2) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA

Many population-based rare variant (RV) association tests, that aggregate variants across a region, have been developed to analyze sequence data. A drawback of analyzing population-based data is that it is difficult to adequately control for population substructure/admixture and spurious associations can occur. For RV this problem can be substantial, because the spectrum of rare variation can differ greatly between populations. A solution is to analyze trio data, parents and a proband, using the transmission disequilibrium test (TDT), which is robust to population substructure/admixture. We extended the TDT to test for RV associations using four commonly used methods. We demonstrate that for all RV-TDT tests using proper analysis strategies, type I error is well controlled even when there are high levels of population substructure or admixture. The power of the RV-TDT tests were evaluated and compared to the analysis of case-control data using a number of different genetic and disease models. The RV-TDT was also used to analyze exome data from 199 Simons Simplex Collection autism trios and an association was observed with the ABCA7 gene. Given the problem of adequately controlling for population substructure/admixture in RV association studies and the growing number of sequenced based trio studies the RV-TDT is extremely beneficial to elucidate the involvement of RVs in the etiology of complex traits.

64
Differential Admixture in Latin American Populations and its Impact on Genetic Structure.

Pedro C. Hidalgo (1) Valentina Colistro (1) Patricia Mut (1) Monica Sans (1)

(1) Departamento de Antropologia Biologica. Facultad de Humanidades. Universidad de la Republica. Uruguay.

Admixture mapping is an alternative approach for analyzing data when population heterogeneity is clear. The process of admixture in the Americas can be seen nowadays as a natural experiment for genetic epidemiology and anthropology, in which polymorphic marker loci are used to infer a genetic basis for traits of interest. Latin American populations are supposed to be ideal for admixture mapping scans (AMSs) with ancestry informative markers (AIMs), as the admixture does not have more than 500 years. But the admixture process is far from being simple and homogeneous in Latin America, and the process itself depending on the characteristics of the populations and regions. We analyze four admixed American populations: Afro-Americans (ASW), Colombians (CLM), Puerto Ricans (PUR), and Mexicans (MEX), each with different proportions of ancestry (European, Native Americans and Africans) in 10 DNA regions: APC, BRAF, MSH2, MSH6, MLH1, MUTYH, PMS2 and three genomic positions: 8q23.3, 16q22.1 and 19q13.11. We used data available in 1000Genome. The AIMs' panel was composed of 307 AIMs developed for the study of Latino American populations (Lace Consortium). We could not find any pattern followed by all admixed populations. Different regions of a same region (exon, intron, 3'…) have different results in different populations. The percentage of Non-Synonymous sites as well as polymorphisms in general varies in admix populations. Linkage disequilibrium blocks are smaller and very variable in "mixed" Latin America populations. LD blocks show similarities (e.g. Colombians and Africans) that not coincide with genetic distances (g.e. Colombians in European clusters). Funded by a EU: HEALTH-2007-2.4.1-14: Studying cancer aetiology in Latin America.

65
Identification of genetic models that predict LDL-C traits using genotype and gene expression data

Emily R Holzinger (1) Scott M Dudek (2) Ronald M Krauss (3) Marisa W Medina (3) Alex T Frase (2) Marylyn D Ritchie (2)

(1) Inherited Disease Research Branch (NIH/NHGRI)
(2) Pennsylvania State University
(3) Children's Hospital Oakland Research Institute

High levels of low-density lipoprotein cholesterol (LDL-C) in the plasma are a major risk factor for cardiovascular disease (CD), the primary cause of death in developed countries. HMG-CoA reductase (HMGCR) inhibitors, or statins, substantially lower LDL-C levels and CD risk. However, statin efficacy is variable from person to person due to environmental and genetic factors. Most genetic analyses of LDL-C levels and statin efficacy have focused on DNA variation. For our study, we combined SNPs and gene expression variables (EVs) to identify variants that contribute to inter-individual variation of LDL-C traits. We conducted the study in various ways to address two main objectives: Obj. 1: To identify EVs and SNPs associated with baseline LDL-C levels. Obj. 2: To identify EVs and SNPs associated with the change in LDL-C before and after simvastatin treatment. For all analyses, we applied a filtering-modeling pipeline using machine learning methods in the ATHENA software package. SNPs and EVs were filtered using a tree-based variable selection method. We used the top filtered variables as inputs for a modeling method to generate parsimonious prediction models. The best performing models for Obj. 1 and Obj. 2 had R-squared values of 0.13 and 0.25, respectively. We assessed the potential biological roles of these models using the functional annotation tool DAVID. One of the top functional clusters included the gene ABCA1, which has previously been associated with statin efficacy. While these models will need to be validated in independent datasets, we were able to combine genomic and transcriptomic data to generate multi-variable models that explain a proportion of the lipid trait variation in this cohort.

66
Genome-wide association study for systolic blood pressure in Brazilian families: the Baependi Heart Study

Andrea RVR Horimoto (1) Nubia E Duarte (1) Silvano C Costa (2) Julia P Soler (3) Mariza de Andrade (4) Jose E Krieger (1) Alexandre C Pereira (1)

(1) Heart Institute, University of Sao Paulo
(2) Exact Sciences Center, State University of Londrina
(3) Mathematics and Statistics Institute, University of Sao Paulo
(4) Department of Health Sciences Research, Mayo Clinic

Background: Hypertension is an common and heritable cardiovascular risk factor. To date, common genetic variants influencing blood pressure in the Brazilian population are unknown. To identify the polygenic basis of this trait, we conducted genome-wide association analyses for systolic blood pressure (SBP).

Methods: The data set consisted of 1,120 subjects distributed in 95 families enrolled in the Baependi Heart Study. These individuals were genotyped in almost 1 million SNPs, using Genome-wide Human SNP Array 6.0 chip (Affymetrix). Systolic blood pressure was calculated as a mean value from three readings. Genotype-phenotype associations of SBP was carried out under polygenic mixed models. Principal components analysis was performed using Eigensoft package.

Results: We found 33 associated markers to SBP under model non-adjusted by principal components, and 1 associated marker when the model was adjusted by principal components at $P < 6.2x10-8$. We are conducting replication analyses to confirm that association.

67
Genome Analysis of Nucleotide Sequences of Novel Avian A(H7N9) Influenza Viruses

Hsin-Hsiung HH Huang (1) Jie JY Yang (1)

(1) University of Illinois at Chicago

Natural vector is a novel way to map nucleotide sequences into R^12 space and we can use it to classify viral nucleotide sequences. Our study shows that the natural vector successfully predicts viral classification information, as well as to identify viral origins. There are 2410 multiple segmented virus sequences in the GenBank collection. We apply our method to analyze H7N9 influenza viruses based on their whole genomes. Up to November 2012, NCBI kept reference sequences for 2410 viruses. Among them, there are 366 viruses containing 2 or more segments. For example, H7N9 virus contains 8 segments. Since February 2013, novel strains of avian-origin influenza A (H7N9) virus have emerged in clinical patients in Zhejiang, China. It attracts many scientists to find the origins of the novel H7N9. We use Natural Vector to find the closest neighbors of these strains. The first neighbor of Influenza A/Zhejiang/DTID-ZJU01/2013(H7N9) is Influenza A/Anas crecca/Spain/1460/2008(H7N9) and the second neighbor is Influenza A/goose/Czech Republic/1848-T14/2009(H7N9). The first neighbor of Influenza A/chicken/Zhejiang/DTID-ZJU01/2013(H7N9) is Influenza A/wild bird/Korea/A3/2011(H7N9) and the second neighbor is Influenza A/spot-billed duck/Korea/447/2011(H7N9). These two Zhejiang H7N9 strains are not close to each other regarding to the multiple segmented results, but they are close in terms of segment-by-segment result. Their M and NS segments are close to each other, so that their M and NS genes probably originate from the same viruses individually.

68
Utilizing Population controls in Rare-Variant Case-Parent Association Tests

Yu Jiang (1) Glen A Satten (2) Yujun Han (3) Erin L Heinzen (3) David B Goldstein (3) Andrew S Allen (1)

(1) Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA
(2) Centers for Disease Control and Prevention, Atlanta, GA 30333, USA
(3) Center for Human Genome Variation, Duke University School of Medicine, Durham, NC 27708, USA

There is a great deal of current interest in detecting associations between human traits and rare genetic variation. To address the low power implicit in single locus tests of rare genetic variants, many rare-variant association approaches attempt to accumulate information across a gene or other genetic unit, often by taking a weighted combination of individual loci contributions to a score or other statistic. Using the 'right' weighted combination is key—an optimal test will up-weight true causal variants, down-weight neutral variants, and will correctly assign the direction of the effect for causal variants. As the true causal loci are unknown, much of the current rare-variant association literature involves identifying flexible approaches to weighting individual loci. Here, we propose a procedure that exploits population controls to estimate the weights to be used in a case-parent rare variant association test. Specifically, we form weights by comparing population control allele frequencies with the allele frequencies in the parents of affected offspring. These weights are then used in a rare-variant transmission distortion test (rvTDT) in the case-parent data. Since the rvTDT is conditional on the parents' data, its use in estimating the rvTDT weights does not affect the validity or asymptotic distribution of the rvTDT. Using simulation, we show that our new control-weighted rvTDT can significantly improve power over rvTDTs that do not use population control information. It also remains valid

under population stratification. We illustrate the approach with an application to childhood epileptic encephalopathy.

69
Robust Rare Variant Association Testing in Samples with Related Individuals

Duo Jiang (1) Mary Sara McPeek (1)

(1) The University of Chicago

The recent development of high-throughput sequencing technologies calls for powerful statistical tests to detect rare genetic variants associated with complex human traits. Sampling related individuals in sequencing studies offers advantages over sampling unrelateds only, including improved protection against sequencing error, the ability to use imputation to make more efficient use of sequence data, and the possibility of power boost due to more observed copies of extremely rare alleles among relatives. With related individuals, familial correlation needs to be accounted for to ensure correct control over type I error and to improve power. Recognizing the limitations of existing rare-variant association tests for related individuals, we propose MONSTER, a robust approach to detecting associations, which adaptively adjusts to the unknown configuration of effects of rare-variant sites. MONSTER also offers an analytical way of assessing p-values, which is desirable because permutation is not straightforward to conduct in related samples. We further propose a pathway-based association test in sequencing studies that exploits the hierarchical structure of gene pathways. We show, through simulation with a wide range of trait models, that MONSTER effectively accounts for family structure, is computationally efficient and compares very favorably, in terms of power, to previously-proposed tests that allow related individuals. We further illustrate the proposed approach using a candidate gene study for high-density lipoprotein cholesterol, where we are able to replicate association with three genes previously linked to the trait.

70
Genome-Wide Association Study of Illicit Drug Abuse: An Example of Using Public Controls from dbGaP

Eric O Johnson (1) Joshua L Levy (2) Nathan C Gaddis (2) Cristie Glasheen (1) Dana B Hancock (1)

(1) Behavioral Health Epidemiology Program, Research Triangle Institute, Research Triangle Park, NC
(2) Research Computing Division, Research Triangle Institute, Research Triangle Park, NC

Abuse of illicit drugs is heritable, but there are no well-established genetic risk factors. We conducted a genome-wide association study (GWAS) of illicit drug abuse using African American and Caucasian cases genotyped on the Illumina Omni1-Quad BeadChip. To assemble a multi-ethnic control set, we used Illumina-genotyped cohorts in the database of Genotypes and Phenotypes (dbGaP). We applied standard quality control to subjects and single nucleotide polymorphisms (SNPs) and compared the potential control cohorts by running GWAS models with arbitrary case/control assignments. Six cohorts had $\lambda<1.05$ for all pair-wise comparisons. We combined 8,753 controls from these 6 cohorts with our 3,159 drug abuse cases and used only SNPs available on all subjects as input genotypes for 1000 Genomes imputation. We ran ethnic-specific GWAS adjusting for 3 ancestry principal components and applied genomic control before combining the ethnic-specific results in a GWAS meta-analysis ($\lambda=1.03$). We identified a significantly associated SNP ($P=4.3x10-8$) that spans a novel chromosome 20 region with evidence for independent replication in two cohorts of illicit drug abuse cases and their own study controls (total N=2,871). This study shows the utility of using dbGaP-derived public controls to conduct a GWAS when only study cases are available and outlines thorough procedures to enrich its validity.

71
Harnessing Web 2.0 Social Networks for Genetic Epidemiology

Joanna Kaplanis (1) Pratheek Nagaraj (2) Barak Markus (1) Daniel MacArthur (3) Alkes Price (4) Yaniv Erlich (1)

(1) Whitehead Institute for Biomedical Research
(2) MIT
(3) Broad Institute
(4) Harvard School of Public Health

Understanding the genetic architecture of complex traits is one of the top missions of human genetics. Emerging lines of studies have highlighted the entangled etiologies of these traits, which can include epistasis, parent-of-origin effects, sex and age interactions, and environmental risk factors. To conduct robust genetic epidemiological analysis, statistical models require sampling substantial amount of data from large families. However, the recruitment of large cohorts of extended kinships is both logistically challenging and cost-prohibitive. We suggest a Big Data strategy to address this challenge: harnessing existing, free, and massive Web 2.0 social network resources to trace the aggregation of complex traits in millions of people and extremely large families. We collected millions of profiles from Geni.com, a genealogy-driven social network (with permission from MIT's IRB and Geni.com). Using this information, we constructed a single pedigree of 13 million individuals spanning many generations up to the 15th century, providing a range of kinships for familial aggregation. In addition, we used Natural Language Processing to convert genealogical information into birth and death locations to obtain a proxy for environmental factors. I will describe this resource which we aim to publish as community resource, the

range of potential phenotypes that can be measured, and analysis of familial aggregation studies for longevity and facial morphology using crowd sourcing approach on Mechanical Turk.

72
Rules for resolving Mendelian inconsistencies in nuclear pedigrees typed for two-allele markers

Sajjad Ahmad Khan

Gene-mapping studies regularly rely on examination for Mendelian transmission of marker alleles in a pedigree, as a way of screening for genotyping errors and mutations. For analysis of family data sets, it is usually necessary to resolve or remove the genotyping errors prior to analysis. At the Center of Inherited Disease Research (CIDR), to deal with their large-scale data flow, they formalized their data cleaning approach in a set of rules based on PedCheck output. We examine via carefully designed simulations that how well CIDR's data cleaning rules work in practice. We found that genotype errors in siblings are detected more often than in parents for less polymorphic SNPs and vice versa for more polymorphic SNPs. Through computer simulation, we conclude that some of the CIDR's rules work poorly in some situations and we suggest a set of modified data cleaning rules that may work better than CIDR's rules.

73
Localising disease regions in genomewide association studies using nonparametric regression models

Pianpool Kirdwichai (1) Fazil Baksh (1)

(1) University of Reading

Although many complex diseases are suspected to be the result of the cumulative action of many loci, each having a small effect, current analysis methodology for genomewide association studies tend to only reliably identify genomic regions with very strong signals of disease gene association. As a result only these regions are taken forward as potential locations of disease genes in most studies. To address this problem, we develop and evaluate a novel method, based on nonparametric regression, which is capable of identifying candidate regions with moderate and weak signals of disease-gene association in data from high dimensional genomewide studies. Our proposed method inherently accounts for the linkage disequilibrium structure in the data through a tuning parameter and assigned weights. A computationally efficient method for obtaining the optimal tuning parameter is proposed and evaluated using both theory and simulations. Results of extensive evaluation and comparison with existing methods show that the nonparametric approach is not only powerful but also leads to substantial reduction in false positive findings. The method is illustrated using data from the Wellcome Trust Case Control Consortium study (2007) of Crohn's disease.

74
Increased Proportion of African Ancestry in Helicobacter pylori Associates with Histopathological Severity in Hosts with High Amerindian Ancestry

Nuri Kodaman (1) Alvaro Pazos (2) Barbara G Schneider (1) M Blanca Piazuelo (1) Rafal Sobota (1) Carrie L Shaffer (1) Keith T Wilson (1) Timothy L Cover (1) Scott M Williams (1) Pelayo Correa (1)

(1) Vanderbilt University
(2) Universidad de Nariño

Helicobacter pylori is the principal cause of gastric cancer, the second leading cause of cancer mortality worldwide. Over 90% of individuals in some Colombian populations are infected with H. pylori, but infection rate does not generally predict cancer prevalence. In particular, residents of the Andean region are 25 times more likely to develop gastric cancer than their coastal counterparts, despite similar rates of infection. We determined the ancestry of H. pylori isolates from the gastric biopsies of 275 Colombian subjects from both the mountain and coastal regions. Most isolates contained genomic regions from four ancestral H. pylori populations: Africa1 (AA1), Europe1 (AE1), Europe2 (AE2), and East Asia (AEA), but these proportions varied with geography. The AA1 cluster was more common in coastal samples (mean=47.9%), and AE2 in mountain samples (mean=50.7%). The human ancestry of the biopsied individuals also varied with geography, with mean proportions of 19.5% European, 57.9% African, and 22.6% Amerindian ancestry in the coastal region, and 30.4% European and 67% Amerindian ancestry in the mountain region. All pairwise correlations between H. pylori ancestry and human host ancestry were significant. Although African H. pylori ancestry correlated negatively with Amerindian human ancestry (r=-0.60), higher proportions of African H. pylori ancestry in individuals of mostly Amerindian descent associated with progression of gastric disease. The significant interaction between H. pylori ancestry and host ancestry indicates that co-evolution has modulated disease risk, and that the disruption of this relationship may account for the risk discrepancy in Colombian populations.

75
CpG methylation and associated phenotypes: Exploring mQTLs in disease

Tess Korthout (1) Pamela Herschberger (1) Marjorie Romkes (2) Jill M Siefried (2) Jianmin Wang (1) Leah Preus (1) Sebastiano Battaglia (1) Song Liu (1) Moray J Campbell (1) Lara E Sucheston-Campbell (1)

(1) Roswell Park Cancer Institute
(2) University of Pittsburgh

Background: In this study we measure SNP-methylation relationships across phenotypes and replicate our significant findings in a cohort of lung cancer patients. Methods: Using a boot-strapping approach we determine if there are phenotypes with an over-representation of significant genome wide associated (GWA) SNPs (taken from the NHGRI GWAS catalogue) in CpG island regions. In two of these over-represented phenotypes (lung cancer and smoking behavior) we measure the correlation between alleles in SNPs and methylation probe levels. Specifically all lung phenotype GWA SNPs and all SNPs in linkage disequilibrium (LD) with these GWA SNPs (within 100kbp of methylation regions) were tested for association with methylation beta values measured in a 27K array in a lung cancer cohort (43 tumor-normal pairs, 46 tumor only and 24 normal only). Results: Over twice as many GWA SNPs are in CpG island regions than expected and 14/369 phenotypes have GWA SNPs in CpG island regions more often than expected ($p<1e-05$). In the lung cancer cohort, our most significant methylation associated SNP, rs707974, resides 8979 bp from rs3117582 (a lung cancer GWAS SNP) and shows a positive association between the G allele and higher levels of BAT3 methylation in lung adenocarcinoma tumor versus normal ($p<.003$). Conclusions: We found GWA SNPs much more frequently in methylation regions than expected. BAT3, which controls apoptosis, has been shown to be associated with lung cancer and we demonstrated that elevated methylation levels in tumor (versus normal) are associated with a nearby SNP in tight LD with a lung GWA SNP. This finding may help explain the strong statistical association of this SNP with lung adenocarcinoma.

76
Pilot whole genome sequencing of germline DNA from 186 breast cancer cases

Peter Kraft (1) Jamie Allen (2) Constance Chen (1) Brennan Decker (2) Jonine Figueroa (3) Steven Hart (4) Sara Lindstrom (1) Jirong Long (5) Meredith Yeager (5) Stephen Chanock (5) Fergus Couch (4) Douglas Easton (2) Christopher Haiman (6) Wei Zheng (5) David J. Hunter (1)

(1) Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA
(2) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
(3) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA
(4) Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
(5) Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
(6) Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

High coverage whole genome sequencing (WGS) has been proposed to identify rare germline mutations associated with complex traits in both coding and regulatory regions. However, WGS studies present many logistical and analytic challenges, including coordinating variant calling and cleaning across multiple sites, accruing sufficient sample size, and designing and interpreting appropriate statistical analyses. We illustrate some of these challenges using a pilot whole genome sequencing study of 143 European-ancestry (EA), early-onset, family-history-positive breast cancer cases, 21 Asian cases and 25 African-American cases from six studies participating in the DRIVE consortium, an NCI-sponsored post-GWAS initiative. Samples from each study were sequenced separately using Illumina HiSeq to an average depth of 30x and called individually using CASAVA. Preliminary analyses identified a missense BRCA1 mutation present in three EA cases that was absent in 4,300 EA subjects in the Exome Sequencing Project ($p=6.8\times10\text{-}5$). A nonsense variant in another gene was present in 11 cases and 0 ESP subjects ($p=5\times10\text{-}16$), but this apparent association was driven by false negative calls in the ESP--the average read depth at this position in the ESP was 1. We describe and present results from a rare-variant burden that uses summary data from individual studies. We discuss future plans, including combined variant calling and analysis of non-coding variants.

77
Obtaining average correlation matrix estimates in incomplete data to optimize pleiotropy predictions

Aldi T. Kraja (1) Michael A. Province (1) XCP-PMI-WG (1) Ingrid B. Borecki (1)

(1) Division of Statistical Genomics, Washington University School of Medicine

Increasing interest in the pleiotropic effects of loci identified via meta-analysis of GWAS results from multiple sources raises the issue of how to focus analytic efforts to elucidate the complex relationship between multiple genetic predictors and multiple correlated phenotypes. Varying numbers of traits and missing value patterns across studies pose challenges in characterizing the overall covariance relationships among these many phenotypes. This work is motivated by a project to identify pleiotropic SNPs contributing to the correlated architecture of 8 metabolic traits and 9 inflammatory biomarkers using data from 14 cohorts, with sample sizes from 950 to 23,000 participants (N total=85,523) within the Cross Consortium Pleiotropy Working Group (XCP). Data were simulated for each component study under a multivariate correlation model based on reported study-wise correlations

and sample sizes, and average correlations computed in the simulated data pooled across studies. This approach is contrasted with combining correlations with Fisher's Z-transform. With the simulation approach, factor analyses were carried out. Ten clusters of metabolic traits and inflammatory markers were predicted for pleiotropy analysis, narrowing the field from 130,816 possible of the same trait combinations. Focusing on these clusters, we carry out correlated meta-analysis of GWAS published results for these combinations of phenotypes to identify pleiotropic genetic variants influencing metabolic and inflammatory traits.

78
Combined influence of SNPs at 8q24 on predisposition to Prostate Cancer in African American Men.

M. Hubert Kuete (1) H. Danawi (1) S. Morrell (1)

(1) Walden University, 100 Washington Ave. S #900, Minneapolis, MN, 55401,USA

Background: African Americans are disproportionally impacted by prostate cancer, with an average incidence rate of 236.0 per 100,000 men across all years from 2005 to 2009, and an average mortality rate of 53.1 per 100,000 men across all years from 2005 to 2009. There is some evidence for an association of Single Nucleotide Polymorphisms (SNPs) located on regions of chromosome 8q24 with the risk for prostate cancer, but this had not been fully evaluated in the African American population.
Methods: Here we assess if an increasing number of SNPs at 8q24, increased the susceptibility to prostate cancer in African Americans. The study design was a case control using the dataset from the Men of African Descent and Carcinoma of the Prostate (MADCaP) consortium. Fifteen 8q24 SNPs were evaluated from cases with prostate cancer (n=3,253) and controls without prostate cancer (n=3,012) for their individual and combined association with prostate cancer.
Results: Eight out of the 15 SNPs were significantly associated with prostate cancer in a single SNP analysis prospective. The most statistically significant SNP from each of the four 8q24 regions was selected and included in a logistic regression analysis. Men with four of the selected 8q24 associated genotypes had an odds ratio for prostate cancer of 7.76 ($p< 0.0001$), as compared with men with no 8q24 associated genotypes. Conclusion: African Americans with four 8q24 associated genetic variants have a cumulative and higher risk of developing prostate cancer compared to those with fewer or no genetic variants at 8q24.

79
Local genetic population matching for genome-wide association studies

André Lacour (1) Tim Becker (1)

(1) German Center for Neurodegenerative Deseases (DZNE), Bonn

Population stratification in samples of genome-wide association and sequenced studies can give rise to large uncertainties in the results of statistical tests. In this work, we propose a novel framework to consider chromosomal-region-specific population matchings. We employ pairwise/groupwise optimal case-control matchings as well as a hierarchical clustering, both based on a genetic similarity score matrix. In order to ensure that the resulting matches obtained from the matching algorithm capture correctly the population structure, we propose two stratum validation methods. We also present a crucial extension to the Cochran--Armitage Trend test, which explicitly takes into account the particular population structure. Association P-values are obtained by within-structure case-control permutations. We assess our framework by simulations of genotype data under the null hypothesis and by a power study. The results are compared with the prominent principal components approach.

80
Regularized Rare Variant Enrichment Analysis for Case-Control Exome Sequencing Data

Nicholas B Larson (1) Daniel J Schaid (1)

(1) Mayo Clinic

Rare variants have recently garnered an immense amount of attention in genetic association analysis. However, unlike methods traditionally used for single marker analysis in GWAS, rare variant analysis often requires some method of aggregation, since single marker approaches are poorly powered for typical sequencing study sample sizes. Advancements in sequencing technologies have rendered next-generation sequencing platforms a realistic alternative to traditional genotyping arrays. Exome sequencing in particular not only provides base-level resolution of genetic coding regions, but also a natural paradigm for aggregation via exons, which may correlate with functional domains in the corresponding coding proteins. Here we propose the use of the sparse group LASSO in combination with aggregation strategies to identify rare variant enrichment in exome sequencing data. By using exon membership as a grouping variable, the sparse group LASSO can be used as a gene-centric analysis of rare variants while also providing a penalized approach toward identifying specific regions of interest. We use simulation studies to evaluate the performance of our approach, including false discovery and sensitivity. Finally, we discuss advantages of the sparse group LASSO on exome sequencing data and outline future applications.

81
Pathway-based analysis for GWAS using the extended propensity score method

Un Jung Lee (1) Stephen J. SJ Finch (1)

(1) Stony Brook University

Many complex diseases are influenced by genetic variations in multiple genes and non-genetic factors. In order to find the association between SNPs and disease, an extension of genomic propensity score (eGPS) (Zhao et al., 2012) was used to correct for bias due to both genetic and non-genetic factors. Pathway analysis, which identifies biological pathway associated with disease outcome, was also used here to improve the power of eGPS. I hypothesize that the type I error rate of this approach will be closer to its nominal value over a wide range of null conditions and that its power will be greater than the sum statistic and principle component analysis based on the simulation study reported here.

82
Evaluation of Classical HLA Allele Prediction Methods in a Sample of European Americans and African Americans

Albert M Levin (1) Indra Adrianto (2) Indrani Datta (1) Michael C Iannuzzi (3) Sheri Trudeau (1) Paul McKeigue (4) Courtney G Montgomery (2) Benjamin A Rybicki (1)

(1) Department of Public Health Sciences, Henry Ford Health System, Detroit, MI
(2) Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK
(3) Department of Medicine, Upstate Medical University, Syracuse, NY
(4) Public Health Sciences Section, University of Edinburgh Medical School, Edinburgh, Scotland

Classical human leukocyte antigen (HLA) allele genotyping remains costly and labor-intensive. This has motivated the development of imputation methods that use genome-wide single nucleotide genotype data and the haplotype structure across this region. While these methods appear to work well in populations of European origin, comparative performance in admixed populations (such as African Americans) has not been evaluated. Using a sample of 480 European American (EA) and 325 African American (AA) individuals from A Case-Control Etiologic Study of Sarcoidosis (ACCESS), we compared results of imputation based on genotype and sequencing data to existing allele data for three HLA class II genes (HLA–DRB1, –DPB1, and –DQB1). In EAs, the newly developed HLA Genotype Imputation with Attribute Bagging (HIBAG) method outperformed the established HLA*IMP approach, with all two- and four-digit allele prediction accuracies exceeding 90%. Among AAs, the classification accuracy for HLA*IMP was higher than the pre-built HIBAG model and BEAGLE model based on the ACCESS AA data but was lower than HIBAG models built using the ACCCESS AA samples. HIBAG ACCESS AA models were significantly less accurate in individuals heterozygous for local West African ancestry ($p \leq 0.04$), but accuracy improved in models including equal numbers of West African and European chromosomes. Including additional variants by SNP imputation and targeted sequencing further improved both overall imputation accuracy and the percentage of high quality calls. Our findings suggest that combining the HIBAG approach with local ancestry and dense variant data can produce highly-accurate classical HLA allele imputation in AAs.

83
Modified Random Forest Algorithms For Analysis of Matched Case Control Data or Case-parent Trio Data

Qing Li (1) Joan E Bailey-Wilson (1)

(1) National Human Genome Research Institute/NIH

Random forests (RF) is a machine-learning method useful to detect complex interactions among genetic markers related to a disease trait based on case-control samples. We propose a new modification of the RF algorithm for matched case-control, or family based (trio) data analysis. RF is an ensemble method, which analyzes data and summarizes results using a large number of classification trees. During the procedure, each classification tree uses a proportion of samples and a subset of predictors. An R package, rpart, has functions implementing classification tree analysis and it can be modified to accommodate different study designs by substituting its functions of classification based on a novel criterion. For ease of implementation, our method utilizes the rpart package to conduct classification tree analysis on a subset of the samples and predictors. Then our ensemble code, also written in R, summarizes results from all trees. For matched case-control, or case-parent trio data, we sample the set of samples (in a matched set, or matched case, pseudo-controls set) to be fit to each classification tree. Different classification criteria are also proposed to accommodate the matched study design. To evaluate our method, we simulated matched case-control, and case-parent trio data, and applied our method to select the top 1% most important predictors. The results are compared with other machine-learning methods applicable for matched case-control data, including RF++, MDR, and trio Logic Regression.

84
A robust and computationally efficient way to integrate bioinformatics and Omics information in large-scale association studies

Dalin DL Li (1)

(1) Cedars-Sinai Medical Center

With the rapid development of bioinformatics and Omics, many datasets with indication or prediction on the function of genes/SNPs are becoming available. Those pieces of information, which can be highly informative, are not utilized in most genome-wide or sequence-based association studies. And although several approaches have been proposed to integrate those in genetic analysis, many of those approaches are cumbersome which makes them impractical for large-scale data. Moreover, many proposed approaches treat the information such as predicted function as a prior, as a consequence those approaches are sensitive to the accuracy and relevance of the bioinformatics and Omics data. Here we propose an alternative way to integrate bioinformatics and Omics information in genetic analysis, in which the biological information is used to individually adjust the significance threshold for each gene/SNP. With a restrain on the sum of the significance thresholds across the genome, the family-wise type I error rate is retained. Our simulation shows that when the independent bioinformatics/biological data are informative, this approach is more powerful than traditional analysis which treats all the genes/SNPs equally, with 5-20% higher power depending on the underlying parameters. While when the independent data are completely non-informative, only slight decrease in power is observed and more importantly, with no inflated type I error rate. This approach can be viewed as a fusion of the hypothesis driven candidate-gene based strategy and hypothesis free genome-wide strategy in a semi-Bayesian way, and can be useful in future integration of bioinformatics, genomics and other Omics.

85
A Bayesian Hierarchical Quantile Regression Model to Prioritize GWAS Results

Wei E Liang (1) David V Conti (1)

(1) University of Southern California

Genome-wide association studies have been a standard method for disease gene/variant discovery in the past decade. However, for common diseases, only a moderate amount of heritability has been explained by the SNPs declared genomewide statistically significant. Rather than a resource-limited approach of increasing the sample size, one alternative approach is to find causal SNPs within the lower Manhattan—the SNPs that just failed to achieve genome-wide significance. Thus, to improve efficiency of GWAS results, we propose a Bayesian hierarchical quantile regression model to incorporate external information with the aim of improving the ranking of causal SNPs. The proposed model examines the extremes of the p-value distribution by adapting a Normal-Exponential mixture presentation of asymmetric Laplace distribution as a prior distribution, which enables us to build an efficient Gibbs sampler. Simulation results show that the proposed model improves the ranking of causal SNPs if the external information is informative (associated with the causality of a SNP) and does not decrease the causal SNP's ranking if the external information is non-informative. We compare this approach to several alternatives, including a filtering framework, and demonstrate that these approaches can worsen the ranking of causal SNPs if the external information is not informative.

86
A Generalized Genetic Random Field Method for Genetic Association Analysis of Sequencing Data

Ming Li (1) Zihuai He (2) Robert C Elston (3) Min Zhang (2) Xiaowei Zhan (2) Changshuai Wei (4) Qing Lu (4)

(1) University of Arkansas for Medical Sciences
(2) University of Michigan
(3) Case Western Reserve University
(4) Michigan State University

With the advance of high-throughput sequencing technologies, it becomes feasible to investigate the influence of the entire spectrum of sequencing variations to complex human diseases. While association studies utilizing the new sequencing technologies hold great promise to unravel novel genetic variants, especially rare variants that contribute to human diseases, the statistical analysis of sequencing data remains great challenge. Advanced analytical methods are in great need to facilitate high-dimensional sequencing data analysis. In this article, we propose a Generalized Genetic Random Field (GGRF) method for association analysis of sequencing data. Similar to other similarity-based methods (e.g., SIMreg and SKAT), the new method has the advantages of avoiding specifying thresholds for rare variants and allowing for testing multiple variants with different directions. Moreover, the method is built on the generalized estimating equation (GEE) framework and thus accommodates to a variety of disease phenotypes (e.g., quantitative and binary phenotypes). Through simulations, we demonstrated that the proposed GGRF attained an improved power over SKAT, under various disease scenarios, especially when the rare variants make more contribution to phenotypes than common variants, or when the proportion of noise variants is low. Compared to SKAT, GGRF also has a nice asymptotic property, and can be applied to small-scale sequencing data without small-sample adjustment. We further illustrated GGRF with an application to a real data from Dallas Heart Study. GGRF was able to detect the association between two candidate genes (i.e. ANGPTL3 and ANGPTL4) and serum triglyceride phenotypes with significant levels higher than SKAT.

87
RNA-Seq Analysis of Alternative Splicing Events in Drosophila melanogaster

Yafang Li (1) Xiayu Rao (2) Chris Amos (3) Bin Liu (2)

(1) The Center for Genetics and Genomics, University of Texas MD Anderson Cancer Center; Center for Genomic Medicine, Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College
(2) The Center for Genetics and Genomics, University of Texas MD Anderson Cancer Center
(3) Center for Genomic Medicine, Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College

Alternative splicing is an important biological process in the generation of derivative functional transcripts with same genomic sequences. RNA-seq technology has been widely used for researches in transcriptomics. The profile of gene differential expression and alternative splicing can be analyzed by the various statistical analytical methods. In this article, we are reporting a protocol for differential analysis of splicing junctions and intron retentions. The DEXSeq, an R bioconductor originally designed for gene exons expression analysis (Anders S, Reyes A, Huber W, Genome Res, 2012 Oct; 22(10):2008-2017), was adopted on splicing junctions and intron retentions analysis. The design of reference files for splicing junctions and intron retentions was introduced to fulfill the DEXSeq analysis. To evaluate the protocol, the public RNA-seq datasets generated from the Drosophila melanogaster S2-DRSC cells with RNAi depletion of RNA binding proteins (Brooks AN et al., Genome Res 2011 Feb; 21(2):193-202) have been analyzed. With the further study on the differential analysis of splicing junctions and intron retentions, sequence motifs were identified at the flanking sequence of significant splicing junctions and intron retentions. Ingenuity Pathway Analysis (IPA) of the vertebrate homologs of the genes with significant splicing events may provide functional information. The results from this study show that DEXSeq package can be applied on alternative splicing analysis on RNA-seq data; conserved DNA sequence motifs may imply the possible roles of the RNA binding proteins during the splicing events.

88 WITHDRAWN

89
Mutational enrichment of cancer-related gene sets in 11 aggressive prostate cancers

Karla J Lindquist (1) Remi Kazma (1) Thomas J Hoffmann (1) Benjamin A Rybicki (2) Albert Levin (2) Pamela L Paris (1) John S Witte (1)

(1) UCSF
(2) Henry Ford Clinic

Prostate cancer is a leading cause of cancer mortality. Many prostate tumors are benign, but some are aggressive and lethal. The mutation profile of aggressive prostate tumors may differ from that of other tumors. To investigate whether somatic mutations in aggressive prostate tumors are more frequent in gene sets previously identified as functionally involved in benign prostate and other cancers, we sequenced the tumor and matched normal tissues of 11 aggressive prostate cancer patients using Complete Genomics' platform. After removing low-confidence calls, we selected mutations within any gene or gene regulatory region (using data from the Encyclopedia of DNA Elements) in the human genome. Then, we determined if 22 gene sets were associated with mutation rates using a mixed-effects Poisson regression model. The gene sets included 18 cancer-related pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG), 2 differential expression indicators from the Prostate Expression Database, and 2 indicators of cancer drug sensitivity or resistance from the Genomics of Drug Sensitivity in Cancer database. Membership in the KEGG prostate cancer pathway was independently associated with higher mutation rates in our samples ($p=0.024$), but stronger predictors of high mutation rates were the KEGG transcriptional misregulation pathway ($p<0.001$), differential expression in response to androgen ($p=0.001$), and two other KEGG cancer pathways (non-small cell lung and endometrial cancers, $p=0.001$ for both). Our work paves the way for future studies to link the mutational enrichment of cancer-related gene sets with the aggressiveness of prostate cancers.

90
False starts and missed opportunities: the importance of good annotation for WGS

Sara Lindstrom (1) Constance Chen (1) Chen Wu (1) David J Hunter (1) Peter Kraft

(1) Harvard School of Public Health

Whole-genome and whole-exome sequencing studies may identify rare alleles that markedly increase risk of complex diseases. The sample sizes needed to identify these alleles with sufficient sensitivity and specificity using statistical association alone exceed realistic near-term sequencing budgets and throughput. In the interim, functional annotation can help interpret sequencing results and prioritize variants for follow-up. However, the positive predictive values of many types of annotation are unknown or modest, and current databases contain both false positive and false negative annotations. To illustrate this, we present results from a pilot whole-genome sequencing study of 20 women from the Nurses' Health Study 1 (NHS1), including 8 women with breast cancer. One of the cases and none of the controls carried a predicted damaging missense mutation in PALB2 (Thr397Ser), a gene known to harbor other high-risk breast-cancer alleles. This mutation was seen once in 4,300 European Ancestry individuals in the Exome Sequencing Project (Fisher's $p=0.003$). Genotyping Thr397Ser in 5,988 breast cancer cases and 7508 controls from the Nurses Health Study 1 and 2 identified four additional carriers, one case and three controls (Fisher's $p=1$). This highlights the importance of screening putative causal

mutations against large control panels, as well as the importance of updating reference databases, such as OMIM and HGMD, that contain entries based on a single case report.

91
Evaluation and correction of low level contamination in variant calling and filtering for NGS data

Hua Ling (1) Kurt Hetrick (1) Elizabeth W Pugh (1) Jane Romm (1) Kimberly F Doheny (1)

(1) Center for Inherited Disease Research, Johns Hopkins University

Cross sample DNA contamination during NGS sequencing can lead to genotyping errors and false positives. It may be present in source DNA, introduced during library preparation or cross-talk during PCR amplification. We evaluated the level of contamination at which variant calling starts to be affected and whether down-sampling alternative alleles to a fixed or estimated contamination level can improve the calling. We chose a set of samples and manually constructed contamination by merging BAM files from different subjects with varying levels and types of combination (related/unrelated, ethnicities, 2-/3-way). Both contaminated and non-contaminated samples were called together using Unified Genotyper. A number of QC metrics (reproducibility, concordance and sensitivity to array data) were employed to evaluate the presence of contamination and quality of variant calls. VerifyBamID was used to estimate contamination levels. The contamination level was well estimated by VerifyBamID despite some variation across different types of contamination. Without any effort of correction, variant calls start to be affected when contamination reaches 3%. Down-sampling was effective in correcting moderate contamination when an accurate estimate of contamination level was provided during variant calling. However, when contamination level reached 10% or more, down-sampling to remove contamination did not work as well and significant numbers of errors remained in the variant calls. Application and integration of verifyBamID with variant calling allows us to detect contamination at early stage and significantly improves calling accuracy.

92
Bayesian Variable Selection for Gene-Gene and Gene-Environment Effects with Hierarchical Constraint

Changlu CL Liu (1) Jianzhong JM Ma (1) Christopher CIA Amos (2)

(1) UT MD Anderson Cancer Center
(2) Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College

Complex diseases such as cancer result from multiple genetic and environmental exposure factors. Genome-wide association studies have been able to localize many causal variants predisposing to many diseases. However, more advanced statistical models are needed to identify and characterize additional genetic and environmental factors and their interactions, which may help us better understand the causal mechanism of complex disease. Here we proposed a Bayesian hierarchical mixture model that allows us to investigate the genetic and environmental effects, gene by gene interactions and gene by environment interactions simultaneously. It is well known that, in many practical situations, there is a natural hierarchical structure between the main effects and interaction effects. The proposed model incorporates this hierarchical structure into the mixture model and thus can remove the irrelevant interaction effects more effectively, resulting in robust and parsimonious models. Our simulation results show that the proposed models well controlled the proportion of false positive and yielded a powerful approach to identify the predisposing effects and interactions. We also illustrated our model in the studies of gene-environment interactions in lung cancer and cutaneous melanoma.

93
Penalized robust analysis of gene-environment interactions in cancer studies

Shuangge Ma (1)

(1) Yale University

High-throughput cancer studies have been extensively conducted, searching for genetic markers associated with outcomes and phenotypes beyond clinical and environmental risk factors. Studies have shown that gene-environment interactions have important implications. The existing methods can be limited in that specific parametric or semiparametric models need to be assumed, making them subject to model mis-specification. In addition, they heavily rely on significance level, whose properties, such as correlation structure, are difficult to establish. In this study, we develop a penalized robust analysis method. The proposed method is robust by not making assumptions on the model format and by adopting rank estimation. It uses penalization for identifying important interactions and “avoids” significance level. A sigmoid approximation is introduced to reduce computational cost. In numerical study, to demonstrate the proposed method, we analyze prognosis data under the AFT (accelerated failure time) model. Simulation shows satisfactory performance of the proposed method. Analysis of two cancer prognosis datasets shows that the proposed method may identify markers with important implications.

94
Genetic Factors that Affect HIV Infection and AIDS Progression

Jessica M Madrigal (1)

(1) University of Illinois at Chicago

HIV-1 infection has rapidly spread worldwide and has become a leading cause of mortality in infectious diseases. Genetic studies of HIV-1 susceptibility have investigated the uniqueness of long-term non-progressors (HIV infected but do not progress to AIDS) and highly exposed but uninfected individuals. The aim of this project was to examine the genetic epidemiological studies in the peer-reviewed literature to better understand the various biological factors that can affect HIV-1 susceptibility and the rate at which HIV-1 infection progresses to AIDS in humans. It is known that host genetic factors play an important role in the outcome of many complex diseases such as AIDS and are also known to regulate the rate of disease progression. Susceptibility to HIV infection and progression to disease are complex traits that can be altered by genetic factors. In this review the role of genetic variation and the major host genes involved in modulating HIV infection and disease progression is presented to provide an enhanced understanding of improvements in diagnostics, prevention and treatment of HIV.

95
Exome sequencing analysis of 10,000 type 2 diabetes cases and controls from five ancestry groups

Anubha Mahajan (1) Jason Flannick (2) Noel Burtt (2) Manuel Rivas (1) Xueling Sim (3) Heather Highland (4) Kyle Gaulton (1) Pablo Cingolani (5) Pierre Fontanillas (2) Tanya Teslovich (3) Andrew Morris on Behalf of T2D-GENES Consortium (1)

(1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom
(2) Broad Institute, Cambridge, MA, USA
(3) University of Michigan, Ann Arbor, Michigan, USA
(4) Human Genetics Center, University of Texas Health Science Center, Texas, USA
(5) McGill University and Génome Québec Innovation Centre, Montréal, Canada

We have undertaken whole-exome sequencing of >10,000 type 2 diabetes (T2D) cases and controls from five major ancestry groups: African American (AA), East Asian (EA), European (EU), Hispanic American (HA), and South Asian (SA). The unique study design will yield a catalogue of coding variation across diverse populations, enabling us to: 1) evaluate the full spectrum of coding variation and assess how it differs across ancestries; 2) assess the genome-wide contribution of coding variation to T2D risk within and across ancestries; and 3) assess the contribution of coding variation to T2D risk at established genome-wide association study (GWAS) loci. Analysis of 9,966 individuals passing quality control identified ~2.4M single nucleotide variants. Only 111K (4.6%) of these are present in all five ancestry groups, 47% of which have minor allele frequency (MAF) >5%. Conversely, the 20%, 18%, 5%, 9% and 21% of variants unique to AA, EA, EU, HA and SA ancestries are largely rare (MAF <1%). Using MANTRA trans-ethnic single variant meta-analysis, we identified association of rs2233580 in PAX4 (log10BF=6.6) with T2D at genome-wide significance. The association was specific to EA samples; p=1.2x10-8, 10% MAF; odds ratio 1.78 (1.46-2.17). Only two copies of the minor allele were observed in any of the other ancestry groups. This variant is independent of the nearby lead SNP at the GCC1 locus (rs6467136, r2=0.022) identified in EA GWAS. When we considered coding variants within established T2D GWAS loci, we observed enrichment for rare (MAF <1%) deleterious variants associated with the disease, although further analyses are required to assess if they fully explain previously reported common lead SNPs.

96
Association between C677T polymorphism of methylene tetrahydrofolate reductase and congenital heart disease: meta-analysis of 7,697 cases and 13,125 controls.

Chrysovalanto Mamasoula (1) R Reid Prentice (2) Tomasz Pierscionek (1) Faith Pangilinan (2) James L Mills (3) Charlotte Druschel (4) Darroch Hall (1) Lawrence C Brody (2) Heather J Cordell (1) Bernard D Keavney (1/5)

(1) Institute of Genetic Medicine, Newcastle University
(2) Molecular Pathogenesis Section, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA
(3) Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA
(4) Congenital Malformations Registry, New York State Department of Health, Troy, NY 12180 and the Department of Epidemiology and Biostatistics, University at Albany School of Public Health, Rensselaer, NY, 12144
(5) Institute of Cardiovascular Sciences, University of Manchester

Background: Association between the C677T polymorphism of the methylene tetrahydrofolate reductase (MTHFR) gene and congenital heart disease (CHD) is contentious. Methods and Results: We compared genotypes between CHD cases and controls, and between mothers of CHD cases and controls. We placed our results in context by conducting meta-analyses of previously published studies. Among 5,814 cases with

primary genotype data and 10,056 controls, there was no evidence of association between MTHFR C677T genotype and CHD risk (OR 0.96 [95% CI 0.87-1.07]). A random-effects meta-analysis of all studies (involving 7,697 cases and 13,125 controls) suggested the presence of a borderline association (OR 1.25 [95% CI 1.03-1.51]; p=0.022), but with substantial heterogeneity among contributing studies (I2=64.4%), and evidence of publication bias. Meta-analysis of large studies only (defined by a variance of the log OR less than 0.05), which together contributed 83% of all cases, yielded no evidence of association (OR 0.97 [95% CI 0.91-1.03]), without significant heterogeneity (I2=0). Moreover, meta-analysis of 1,781 mothers of CHD cases (829 of whom were genotyped in this study) and 19,861 controls revealed no evidence of association between maternal C677T genotype and risk of CHD in offspring (OR 1.13 [95% CI 0.87-1.47]). There was no significant association between MTHFR genotype and CHD risk in large studies from regions with different levels of dietary folate. Conclusions: The MTHFR C677T polymorphism, which directly influences plasma folate levels, is not associated with CHD risk. Publication biases appear to substantially contaminate the literature with regard to this genetic association.

97
Whole Exome Sequencing of Familial Bicuspid Aortic Valve

Lisa J Martin (1) Valentina Pilipenko (1) Kenneth Kaufman (1) Mehdi Keddache (1) D Woodrow Benson (2)

(1) Cincinnati Children's Hospital Medical Center
(2) Children's Hospital of Wisconsin

Bicuspid aortic valve (BAV) is the most common congenital cardiovascular malformation (CVM). In spite of highly heritability, few causal variants have been identified. The purpose of this study was to examine whole exome sequencing's (WES) utility to identify BAV causal variants. WES was performed on 17 individuals from a single multiplex family (BAV = 3, other CVM = 3). Post-GATK quality control metrics (QCM) were established after examining the relationship between Mendelian error (ME) rate and coverage, quality score, and call rate. We applied 4 variant selection strategies (affected cousins (n=2), nuclear subfamily (n=5), case control (n=14), and linkage (n=14)) on data passing QCM and with minor allele frequencies less than 10%. Post GATK, ME rate was 16.8%. After QCM, ME rate was 2.8%. We based affected status on presence of BAV, and found there was an exponential inverse relationship between number of variants identified and number of individuals included in variant selection (cousins = 4409; nuclear = 416; linkage = 3). Ignoring relatedness increased variants identified (case control = 107). These results demonstrate QCM reduces error rates. Additionally, as the variants identified using linkage were unlikely candidates (noncoding variants, not in candidate genes) thus would not be detected with current bioinformatic screens, these data support the value of sequencing larger numbers of individuals in families.

98
Effects of Waterpipe Smoking on gene expression

Zahra ZM Montazeri (1) Hoda HE El-Katerji (1) James JG Gomes (1) Julian JL Little (1)

(1) Ottawa University

Objectives: Recently, a sharp rise in waterpipe smoking has been observed in North America and Europe, especially among young adults. There is a belief that waterpipe is the healthiest way to smoke tobacco. We studied the effects of waterpipe smoking on gene expression among young waterpipe smokers. Methods: Adverse health effects of waterpipe smoking have been reported in different studies, but this evidence has been appraised as of poor quality. 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 among young waterpipe smokers in Ottawa, Canada. We asked participants to respond to a questionnaire and to provide a saliva sample. DNA extracted from these samples is being analyzed using PCR arrays. Results: We selected 18 genes to study the potential carcinogenic effects of waterpipe smoking on cancer based on gene expression. They have been selected based on the fact that they are induced by cigarette smoking as well as they are in the pathways of cancer diseases; genes of xenobiotic metabolism are also included. The main inclusion criteria was that the individual was between the age of 18 and 25 and reported that they smoked waterpipe. The fold change between before and after one hour and a half of smoking waterpipe, effect size, ranged between 0.02 and 34.42. Conclusions: Results could be used to predict the health effects of waterpipe smoking. This research will be used in knowledge translation to enable public health professionals and policy makers to make informed decisions about the control of waterpipe smoking, including potential prevention strategies and cessation interventions.

99
Exploiting interestingness in a computational evolution system for the genome-wide genetic analysis of Alzheimer's Disease

Jason H Moore (1) Douglas P Hill (1) Andrew Saykin (2) Li Shen (2)
(1) Institute for Quantitative Biomedical Sciences, Dartmouth College
(2) Department of Radiology and Imaging Sciences, Indiana University School of Medicine

Susceptibility to Alzheimer's disease is likely due to complex interaction among many genetic and environmental factors.

Identifying complex genetic effects in large data sets will require computational methods that extend beyond parametric statistical methods such as logistic regression. We have previously introduced a computational evolution system (CES) that uses genetic programming (GP) to represent genetic models of disease and to search for optimal models. The CES approach differs from other GP approaches in that it is able to learn how to solve the problem by generating its own operators. A key feature is the ability for the operators to use expert knowledge to guide the stochastic search. We have previously shown that CES is able to discover nonlinear genetic models of disease susceptibility in both simulated and real data. The goal of the present study was to introduce a measure of interestingness into the modeling process. Here, we define interestingness as a measure of non-additive gene-gene interactions. That is, we are more interested in those CES models that include attributes that exhibit synergistic effects on disease risk. To implement this new feature we first pre-processed the data to measure all pairwise gene-gene interaction effects using entropy-based methods. We then provided these pre-computed measures to CES as expert knowledge and as one of three fitness criteria in three-dimensional Pareto optimization. We applied this new CES algorithm to an Alzheimer's disease data set with approximately 520,000 genetic attributes. We show that this approach discovers more interesting models with the added benefit of improving classification accuracy.

100
Fine-mapping type 2 diabetes susceptibility loci with the Metabochip

Andrew P Morris (1) Tanya M Teslovich (2) Teresa Ferreira (1) Anubha Mahajan (1) Yeji Lee (2) Nigel W Rayner (1) Colin N A Palmer (3) David Altshuler (4) Michael Boehnke (2) Mark I McCarthy (1)

(1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
(2) Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA.
(3) Medical Research Institute, University of Dundee, Ninewells Hospital, Dundee, UK.
(4) Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA.

We combined Metabochip data from 25,154 type 2 diabetes (T2D) cases and 50,269 controls of European ancestry, supplemented by imputation up to the 1000 Genomes Project reference panel (March 2012 release), to facilitate fine-mapping of 39 established loci for the disease. We aimed to: (i) delineate association signals in established loci arising from multiple independent causal variants; (ii) assess the evidence for association with low-frequency (LF) variants, minor allele frequency (MAF) <5%, in established loci; and (iii) define "credible sets" of SNPs that account for 99% of the probability of including the causal variant at each association signal. We undertook approximate conditional analyses to identify loci with multiple SNPs at nominal significance ($p<10^{-5}$), reflecting increased prior odds of association in Metabochip fine-mapping regions. The association signals mapping to TCF7L2 and KCNQ1 can be explained by three causal variants, whilst those at DGKB, CDKN2A/B, KCNJ11, HMGA2, HNF1A and GIPR can be delineated by two. The conditional analysis highlighted LF associated variants mapping to TCF7L2 (rs140820620, MAF=1.9%, $p=6.5\times10^{-13}$; rs180726800, MAF=2.0%, $p=1.9\times10^{-6}$), KCNJ11 (rs61763083, MAF=0.3%, $p=7.5\times10^{-7}$), HMGA2 (rs116521220, MAF=0.6%, $p=6.3\times10^{-6}$), and HNF1A (rs1800574, MAF=2.6%, $p=1.5\times10^{-6}$), all independent ($r^2<0.05$) of the common lead SNPs at these loci. The 99% credible sets map to <10kb at MTNR1B (rs10830963 only), CDKN2A/B (both signals map to <2kb), TCF7L2 (primary signal includes 3 SNPs, 4.3kb), HNF1B (7 SNPs, 5.8kb), and GCK (3 SNPs, 9.8kb). Our study has provided insights into the genetic architecture of T2D at established loci and prioritised regions for further investigation.

101
Slicing the Genome: A New Approach to Association in Complex, Longitudinal Diseases

Anthony Musolf (1) Douglas Londono (1) Alejandro Q. Nato (2) Philippe Vuistiner (3) Carol A. Wise (4) Lei Yu (5) Stephen J. Finch (6) Murielle Bochud (3) Tara C. Matise (1) Derek Gordon (1) Pascal Bovet (7)

(1) Department of Genetics, Rutgers University
(2) Divison of Medical Genetics, University of Washington
(3) Swiss Institute of Bioinformatics
(4) Seay Center for Musculoskeletal Research, Texas Scottish Rite Hospital for Children
(5) Center for Alcohol Studies, Rutgers University
(6) Department of Applied Mathematics and Statistics, Rutgers University
(7) Ministry of Health and Social Services of the Seychelles

We have published a method that tests for association between a longitudinal phenotype and genetic variants. The method uses growth mixture models to determine longitudinal trajectory curves. The Bayesian posterior probability (BPP) of belonging to a specific curve is used as a quantitative phenotype in association. Though the method proves to be powerful for a single causal variant, power decreases when more than one causal variant is used. We present a new method designed to detect multiple causal SNPs associated with longitudinal phenotypes. This method can be used for both family-based and population-based studies and can use covariates. Instead of performing individual association tests on each SNP, we slice the genome into non-overlapping blocks of 50 SNPs. A p-value is obtained for each block via the SumStat method, developed by Jurg Ott and colleagues. Since SumStat is a population only test, we use a modified procedure (TDT-HET) to test for

family-based association. We test various scenarios in our simulations, including four causal variants located within a block and eight causal variants spread between blocks on different chromosomes. We also introduce environmental covariates. Our data set is highly stratified to ensure robustness in the presence of population stratification. We report that our simulations: 1) appear to maintain the proper type I error and 2) have greater than empirical 75% power for most simulations. These results suggest that our method can detect multiple causal SNPs located in multiple regions across the genome. We believe that is method will be use to researchers who are studying complex diseases that display longitudinal phenotypes as it allows for potentially high power for association of causal loci.

102
'Filter Feeding': Principled exploratory filtering approaches for sequence data to identify variants, genes, and regions for genetic follow-up studies.

Adam C Naj (1)

(1) University of Pennsylvania Perelman School of Medicine

Large scale sequencing projects using many cases and controls to identify both rare and common variants contributing to complex disease risk remain cost-prohibitive for most studies. Among the feasible approaches being implemented widely are targeted resequencing, whole exome sequencing, and whole genome sequencing of small numbers of samples (tens or hundreds of cases and sometimes controls) and the use of filtering strategies on the resulting data to facilitate follow-up genotyping of candidate variants or limited fine-mapping sequencing of candidate genes or regions for association analyses in larger datasets. Principled filtering strategies are thus critical in addressing potential type I and type II errors in identifying variants/genes/regions for follow-up. In this presentation, we will review filtering approaches for examining small-scale sequencing data in both family-based and unrelated case-control data, as well as sample selection strategies for both primary sequencing and follow-up genotyping or sequencing projects. We will discuss the application in filtering of data from multiple annotation sources, as well as preliminarily highlight the value and potential applications of new data on functional genetic elements arising from the ENCODE Project. We will explore the application of these approaches to several on-going genetic studies of late-onset Alzheimer's disease.

103
Genome-wide copy number variation and breast cancer in African American women

Heather M Ochs-Balcom (1) Christophe G Lambert (2) Jean Wactawski-Wende (1) Rowan Chlebowski (3) Lara E Sucheston (4)

(1) Department of Social and Preventive Medicine, University at Buffalo
(2) Golden Helix, Inc.
(3) Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center
(4) Division of Cancer Prevention and Population Sciences, Roswell Park Cancer Institute

Inherited copy number variation has been implicated in complex phenotypes, including cancers. We measured the association of genome-wide copy number variation with invasive breast cancer in African American participants from the Women's Health Initiative. We generated normalized log2ratios from raw Affymetrix 6.0 array data for 272 cases and 6667 non-cases; filtering based on quality control metrics yielded 230 cases and 5200 non-cases. We analyzed the association of single probes and continuous segment means with breast cancer status. Fifty single probe associations were significant at p=5.0x10-8; three probes (CN_498754, rs2297687, CN_371897) were significant in both the whole and filtered samples. CN_498754 is nearby GATA3 and ITIH5 that are involved in luminal epithelial cell differentiation in the mammary gland and tumor metastasis, respectively. Permutation tests on 34 associations with continuous segment means that were genome-wide significant in both samples were no longer significant; the most significant probes in the filtered sample were NRG3 and ATP7B (p-values=1.2x10-5 and p=2.07x10-4, respectively), where the mean of log ratios was lower in cases suggesting that deletions may be driving susceptibility. These two genes were borderline significant in Cytoscape GoBingo analysis (p=5.21x10-2) and fall in the GO category 30879 for mammary gland development. Genome-wide CNV signals reported herein deserve follow-up given their potential functional significance.

104
Network-guided random forests for the detection of gene-gene interactions

Qinxin Pan (1) Ting Hu (1) James D Malley (2) Angeline S Andrew (1) Margaret R Karagas (1) Jason H Moore (1)

(1) Institute for Quantitative Biomedical Sciences, Dartmouth College
(2) Center for Information Technology, National Institutes of Health

Common diseases are driven by a multifactorial interplay of genetic and environmental factors. As such, new bioinformatics methods are needed that embrace the complexity of the genotype-phenotype mapping relationship. One of the primary reasons that gene-gene interactions are not more commonly investigated is that exhaustive evaluation of all possible combinations of genetic variants is computationally expensive. Network science is emerging as a useful approach for characterizing the space of pairwise interactions systematically,

which can be informative for searching higher-order interactions by prioritizing genetic attributes clustered together in the networks. Meanwhile random forests (RF) has been successfully applied in many studies although it is known to rely on marginal main effects. In the present study, we introduce and evaluate a hybrid algorithm, network-guided random forests, which overlays the neighborhood structure of a gene-gene interaction network onto the growth of the forest. By embedding the unsupervised process of network building into the supervised process of prediction, we are able to use pairwise variable relationships and prioritize searching the space of higher-order interactions. Applying this approach to a population-based genetic study of bladder cancer susceptibility, we found that network-guided random forest produces trees with lower error, reduced computational cost and increased interpretability. Finally, it highlights variables that are in higher-order interactions, which can be further validated using the explicit test of interactions. This approach opens the door to bioinformatics analysis of high-order gene-gene interactions.

105
A comparison of penalized regression methods for prediction modeling in a large-scale candidate gene study

Angela P Presson (1) Kristin L Ayers (2) Jennifer S Herrick (1) Martha L Slattery (1)

(1) University of Utah, Division of Epidemiology, Department of Internal Medicine, 295 Chipeta Way Salt Lake City, Utah 84108, USA
(2) Institute of Genetic Medicine, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne, NE1 3BZ, UK

Penalized regression (PR) methods such as the lasso select concise predictors in large-scale genetic studies. The group lasso (GL) and sparse group lasso (SGL) methods can incorporate gene information in model selection. We show that GL and SGL improve prediction accuracy over the lasso in a breast cancer candidate gene study. Our data included 33 non-SNP predictors and 277 SNPs (in 34 genes) for 3560 cases and 4134 controls. We analyzed up to 3-way interactions for the top 30 predictors at a variety of penalty strengths λ using seven different PR models in Mendel v12, including lasso, GL, and the SGL at five values of the mixing parameter p which controls the ratio of the group to sparsity penalty. We compared prediction accuracies across λs using 10-fold cross-validation (CV), and important predictors were identified by fixing the model size (MS) at 10 and performing 5x10-fold CV. We analyzed prediction accuracy, MS, and frequencies of predictors across folds. Across λs, p=0.5 had top accuracy most frequently followed by other SGL models. On average, actual MSs of SGL and GL were about 5% and 15% larger than lasso, respectively. Prediction accuracies at MS=10 were 61.9-62.3%, whereas accuracies at MS=200 were 81.6-84.7%. Our PR analysis showed that SGL and GL had better prediction accuracies but yielded larger models than lasso, and accuracy increased with MS. Thus, in studies with many small-effect predictors, it appears justified to choose λ for a desired MS.

106
Age-varying genetic effects cause missing heritability

Jesse D Raffa (1) Elizabeth A Thompson (1)

(1) Department of Statistics, University of Washington

Traditionally studies estimating the heritability of a quantitative trait have assumed that heritability is constant over age. We consider situations where a latent multivariate trait governs both baseline values (intercept) and changes over age (slope) of a single observed phenotype. In such a situation, after adjusting for the fixed effects of age and other confounders, heritability varies over age. Ignoring this variation over time generally underestimates the true heritability. In a cross-sectional analysis using only one time point per individual, the magnitude on this bias depends on: the sampling process of the observation ages and the magnitude of the random variation of the slope. If the sampling ages are selected independent of the latent trait values, the expectation of the estimated additive variance will be reduced by var(sampling age)var(slope), resulting in an underestimate of the true heritability, unless var(sampling age) or var(slope) equal zero. The variance of the sampling age would be close to zero when all individuals are approximately the same age, and under such a situation, no bias results, but heritability estimates will vary across studies which use different ages. When observation ages are dependent on the latent traits (e.g., case-control studies), the magnitude of the bias in the estimate of heritability increases. Finally, when the relatedness of individuals is misspecified, the bias induced by the age varying genetic effects on the heritability estimate increases. Under realistic scenarios for sampling ages and genetic effects, the bias created by ignoring these age varying effects can be up to 20%. Longitudinal models, methods and software are needed to efficiently address these and other issues.

107
The PGx project: design and implementation

Laura J. Rasmussen-Torvik (1) Joshua Denny (2) Marc S. Williams (3) Brendan Keating (4) Ariel Brautbar (5) Cindy Prows (6) Shannon Manzi (7) Suzette J. Bielinski (8) Stuart Scott (9) James Ralston (10) Simona Volpi (11) Adam Gordon (12) Jonthan Haines (2) Dan Roden (2)

(1) Northwestern University Feinberg School of Medicine
(2) Vanderbilt University
(3) Geisinger Health System
(4) University of Pennsylvania
(5) Marshfield Clinic

(6) Cincinnati Children's Hospital Medical Center
(7) Boston Children's Hospital
(8) Mayo Clinic
(9) The Mount Sinai Hospital
(10) Group Health Research Institute
(11) NHGRI
(12) University of Washington

Design: One vision of the human genome project is that genetics will guide preventive and therapeutic decisions. Many common pharmacogenetic variants have been discovered, but the importance of rare variation in these genes is not well understood. There is also now a critical need to determine the best application of pharmacogenetic testing into routine clinical care. The PGx project is a partnership between the eMERGE network and PGRN designed to pursue three aims: 1) to deploy PGRNSeq, a targeted next generation sequencing platform, in a population at high likelihood of being prescribed certain medications, 2) to create a repository of information drawn from electronic health records (EHR) to examine outcomes in patients having rare pharmacogenetic variants of interest, and 3) to perform a pilot study implementing pre-emptive genetic testing and decision support in EHR for a number of validated pharmacogenetic gene/drug pairs. Implementation: The PGx project is enrolling over 9000 participants at 10 clinical sites. All participants will by genotyped on PGRNSeq and have information about demographics, diagnoses, and medications placed into a central repository. Each PGx site is working with local practitioners to determine which pharmacogenetic gene/drug pairs will be implemented for pre-emptive genetic testing and to develop decision support within the framework of their EHR. Process outcomes will be examined across sites.

108
WITHDRAWN

109
Estimation of statistical power to detect genetic association in longitudinal data using mixed models

Ghislain Rocheleau (1) Loïc Yengo (2) Philippe Froguel (3)

(1) Lille 2 University, Lille, France
(2) CNRS 8199-Institute of Biology, Pasteur Institute, Lille, France
(3) Department of Genomics of Common Disease, Imperial College London, London, UK

To detect novel loci associated to quantitative traits, the current approach is based on simple linear regression in a cross-sectional design. We extend this approach by considering the association between a locus and longitudinal measurements of the trait over time. One of the most natural extensions is to use the random coefficients model, more specifically the random slope and random intercept model, possibly including an interaction term between time and genotype at that locus. We derive closed-form formulas for the statistical power of testing the interaction term, the locus effect and the time effect in the random slope and random intercept model. Effects of parameters influencing statistical power of these tests, i.e. number of subjects, number of repeated measures per subject, minor allele frequency, intercept and slope variance-covariance matrix, are assessed mathematically, and by means of simulated genotype and phenotype data. General formulas for power estimation when considering other choices of the variance-covariance matrix incorporating within- and between-subjects variability are also discussed. We illustrate and apply our results on genotype and phenotype data coming from the French cohort D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d'Insulino-Résistance).

110
Empirical kinship estimation in the French Canadian founder population

Marie-Hélène Roy-Gagnon (1) Héloïse Gauvin (2) Jean-François Lefebvre (3) Catherine Laprise (4) Hélène Vézina (5) Damian Labuda (2)

(1) Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario; and CHU Sainte-Justine Research Center, Montreal, Canada
(2) CHU Sainte-Justine Research Center, Montreal, Canada; and Department of Social and Preventive Medicine, Université de Montréal, Montreal, Canada
(3) CHU Sainte-Justine Research Center, Montreal, Canada
(4) Department of Fundamental Sciences, Université du Québec à Chicoutimi, Chicoutimi, Canada
(5) Department of Human Sciences, Université du Québec à Chicoutimi, Chicoutimi, Canada

Several identity-by-descent (IBD) estimation methods have recently been developed, taking advantage of high-density genotype data. Most are designed for outbred populations and unilineal relationships. In founder populations like the French Canadian (FC) population of Quebec (Canada), any 2 individuals likely share many common ancestors. We examined the impact of including IBD shared at more than 2 alleles in a pair of individuals on kinship estimation in the FC founder population. We used genotype data from Illumina HumanHap650Y arrays and genealogical data from the BALSAC population register on 140 individuals from 7 sub-populations of Quebec. We used the IBDLD software, which allows estimation of all nine condensed identity states. We also used GERMLINE and Beagle Refined IBD. We compared empirical to genealogical kinship estimates using intraclass correlation coefficients (ICC) and correlated IBD estimates to genealogical characteristics including number of and distance to ancestors. Using IBDLD, we found that individuals shared more than 2 alleles IBD on 0.002 to 0.06% of their genomes on average (up to 2.7%) depending on the sub-population. We

found high ICCs (0.74-0.87) between empirical and expected kinship in the sub-populations with higher levels of relatedness. IBD sharing at more than 2 alleles did not improve the ICCs but was significantly associated with genealogical characteristics, and could thus provide information on relatedness in founder populations.

111
Multiplicative and additive gene-environment interaction between common breast cancer susceptibility loci and established environmental risk factors

Anja Rudolph (1) Montserrat Garcia-Closas (2) Nasim Mavaddat (3) Nilanjan Chatterjee (4) Mark Brook (5) Doug Easton (3) Jenny Chang-Claude (1)

(1) Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany
(2) Division of Genetics and Epidemiology, Institute of Cancer Research and Breakthrough Breast Cancer Research Centre, London, UK
(3) Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK
(4) Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA
(5) Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, UK

Over 70 breast cancer (BC) susceptibility loci have been identified to date. We tested for multiplicative and additive gene (G)-environment (E) interactions using a polygenic risk score (PRS) of 77 single nucleotide polymorphisms. Up to 23,667 invasive BC cases and 25,043 controls from 19 studies in the Breast Cancer Association Consortium were analyzed. The risk factors considered were age at menarche (≤13/>13 years), nulliparity, age at first pregnancy (≤25/>25 years), postmenopausal BMI (≤25/>25 kg/m2), current use of postmenopausal estrogen-progesterone therapy (EPT), current smoking, and cumulative lifetime alcohol consumption (≤12/>12 g/day). We evaluated multiplicative interaction using an Empirical Bayes model fitting procedure. Additive interaction was assessed by calculating relative excess risk due to interaction incorporating the G-E independence assumption to enhance efficiency. We observed three supra-additive interactions with PRS at $P < 0.05$ for categories of PRS (dichotomized or in quartiles). The expected and observed joint odds ratios for a dichotomized PRS at the median and dichotomous exposures were 2.29 vs 2.80, respectively; $P = 0.004$ for current EPT use; 2.22 vs 2.54; $P = 0.002$ for current smoking; and 2.14 vs 2.36; $P = 0.002$ for nulliparity. No significant GxE interaction on the multiplicative scale was found. In conclusion, our study provides evidence for additive GxE interactions between BC risk factors and common BC susceptibility loci. These findings indicate that the expected absolute risk reduction from changes in EPT use and smoking would be larger in women at higher than lower polygenic risk.

112
GenomeBrowse: Visual analytics and false-positive discovery for DNA and RNA-seq NGS data

Gabe G. Rudy (1) G. Bryce Christensen (1) Sam S. Gardner (1) Mike M. Thiesen (1) Autumn A. Laughbaum (1)

(1) Golden Helix

High-throughput sequencing (HTS) has recently provided price competitive alternatives to microarrays for both RNA expression profiling with the RNA-seq protocol and DNA genotyping with whole genome and whole exome sequencing. Although the bioinformatics tools have matured for secondary analysis of sequence data, including alignment, variant calling, and gene and transcript level quantification, the outputs of these tools often require inspecting the "raw read alignments" for putative variants and genes with interesting expression profiles. Investigating these variants in their VCF format and the alignments in BAM format allows for detection of false-positives as well as aiding the interpretation process by providing a rich genomic context.

113
Power of Family based Association Designs in Large Pedigrees Using Imputation

Mohamad M Saad (1) Ellen E Wijsman (1)

(1) University of Washington

In the last few years, the "Common Disease-Multiple Rare Variants" hypothesis has received much attention, especially with availability of next generation sequencing. Family based designs are well suited for discovery of rare variants with large effect, with large and carefully selected pedigrees enriching for multiple copies of such variants. However, sequencing a large number of samples is still prohibitive, despite the decreasing of sequencing costs, and the presence of multiple variants among families requires use of methods that allow for allelic heterogeneity. We propose a cost-effective strategy (pseudo-sequencing) to detect association with rare variants in extended families. This strategy consists of: sequencing a small subset of subjects, genotyping the remaining subjects on a set of sparse markers, and imputing the untyped markers in the remaining subjects conditional on sequenced subjects. We used a recent imputation method (GIGI: Genotype Imputation Given Inheritance), which is able to efficiently handle large pedigrees and to accurately impute rare variants. To test the association, we used famSKAT, which accounts for family relationship and allelic effect heterogeneity. We simulated sequence data on a large collection of big pedigrees and compared the power of

association test for: pseudo-sequence data, sequence data used for imputation, and all subjects sequenced. We also compared, within the pseudo-sequence data, the power of association test using best-guess genotypes and allelic dosages. Our results show that the pseudo-sequencing strategy can considerably improve the power to detect association. They also show that the use of allelic dosages results in much higher power than use of best-guess genotypes.

114
Use of multinomial regression model to identify loci underlying diseases with variable age of onset

Chloé Sarnowski (1) Marie-Hélène Dizier (2) Ismaïl Ahmed (3) Patricia Margaritte-Jeannin (2) Mark Lathrop (4) Florence Demenais (2) Emmanuelle Bouzigon (2) EGEA cooperative group

(1) INSERM, UMR 946, Paris, France & Univ. Paris Sud, Paris, France
(2) INSERM, UMR 946, Paris, France & Univ. Paris Diderot, Paris, France
(3) Univ. Paris Sud, Paris, France & CESP, INSERM, UMRS 1018, Villejuif, France
(4) McGill Univ., Montréal, Canada & CEA/CNG, Evry, France

Asthma is a heterogeneous disease and age of onset is one of the simplest features that can be used to differentiate asthma phenotypes. To characterize the genetic factors influencing asthma age-of-onset, we conducted a GWAS using a multinomial regression model applied to 750 asthmatics categorized according to their age-of-onset and 1,085 non-asthmatics from the French EGEA study with HapMap2 imputed data. Asthmatics were split into four specific age-of-onset sub-phenotypes: A) age-of-onset ≤ 4yrs (early-onset), B) 5-12yrs (before puberty), C) 13-20yrs (between puberty and adulthood) and D) > 20yrs (adult-onset). First, we applied an association test allowing heterogeneity of SNP effect between sub-phenotypes (Morris et al. Genet Epidemiol 2010) and detected 60 SNPs with P-value ≤10-5. Then, we tested whether these SNPs had a heterogeneous effect among the four sub-phenotypes. We identified 53 SNPs located in 16 regions with an interclass heterogeneity P-value ≤10-3. Among these regions, six had intra-class association P-values ≤10-5. We confirmed the specific association between 17q12-q21 genetic variants and early-onset asthma (P=10-6) (Bouzigon et al. N Engl J Med 2008). We also detected five new regions with SNP effect restricted to one asthma age-of-onset sub-phenotype: 9q34 with phenotype A (P=5x10-6), 3q25 with phenotype B (P=2x10-7), 1p13-p12, 3p22 and 3q27-q28 with phenotype C (P≤3x10-7). This analysis will be extended to GABRIEL Asthma consortium datasets. Thus, taking into account the age of onset in a multinomial regression framework can be a powerful approach to identify new loci underlying complex diseases. Funded by FRSR, ANR-GEWIS-AM, GABRIEL & Région Ile de France.

115
Evaluation of statistical interactions for binary traits

Jaya M Satagopan (1) Sara H Olson (1) Robert C Elston (2)

(1) Memorial Sloan-Kettering Cancer Center
(2) Case Western Reserve University

There has been a long-standing interest in the evaluation of interactions between risk factors in epidemiology and genetics. Statisticians generally define interaction as a departure from additivity in a linear model on a certain scale of measurement of the outcome. Certain interactions, which correspond to strictly monotonic curvilinear effect of the risk factors on the outcome, may be eliminated via an invertible transformation of the outcome. When the outcome is binary, the transformation corresponds to a link function. This work examines monotonicity properties of curvatures induced by multiple risk factors to obtain useful insights into the meaning of statistical interactions for binary disease traits. It is shown that, when the interactions are strictly monotonic curvatures, the relationship between the risk factors and the binary outcome is additive under a Box-Cox family of transformation of the disease odds. This effectively means that disease risk increases (or decreases) at a rate that is substantially higher than that postulated by a logistic distribution. These properties and their practical implications are illustrated using published data sets from case-control studies of bladder cancer, advanced colorectal adenoma, and endometrial cancer. These empirical illustrations provide some useful interpretations of interaction between smoking and NAT2 acetylation in bladder cancer and advanced colorectal adenoma, and interaction between estrogen-related factors in endometrial cancer.

116
Challenges in the imputation and data analysis of very large genotype-to-phenotype projects

Petteri Sevon (1) Timo PJ Kanninen (2) Lauri MA Eronen (3)

(1) Biocomputing Platforms Ltd

As the number of subjects in GWAS projects are reaching 100,000 and beyond, data analysis - especially imputing and downstream analysis of probabilistic imputed genotype data, and data format conversions (if transposing the data matrix is required) may become a research bottleneck, even when using external calculation cluster. In theory, for most analysis tasks performance can be increased by distributed computing - partitioning the data by genomic regions and by subjects. However, limited bandwidth between data storage and computing resources, disk system performance and non-

parallelizable component of splitting the data often significantly reduce the gains of parallelization.

To facilitate a massively parallel imputation process and further downstream analysis of imputed data, we have developed a workflow where instead of very large, monolithic files, data is partitioned into tiles of manageable size (fitting the random access memory), each covering a subset of the genome and subjects. Results of the imputation can be directly stored as compressed tiles, preferably on a storage close to the calculation servers. If downstream analysis is partitioned according to existing tiles (or multiples of them), the cost of splitting data will be eliminated completely.

In this presentation we evaluate feasibility, costs and parallelization performance of the proposed architecture on Amazon cloud environment by imputing a dataset of 100,000 subjects and performing association analysis on the imputed data using BC|GENOME 4.0 software platform, which implements the proposed workflow.

117
Test of Rare Variant Association Based on Affected Sib-pairs

Qiuying Sha (1) Shuanglin Zhang (1)

(1) Michigan technological University

With the development of sequencing techniques, there is increasing interest to detect associations between rare variants and complex traits. Quite a few statistical methods to detect associations between rare variants and complex traits have been developed for unrelated individuals. Statistical methods for detecting rare variant associations under family-based designs have not received as much attention as methods for unrelated individuals. Recent studies show that rare disease variants will be enriched in family data and thus family-based designs may improve power to detect rare variant associations. In this article, we propose a novel test to test association between the optimally weighted combination of variants and trait of interests for affected sib-pairs. The optimal weights are analytically derived and can be calculated from sampled genotypes and phenotypes. Based on the optimal weights, the proposed method is robust to directions of effects of causal variants and is less affected by neutral variants than existing methods do. Our simulation results show that, in all the cases, the proposed method is substantially more powerful than existing methods based on unrelated individuals and existing methods based on affected sib-pairs.

118
Testing Hardy-Weinberg Equilibrium Conditional on Admixture Reveals Natural Selection in Admixed Populations

Daniel Shriner (1) Adebowale Adeyemo (1) Charles N. Rotimi (1)

(1) National Human Genome Research Institute

The Hardy-Weinberg principle provides expectations for genotype frequencies given allele frequencies, with a key assumption of random mating. Admixture resulting from interbreeding between previously isolated parental populations could lead to violation of this assumption. Here, we developed a new version of the test of Hardy-Weinberg equilibrium (HWE) that is explicitly conditioned on admixture, investigated the impact of admixture on HWE, examined whether ancestry-informative markers disproportionately fail HWE, and used the new structured HWE test to detect loci with evidence of natural selection. By computer simulation, we show that the structured test has control of the type I error rate and is valid across the entire range of allele frequencies, mixing proportions, and number of parental populations. To assess the implications for analysis of genome-wide genotype data, we applied the test to data comprising 806,646 autosomal markers in 1,016 unrelated admixed African Americans. By conditioning on admixture, we detected 85 loci with evidence of positive, directional selection, 369 loci with evidence of balancing or frequency-dependent selection, and 148 loci with evidence of purifying selection. We also found that markers highly differentiated among the parental populations of an admixed population and therefore ancestry-informative did not disproportionately fail our structured test of HWE. Our new test and results have implications for quality control of genome-wide genotype and sequence data as well as the study of population genetics in admixed populations.

119
GWAS with longitudinal data on your notebook – computationally fast strategies for linear mixed models and linear regression.

Karolina Sikorska (1) Paul Eilers (1) Emmanuel Lesaffre (1)

(1) Erasmus Medical Center, Rotterdam, Netherlands

We suggest an approach which makes an analysis of a GWAS with longitudinal data feasible on a single computer. To account for a correlation between observations from the same individual, a linear mixed model (LMM) is used to identify SNPs influencing evolution of a trait over time. However, fitting millions of LMMs results in prohibitively long computation time. We propose the conditional two-step (CTS) approach as a fast approximation of the p-values for the SNP*time interaction effect. Our idea is based on the conditional linear mixed model (CLMM) in which longitudinal effects are fitted properly, regardless any misspecifications in the baseline characteristic of the individuals. In the first step we fit a reduced CLMM omitting SNP*time interaction term. In the second step, the estimated random slopes (BLUPs) are regressed on SNPs. The approximation given by the CTS is precise even when the data are unbalanced. We illustrate, through many examples, the

advantages of using CLMM as a preliminary step. To speed up the second step we propose a new approach for a quick analysis of many linear regression models. Our method analyses many SNPs at the same time using highly optimized matrix operations in R software. We call this approach semi-parallel to distinguish it from the multiprocessor parallel computing. We can fit regression models for 1M SNPs, 10K individuals within 10 minutes on an average laptop, working on blocks of SNPs. Quick data access is an important issue. We discuss our solutions based on array-oriented binary files. Combination of the CTS and semi-parallel approach speeds up the computations immensely, reducing demands on computing resources.

120
Meta-Analysis of Genome-Wide Association Studies in Myopia in Eight Populations

Claire L Simpson (1) Robert Wojciechowski (2) Virginie J.M. Verhoeven (3) Pirro Hysi (4) Maria Schache (5) Mohsen Hosseini (6) Laura Portas (7) Federico Murgia (7) Konrad Oexle (8) Andrew Paterson (6)

(1) NHGRI/NIH
(2) Johns Hopkins University
(3) Erasmus University
(4) King's College London
(5) University of Melbourne
(6) University of Toronto
(7) Istituto di genetica delle popolazioni, Consiglio Nazionale delle Ricerche
(8) Institut für Humangenetik, Technische Universitat Munchen

Myopia is a common refractive error which affects at least a third of most populations. Both genetic and environmental factors influence myopic development. It has a significant impact on the lives of affected individuals and carries high economic costs associated with treatment, loss of productivity and co-morbidity from vision impairment. Recent genome-wide association studies (GWAS) have identified a number of loci associated with myopia and refractive error. Here we report results of a large meta-analysis of myopia in 8 cohorts, for a total of 16,325 individuals of European ancestry and replication in a further 8 cohorts for a total of 7953 individuals. Genotypes in each population were imputed to HapMap2 and analyzed separately by each group. Cases were defined as a spherical equivalent of -1 diopters (D) or worse and controls were defined as > 0D. Individuals between 0 and -1D were coded as unknown. Analyses were performed including age, sex and years of education, plus the first 3 principal components to adjust for population structure. Meta-analysis was performed in METAL using the sample size schema. Due to large differences in numbers of cases and controls for some studies, effective sample sizes were calculated using the formula recommended by the authors of METAL. Genomic control was used to adjust for any residual structural differences between populations. SNPs with P < 1e-5 were identified and all SNPs within 500kb each side of that SNP selected for replication. The replication threshold was set by calculating the effective degrees of freedom using the Ramos method. SNPs were considered to replicate where the p value < 0.0026. Results of the meta-analyses will be presented.

121
Identification of a Functional Variant in ADCY3 Associated with Fat Mass through Phenotypic Refinement and Genome-Wide Association

Evie Stergiakouli (1) Romy Gaillard (2) Jeremy M Tavaré (3) H Rob Taal (4) David M Evans (5) John P Kemp (5) Susan M Ring (1) Vincent W V Jaddoe (2) George Davey Smith (5) Nicholas J Timpson (5)

(1) Avon Longitudinal Study of Parents and Children (ALSPAC), School of Social & Community Medicine, University of Bristol, Bristol, UK
(2) Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands; Department of Paediatrics, Erasmus Medical Center, Rotterdam, the Netherlands; Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
(3) School of Biochemistry, University of Bristol, Bristol, UK
(4) The Generation R Study Group, Erasmus Medical Center, Rotterdam, The Netherlands; Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands; Department of Paediatrics, Erasmus Medical Center, Rotterdam, the Netherlands
(5) Avon Longitudinal Study of Parents and Children (ALSPAC), School of Social & Community Medicine, University of Bristol, Bristol, UK; Medical Research Council (MRC) Centre for Causal Analyses in Translational Epidemiology (CAiTE), School of Social & Community Medicine, University of Bristol, Bristol, UK

Genetic association studies tend to be undertaken for recognised phenotypes based on convention. However, the use of more biologically proximal phenotypes can reduce the noise associated with genetic signals and also yield marked differences in genetic profile. Body mass index (BMI; weight(kg)/height(m2)) has been widely used for clinical and research purposes despite known limitations of the square of height to precisely account for covariation. We aimed to challenge conventional use of BMI by assessing the contribution of common genetic variants to a redefined measure of weight for given height using the appropriate power term for this age group. Genome-wide association studies (GWAS) of BMI and BMI adjusted for height in 5,089 participants from the Avon Longitudinal Study of Parents and Children (ALSPAC) revealed marked differences in the genomic profile of these traits. When adjusting for height, SNPs in ADCY3 (adenylate cyclase 3) showed comparable strength of association to the fat mass and obesity related locus,

FTO, (rs11676272 (0.28kg/m2 change per allele G (0.19,0.38), p=6.07 x 10-9). This result was replicated in an independent sample of 2,089 children from the Generation R Study. The association of rs11676272 with BMI is determined by fat mass specifically and is driven by a missense variant likely to alter ADCY3 activity. ADCY3 mRNA expression is prominent within nuclei of the hypothalamus involved in regulation of energy homeostasis. In this study we identified a likely functional locus associated with fat mass in childhood by improving measurement of the phenotype studied.

122
Comparing Common Multipoint Linkage Methods: The Untold Story

William C. L. Stewart (1) Susan E. Hodge (1) David A. Greenberg (1)

(1) Nationwide Children's Hospital

Third generation sequencing technologies are just around the corner, which means that researchers will soon enjoy longer sequence reads, lower error rates, and significant reductions in per sample costs. Furthermore, since these technologies will generate high-throughput genotypes and haplotypes, the power of family-based studies is likely to increase. However, to realize the full potential of any family-based analysis, researchers must first understand the statistical and computational complexities at hand. For example, our simulations show that the statistical power of three widely used methods (i.e. the maximized maximum lod (MMLS) (Greenberg et al. 1989), the nonparameteric linkage statistic—Zlr (Kong and Cox 1997), and the maximized lod (MOD) score (Dietter et al. 2007)), is actually quite similar, despite the fact that the mathematical models underlying these methods are considerably different, and despite the fact that one method in particular is used almost to the exclusion of the other two. Furthermore, when calculations are run on a typical desktop computer, MMLS and Zlr have comparable run times, and both statistics are several orders of magnitude faster than the MOD. However, in the context of parallel computing, MMLS is (on average) four times faster than Zlr, with slight differences in power depending on the degree of missing data. Another, attractive feature of the MMLS statistic is that it provides information about both mode of inheritance and heterogeneity. For the analysis of dense SNP data, our software package EAGLET computes MMLS (or optionally Zlr), and it is freely available from the web at: http://www.mathmed.org/wclstewart/SOFT/soft.html.

123
Adaptive Pathway-Based Methods for Association Testing

Yu-Chen Su (1) Juan Pablo Lewinger (1) James W Gauderman (1)

(1) University of Southern California

With a typical sample size, a single genomewide association study (GWAS) can only detect variants conferring a modest effect on risk. Standard one-variant-at-a-time methods for analyzing GWAS have very low power to detect individual variants with smaller effects. However, a set of variants with small effects may become detectable if grouped together and analyzed jointly — the premise of set-based methods. An appealingly set-based test is the rank truncated product (RTP) method, which uses Fisher's p-value combination method to summarize a pre-defined number of variants with the smallest individual p-values. The performance of RTP depends critically on the unknown proportion of associated variants among those tested. A set-based test exhibiting better power in detecting associated variants is obtained by selecting the number of variants to combine in an adaptive, data-driven fashion. A popular adaptive set-based method is the adaptive rank truncated product (ARTP), which finds the subset of variants giving the best evidence of association. We compared the standard ARTP, several new variations of ARTP we introduce, and other adaptive and non-adaptive in a comprehensive simulation study. In our simulations we assumed a set of 162 typed SNPs based on the LD structure of the prostaglandin E receptor 3 (PTGER3) gene and a range from 1 to 10 causal SNPs that are either genotyped or in LD with genotyped markers. We used permutations to assess significance and provide a level playing field comparison across all methods. Our results show that the standard ARTP was generally the most powerful method, with the LASSO, Global Model of Random Effects (GMRE), minimum p-value (MinP), and new variations of ARTP having all similar power to standard ARTP.

124
Control of type I error of single SNP marker linkage in H-E regression

Xiangqing Sun (1) Robert C Elston (1)

(1) Department of Biostatistics and Epidemiology, Case Western Reserve University

Various versions of Haseman-Elston (H-E) regression can detect linkage between a quantitative trait and genetic markers, and their implementation in the S.A.G.E. program SIBPAL allows calculation of permutation P-values on the assumption that the identity by descent sharing of alleles within sibships, and across sibships of the same size, is exchangeable. This assumes that the parental information is the same across all sibships; if not, the permutation should be conditional on the mating type, which may be unknown. This problem is exacerbated when we test linkage to a single diallelic SNP, which has been shown to be a powerful way to locate causal genes. An approximate way to condition on the six mating types, when the parental genotypes are known, would be to condition on the five classes defined by the number of minor

alleles carried by the mating pair. In this study, we examine controlling the type I error by conditioning on the number of minor alleles carried by each sibpair, which introduces two opposing biases: that due to conditioning on only five instead of six mating type classes, and that due to conditioning on the class of the sibpair instead of on the class of the sibship parents. This conditioning can be accomplished by including the sum of the sibpair's minor alleles as a covariate in the H-E regression, and we study this by extensive simulation. The results indicate that this adjustment can control the type I error in both the original H-E regression and the asymptotically most powerful version that uses the estimated weights to average of the squared sib-pair trait difference and the squared sib-pair mean-corrected trait sum (denoted W4 in SIBPAL).

125
Multimodal distribution of DNA methylation sites

Yan V. Sun (1)

(1) Department of Epidemiology, Rollins School of Public Health, Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, USA

Epigenetics refers to the heritable molecular modifications that are unrelated to primary DNA sequence, and can be modified by environmental exposures. DNA methylation (DNAm) is an essential epigenetic mechanism for normal development and is linked to the chronic diseases. The methylation levels may be shifted between the unmethylated and methylated states, which may show multimodality at the population level. Most methylome-wide analyses overlooked the multimodal distribution of the methylation levels, which may require different analytical strategy for the association analysis depending on the choice of statistical models. To identify the multimodal DNAm sites in human samples, we implemented the dip test proposed by Hartigan, and applied to over 470,000 DNAm sites in a sample of 656 adults. Since X chromosome inactivation is a known mechanism causing bimodal distribution of methylation levels between males and females, we excluded all sex-associated sites and all sites located on the X chromosome. The dip statistic is asymptotically larger for the uniform distribution than for any distribution in a wide class of unimodal distributions. We assessed the empirical p-values of the dip statistic based on 108 uniform distributions. We identified 4,084 multimodal sites with Bonferroni corrected empirical p-value < 0.05, among which 1,400 have a neighboring SNPs mapped to the probes. These multimodal distributed sites may indicate the unique molecular mechanism affecting the DNA methylation. A catalog of the multimodal DNAm sites is needed to facilitate proper analytical approach, and to understand the roles of DNAm in the growing field of epigenetic epidemiology.

126
How effective is meta-analysis as compared to mega-analysis of the pooled data for identifying gene-environment interactions?

Yun Ju Sung (1) Karen Schwander (1) Donna K Arnett (2) Sharon L.R. Kardia (3) Tuomo Rankinen (4) Claude Bouchard (4) Eric Boerwinkle (5) Steven C. Hunt (6) Dabeeru C. Rao (1)

(1) Division of Biostatistics, Washington University School of Medicine in St. Louis, St Louis, MO
(2) Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL
(3) Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI
(4) Human Genomics Laboratory, Pennington Biomedical Research Center, Baton Rouge, LA
(5) Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston TX
(6) Cardiovascular Genetics, Univeristy of Utah, Salt Lake City, UT

Meta-analysis combining summary results from multiple studies is a standard practice in GWAS. Gene-environment interactions are important, as they can explain a part of the missing heritability and identify individuals at high risk for disease. Analysis of interactions requires a large sample size, which is possible through consortia with multiple studies. However, whereas meta-analysis of main effects was shown to provide comparable results as mega-analysis of the pooled data, the same is not known for interactions.

127
Type I Error in Regression-based Genetic Model Building

HEEJONG HS SUNG (1) ALEXA AJMS SORANT (1) BHOOM BS SUKTITIPAT (1) ALEXANDER AFW WILSON (1)

(1) Genometrics Section, Inherited Disease Research Branch, NHGRI/NIH, Baltimore MD

The task of identifying genetic variants contributing to trait variation is increasingly challenging, given the large number and density of variant data being produced. Current methods of analyzing these data include regression-based variable selection methods which produce linear models incorporating the chosen variants. For example, the Tiled Regression method begins by examining relatively small segments of the genome called tiles. Selection of significant predictors, if any, is done first within individual tiles. However, type I error rates for such methods haven't been fully investigated, particularly considering correlation among variants. To investigate type I error in this situation, we simulated a mini-GWAS genome

including 306,097 SNPs in 4,000 unrelated samples with 100 non-genetic traits. 53,060 tiles were defined by dividing the genome according to recombination hotspots. Stepwise regression and LASSO variable selection methods were performed within each tile. Type I error rates were calculated as the number of selected variants divided by the number considered, averaged over the 100 phenotypes. Overall rates for stepwise regression using fixed selection criteria of 0.05 and LASSO minimizing mean square error were 0.04 and 0.12, respectively. Considering separately each combination of tile size and multicollinearity (defined as 1 – the determinant of the genotype correlation matrix), observed type I error rates for stepwise regression tended to increase with the number of variants and decrease with increasing multicollinearity. With LASSO, the trends were in the opposite direction. Different ways of choosing selection criteria were investigated for both methods.

128
Probability machines for quality control of called variants in next generation sequencing data

Silke Szymczak (1) Hua Ling (2) Kurt Hetrick (2) Margaret M Parker (3) Qing Li (4) Cheryl D Cropp (4) Alan F Scott (2) Terri H Beaty (3) Joan E Bailey-Wilson (4)

(1) Institute of Clinical Molecular Biology, Kiel University, Germany; Inherited Disease Research Branch, National Human Genome Research Institute, NIH, USA
(2) Center for Inherited Disease Research, Johns Hopkins University, USA
(3) Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, USA
(4) Inherited Disease Research Branch, National Human Genome Research Institute, NIH, USA

Next generation DNA sequencing technologies are a promising tool to identify rare genetic variants controlling susceptibility of complex diseases. Sequencing all exons or even genomes of many individuals usually identifies many thousands of rare variants and some of them are sequencing or alignment artifacts. Quality control to detect variants of low quality is therefore mandatory. However, research is still needed to determine the best quality control algorithms. We present a new quality control pipeline based on a multi-sample VCF file generated by GATK. For each single nucleotide variant (SNV), a probability of having low quality is estimated using probability machines based on random forests. Training data is generated by comparing HapMap genotypes in our study with data provided by the 1000 genomes project to detect mismatches. Quality characteristics of SNVs that match or do not match in the HapMap samples are used to train the machine and then the machine is used to predict the probability of low quality for the SNVs in the experimental data. We evaluate our new approach on a whole exome and a whole genome sequencing study and compare results with GATK's variant quality score recalibration.

129
Next generation association studies in isolated populations

Ioanna Tachmazidou (1) George Dedoussis (2) Lorraine Southam (3) Aliki-Eleni Farmaki (2) Graham RS Ritchie (4) Denise Xifara (5) Angela Matchan (1) Konstantinos Hatzikotoulas (3) Nigel W Rayner (3) Yuan Chen (1)

(1) Wellcome Trust Sanger Institute, Hinxton, UK
(2) Harokopio University Athens, Athens, Greece
(3) Wellcome Trust Sanger Institute, Hinxton, UK; Wellcome Trust Centre for Human Genetics, Oxford, UK
(4) Wellcome Trust Sanger Institute, Hinxton, UK; European Bioinformatics Institute, Hinxton, UK
(5) Wellcome Trust Centre for Human Genetics, Oxford, UK; Department of Statistics, Oxford, UK

Population isolates can enhance the power to detect association at low-frequency and rare sequence variation, because of potentially increased allele frequency and extended linkage disequilibrium. We have collected samples from two isolated populations in Greece (HELlenic Isolated Cohorts study, www.helic.org). All samples (n~3000) have information on a wide array of anthropometric, cardiometabolic, biochemical, haematological and diet-related traits. All individuals have been typed on the Illumina OmniExpress and HumanExome Beadchip platforms, and 250 individuals have been whole-genome sequenced at 4x depth. We are whole-genome sequencing all individuals at 1x depth. Using the exome-chip data from 1267 individuals, we find genome-wide significant evidence for association between R19X, a functional variant in APOC3, with increased HDL and decreased triglyceride levels. We find that around 3.8% of samples are heterozygous for the mutation. This cardioprotective variant has previously been associated with HDL levels in the Amish founder population. R19X is rare (<0.05% frequency) in outbred European populations. The increased frequency of R19X enables discovery of this lipid traits signal at genome-wide significance in a small sample size. The sample size needed to detect this in an outbred European population is 67,000. Our work provides a proof of principle of the value of isolated populations in detecting transferable rare variant associations of high medical relevance.

130
Exact p-values for SNPs accounting for testing of multiple genetic models

Rajesh Talluri (1) Jian Wang (1) Sanjay Shete (1)

(1) MD Anderson Cancer Center

In Genome wide association studies (GWAS) hundreds of thousands of SNP's are tested for association with a particular phenotype. There are several methods to account for multiple comparisons. However, typically, investigators also test each SNP using multiple genetic models. Association testing using the Cochran-Armitage trend test is commonly performed assuming an additive, dominant, or recessive genetic model for each SNP. Thus, each SNP is tested multiple times. The most popular, but incorrect, way of reporting p-values for a SNP has been to report the smallest p-value obtained from the three tests corresponding to the three genetic models which inherently leads to high type 1 error rate. Because of the small number of multiple tests (three) and high correlation (functional dependence) between the tests, the procedures available to account for multiple comparisons are conservative. We propose a method to calculate the exact p-value for testing a SNP with different genetic models. We performed simulations to demonstrate control of Type 1 error and power gains using the proposed approach.

131
Binary Trait Analysis in Sequencing Studies Under Trait-Dependent Sampling

Zheng-Zheng Tang (1) Dan-Yu Lin Lin (1)

(1) University of North Carolina at Chapel Hill

In sequencing study, it is a common practice to sample only the subjects with the extreme values of a quantitative trait. This is a cost-effective strategy to increase power in the association analysis. In the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP), subjects with extremely high or low values of body mass index (BMI), low-density lipoprotein (LDL) or blood pressures (BP) were selected for whole-exome sequencing. For a binary trait of interest, the standard logistic regression-even adjust for the trait of sampling-can give misleading results. We present valid and efficient methods for association analysis under trait-dependent sampling. The validity and efficiency of the proposed methods are demonstrated through extensive simulation studies and ESP real data analysis.

132
A population-based analysis of clustering identifies a strong genetic contribution to recurrent prostate cancer

Craig C Teerlink (1) Quentin C Nelson (1) Neeraj Agarwal (1) Robert A Stephenson (1) Lisa A Cannon-Albright (1)

(1) University of Utah

Prostate cancer is a common and often deadly cancer. We hypothesize that there are rare segregating variants responsible for high-risk prostate cancer pedigrees, but recognize that within-pedigree heterogeneity may contribute significant noise that overwhelms signal. Here we introduce a new method to identify potentially homogeneous subsets of prostate cancer based on clinical tumor characteristics demonstrating the best evidence for familial aggregation. The Genealogical Index of Familiality (GIF) is used to show evidence for significant familial clustering of a trait in a population genealogy. The subset GIF is a new method that tests for excess familial clustering of subsets of particular prostate cancer cases compared to all prostate cancer cases. Consideration of the familial clustering of 8 clinical subsets of prostate cancer cases compared to the expected familial clustering of all prostate cancer cases identified four subsets of prostate cancer cases with evidence for familial clustering significantly in excess of expected. These subsets include prostate cancer cases diagnosed before age 50 years, prostate cancer cases who survived at least 10 years after diagnosis, prostate cancer cases with BMI greater than or equal to 30, and prostate cancer cases for whom prostate cancer contributed to death. This analysis identified several subsets of prostate cancer cases that cluster significantly more than expected when compared to all prostate cancer familial clustering. A focus on high-risk prostate cancer pedigrees with these characteristics may enhance statistical power in variant identification studies.

133
Unravelling the genetics of lung function and COPD: from cases to cohorts to biobanks

Martin D. Tobin (1)

(1) University of Leicester

Mutations in alpha-1-antitrypsin have long been known to cause chronic obstructive airways disease (COPD, in which lung function measures show obstructive airflow), but other genetic determinants of COPD have been poorly understood. Over the last 5 years various study designs and analytic approaches have been employed to attempt to detect genetics of lung function and COPD, with at least 26 loci associating with lung function in large genome wide association studies (GWAS) undertaken by the SpiroMeta and CHARGE consortia, many of which are associated with COPD. As with many other traits, much of the heritability of lung function remains unexplained. Sequencing studies, copy number variation association studies and early exome chip studies are in progress. I will compare the relative merits – and drawbacks – of the various approaches we have employed to date. Across all complex traits, options to employ newer technologies to explore a more complete distribution of the allele frequency spectrum present substantial costs and design challenges. Here I outline the design and progress of UK BiLEVE, a consortium established to study lung function related phenotypes in 50,000 individuals in UK Biobank using a custom design Affymetrix array, designed to impute common and low frequency variants, with direct typing of rare, putatively functional variants.

134
Comparing the Haplotype Distributions between Populations

Liping Tong (1) Bamidele Tayo (1) Richard Cooper (1)

(1) Loyola University Chicago

Accurate characterization of haplotype structure and diversity is a key challenge in statistical genetics. We propose a new statistic to assess and compare the haplotype variations among populations which is particularly suited to this emerging challenge. We first describe the properties and calculations of this method. Subsequently, using simulation studies, we show that the proposed method is more powerful than the chi-square test statistic when comparing haplotype distributions under the following two circumstances (1) when variations of haplotype distances are not balanced (2) when haplotypes are tainted by accumulated mutations or genotyping errors. We also performed simulations to show that the proposed method can be applied to case-control association study and can be much more efficient than the single locus association test by greatly decreasing number of multiple tests. Finally, we applied our method to the Human Genome Diversity Project (HGDP) and HapMap3 data for SNPs on chromosome 2 in the region surrounding the LCT gene. Our results showed that 726 pairs of populations (out of 780) can be distinguished (p-value < 0.05) using the 127 SNPs surrounding the LCT gene.

135
Challenges in estimation of genetic effects from multiple cases family studies

Roula R Tsonaka (1) Renaud R Tissier (1) Jeanine J Houwing-Duistermaat (1)

(1) Leiden University Medical Center

Estimation of genetic effects in multiple cases family studies is often complicated by the outcome-dependent sampling and the within families correlation. An approach to deal with both of these features is the ascertainment corrected mixed-effects model. However, for small sample sizes and when the disease is rare convergence issues may arise, because the data do not provide sufficient information to estimate all model parameters. To overcome such numerical difficulties, we propose a penalized maximum likelihood estimation procedure which reliably estimates the model parameters in small family studies by using external population-based information. Extensive simulations have shown this to be a promising approach to reliably estimate the quantities of interest in small family studies. Finally, we illustrate our proposal using the Genetics In Familial Thrombosis study, which considers families with at least two affected siblings to contribute to the search for novel genetic risk factors for venous thrombosis.

136
Mediation of Genetic Effects from Nordic Twin Registry

Ayse Assistant Prof Dr Ulgen (1) Jacob Associate Prof Dr Hjelmborg (2) Wentian Research Professor Li (3)

(1) Eastern Mediterranean University, Faculty of Medicine, Famagusta, Mersin-10-Turkey
(2) University of Southern Denmark, Institute of Public Health, Odense, Denmark
(3) Robert S Boas Center for Genomics and Human Genetics,NY, USA

We apply a genetic modeling via simulation to twin data from Nordic Twin Registry for obesity and cancer related measurements. For obesity, we use both BMI and other quantities derived from BMI to measure the weight growth. More specifically, the log(BMI) at baseline is regressed over other factors: ln(BMI_ij)= beta_i + alpha_i *j +gamma*age_i+epsilon_ij, for individual i and timepoint j (time since the baseline, in years). The alpha_i is then the log-weight growth rate (Hjelmborg et al. Obesity, 16(4), 2008). For cancer data, we use phenotypes available in the registry. We follow the genetic modeling of twin data proposed by (Dite et al. and Stone et al; Cancer Epidemiol Biomarkers Prev; 17(12), 2008 and 17(12), 2012, respectively.) In this modeling, the phenotype of a twin in a twin pair is regressed over both twins' co-variates. If the two twins in a twin pair are labeled as 1 and 2, Y denotes phenotype and X co-variate, then E(Y)= alpha + beta_1 X_1 +beta_2 X_2. It was shown that by varying a covariate experimentally, the expected value of the phenotype measure would change. In our analysis, we assume a bivariate normal distribution for both (Y1,Y2) and (X1,X2). We treat phenotype measurements such as BMI, growth rate, at the baseline time as X, and that at the later time as Y. This approach would incorporate measurements at two points along a time course, thus enhance the power to detect the genetic component. We also introduce a random effects model for the stratification effects.

137
Replication of large-scale epistasis studies: an example on ankylosing spondylitis

Kristel Van Steen (1) Kyrylo Bessonov (1) Elena Gusareva (1)

(1) University of Liège: Montefiore Institute/GIGA-R

Ankylosing spondylitis (AS) is a common form of inflammatory arthritis occurring in approximately 5 out of 1,000 adults of European descent. Recently, the Australo-Anglo-American Spondyloarthritis Consortium and the WTCCC2 [1] showed that polymorphisms of ERAP1 only affect AS risk in HLA-B27-positive individuals, hereby establishing an interaction between ERAP1 and HLA in the

TASC, WTCCC2 and replication datasets. In this study, we use the aforementioned data from WTTCC2 on AS to address unresolved issues when performing large-scale SNP-SNP interaction studies, so as to better guarantee "stable" and "truly replicable" results. These issues are 1) the choice of variable selection method (e.g., known loci versus know pathways), 2) the choice of SNPs representing a genomic region (e.g., SNPs with modest versus negligible LD between them), 3) the choice of analysis method (e.g., regression-based versus data-reduction (non-parametric) based), 4) the choice of multiple testing correction (e.g., ignoring potentially varying SNP-SNP test distributions versus properly accounting for them). We show that even modest changes in 1)-4) may give rise to quite varying epistasis findings for AS, and motivate some "optimal" choices via extensive simulation studies. This undersese have led to additional epistasis signals (rs1058026 x rs2523608 (HLA-B), rs1632948 x rs2763979 (HLA-G x HSPA1B), and rs11244 x rs1015166 (HLA-DOB x TAP2), involving genes in the antigen presentation processes linked to the pathology of AS [4]. [1] Nature Genetics 2011; 43(9):919; [2] Hum Hered 2005; 59:79–87; [3] Ann Hum Genet 2011; 75:78-89; [4] Joint Bone Spine 2012; 79(3):243-8.

138
Performance of two imputation methods on large scale data: experiences in the eMERGE network

Shefali S Verma (1) Gretta J Armstrong (1) Dana C Crawford (2) Yuki Bradford (2) Mariza de Andrade (3) Iftikhar J Kullo (3) Gerard Tromp (4) Helena S Kuivaniemi (4) Loren Armstrong (5) Geoffrey Hayes (5) Brendan Keating (6) David R. Crosslin (7) Gail P. Jarvik (7) Bahram Namjou (8) Ebony B. Bookman (9) Rongling Li (9) Marylyn D. Ritchie (1)

(1) Center for Systems Genomics, The Pennsylvania State University, University Park, PA
(2) Center for Human Genetics Research, Vanderbilt University, Nashville, TN
(3) Mayo Clinic, Rochester, MN
(4) Geisinger Health System, Danville, PA
(5) Northwestern University, Chicago, IL
(6) Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA
(7) University of Washington, Seattle, WA
(8) Cincinnati Children's Hospital Medical Center, Cincinnati, OH
(9) Division of Genomic Medicine, National Human Genome Research Institute

The eMERGE network, an NHGRI funded initiative comprises nine sites each with DNA biobanks linked to electronic health records (EHRs). Approximately 39,206 unique DNA samples have been genotyped using either Affymetrix or Illumina genome-wide SNP arrays. Led by the Coordinating Center and the eMERGE genomics workgroup, we have developed an imputation pipeline for merging genomic data across the different SNP arrays used by the eMERGE sites, to maximize sample size and the power to detect associations. We performed imputation using the 1000 Genomes Cosmopolitan reference panel – which includes 1092 individuals and over 36 million SNPs. We compared accuracy of imputation results from two software packages - Beagle and Impute2 (phasing performed with ShapeIt2). For the comparison we used the following metrics: accuracy of imputation, allelic R2 (estimated correlation between the imputed and true genotypes for all imputed SNPs), and relationship between allelic R2 and minor allele frequency across all imputed SNPs.

139
Exploring the Relationship between Immune System Related Genetic Loci and Complex Traits and Disease through a Phenome-Wide Association Study (PheWAS)

Anurag Verma (1) Helena S. Kuivaniemi (2) Gerard C. Tromp (2) David J. Carey (2) Glenn S. Gerhard (1) James E. Crowe (3) Marylyn D. Ritchie (1) Sarah A. Pendergrass (1)

(1) The Pennsylvania State University, University Park, PA, USA
(2) Geisinger Clinic, Danville, PA, USA
(3) Vanderbilt University, Nashville, TN, USA

Exploring the relationship between immune-system related genetic loci and a wide array of phenotypes, including the presence or absence of disease, provides a way to elucidate more about interrelationships between the immune system and outcome traits and diagnoses and identify pleiotropic loci. To explore these connections further and identify novel associations and pleiotropy, we performed a Phenome-Wide Association Study (PheWAS), calculating comprehensive associations between 132,467 single nucleotide variants (allele frequency > 0.01) selected for previously known associations with autoimmunity and the immune system and 480 clinical diagnoses. To define case-control status we used ICD9 diagnosis codes from 3035 subjects using de-identified electronic medical records (EMRs) from the Geisinger Clinic MyCode biorepository. We required 10 or more case subjects for ICD9 code inclusion in our study, and used logistic regression for all associations with models adjusted for age and sex. We found a total of 2988 SNP-diagnosis associations with an exploratory p-value cutoff < 0.001, and 55 SNPs exhibiting association with more than one diagnosis code at this cutoff. The most significant association was for the SNP rs6025 and the diagnosis of "venus thrombosis" (P-value 3.5x10-8, 55 cases, 2981 controls). Other SNPs had significant associations with diagnoses such as "syncope and collapse", "rheumatoid arthritis", "acute pancreatitis", and "abnormal loss of weight". Further work will include using additional diagnosis codes available within this EMR, as well as seeking replication of the results of this study in an independent EMR based dataset through the Vanderbilt University Medical Center BioVU repository.

140
Human Genetics of Mycobacterium ulcerans infection (Buruli Ulcer): results from the first genome-wide association study

Quentin Vincent (1) Marie-Francoise Ardant (2) Julien Guernon (3) Jacques Gnossike (2) Ioannis Theodorou (3) LAurent Abel (1) Laurent Marsollier (4) Annick Chauty (2) Alexandre Alcaïs (1)

(1) Laboratory of Human Genetics of Infectious Diseases, Institut National de la Recherche Médicale U980 (INSERM), Université Paris Descartes, Sorbonne Paris Cité, Paris, France
(2) Centre de Détection et de Traitement de l'Ulcère de Buruli (CDTUB), BP 191, Pobè, Benin and fondation Raoul Follereau, Paris, France
(3) Laboratory of Immunity and Infection, INSERM UMR-S 945, UPMC Univ Paris 06, Groupe Hospitalier Pitié-Salpêtrière AP-HP, Paris, France.
(4) Institut National de la Recherche Médicale U892 (INSERM) et CNRS U6299, équipe 7, Université et CHU d'Angers, Angers, France

Buruli ulcer (BU), caused by Mycobacterium ulcerans, is a devastating emerging infectious disease and the third most common mycobacterial disease worldwide, after tuberculosis and leprosy. We report the results of the first genome-wide association study of BU.. Methods 408 PCR-confirmed BU cases diagnosed over the last 10 years at the Centre de Détection et de Traitement de l'Ulcère de Buruli in Pobe, Benin, Western Africa and 408 village-matched exposed controls were concomitantly enrolled and genotyped using the Illumina Omni2.5 array. Population stratification, cryptic relatedness, plate-effect and several other complex QC criteria were first assessed and eventually accounted for. Association studies combined several approaches, including the use of mixed models as implemented in the GEMMA software. After stringent QC filtering Stringent 402 cases and 401 controls were kept in the analysis. Principal Component Analysis (PCA) confirmed the Yoruba origin of all individuals of the cohort. PCA and pairwise IBS estimates revealed some level of cryptic familial relatedness between enrolled individuals and its impact on the analysis was assessed. Promising association signals were identified in several genes and a replication study in an independent sample from the same population is in progress. This first GWAS of BU ever implicated unanticipated pathways as key players in the physiopathology of the disease and highlights the power of forward genetics to dissect the genetic architecture of a neglected emerging tropical disease at an unprecedented molecular level.

141
GEE-based SNP Set Association Test for Continuous and Discrete Traits in Family Based Association Studies

Xuefeng Wang (1) Seunggeun Lee (1) Xiaofeng Zhu (2) Susan Redline (3) Timothy W. Yu (4) Christopher A. Walsh (4)

(1) Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA 02115
(2) Case Western Reserve University
(3) Department of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA
(4) Boston Children's Hospital, Boston, MA 02115, USA

Genetic association studies of related individuals or which use a family-based design provide opportunities to detect genetic variants that complement studies of unrelated individuals. Most statistical methods for family association studies for common variants are single-marker-based, which test one SNP a time. In this paper, we consider testing the effect of a SNP set, e.g., SNPs in a gene, in family studies. Specifically, we propose a Generalized Estimating Equations (GEE)-based kernel association test (KAT), a variance component-based testing method, to test for the association between a phenotype and multiple variants in a SNP set jointly using family samples. The proposed approach allows for both continuous and discrete traits, where the correlation among family members is taken into account through the use of an empirical covariance estimator. We derive the theoretical distribution of the proposed statistic under the null and develop analytical methods to calculate the p-values. We also propose an efficient resampling method for correcting for small sample size bias in family studies. The proposed method allows for easily incorporating covariates and SNP-SNP interactions. Simulation studies show that the proposed method properly controls for type-I error rates under both random and ascertained sampling schemes in family studies. We demonstrate through simulation studies that our approach has superior performance for association mapping compared to the single marker based minimum p-value GEE test for a SNP set effect over a wide range of scenarios. We illustrate the application of the proposed method using data from the Cleveland Family GWAS Study and an exome sequencing study of autism.

142
X-chromosome Genetic Association Test Accounting for X-inactivation, Skewed X-inactivation, and Escaping of X-inactivation

Jian Wang (1) Robert Yu (1) Sanjay Shete (1)

(1) The University of Texas MD Anderson Cancer Center

X-inactivation is the process in which one of the two copies of the X-chromosome in females is randomly inactivated to achieve the dosage compensation of X-linked genes between males and females. That is, 50% of the cells have one allele inactive and the other 50% of the cells have the other allele inactive. However, recent studies have shown that skewness or non-random X-inactivation is a biological plausibility in which more than 50% of cells have the same allele inactive. Also, some of the X-chromosome genes escape the X-inactivation, i.e., both alleles are active in all cells. Current statistical tests for X-chromosome association studies can either account for random X-inactivation (e.g., Clayton's approach) or escaping of X-inactivation (e.g., PLINK software). Because the true X-inactivation process is unknown and differs across different regions on X-chromosome, we proposed a unified approach of maximizing likelihood over all such biological possibilities: X-inactivation, X-inactivation skewness and escaping of X-inactivation. A permutation procedure was developed to assess the significance of the approach. We conducted simulation studies to compare the performance of the proposed approach with Clayton's and PLINK approaches. The results showed that the proposed approach has higher powers in most scenarios with well controlled type I errors. We also applied the approach to analyze the X-chromosome data for head and neck cancer.

143
Genetic Prediction of Quantitative Lipid Traits: Comparing Shrinkage Models to Polygenic Score Models

Helen R Warren (1) Juan-Pablo Casas (2) Frank Dudbridge (1) John C Whittaker (3)

(1) London School of Hygiene & Tropical Medicine
(2) London School of Hygiene & Tropical Medicine and UCL
(3) London School of Hygiene & Tropical Medicine and GSK

This project focuses on the statistical methodology for the genetic prediction of quantitative phenotypic traits related to disease risk. The aim is to investigate the benefits of incorporating much larger sets of SNPs into the models, not only the confirmed genetic determinants. Multivariate shrinkage models analysing all genotyped SNPs simultaneously, (e.g. Ridge Regression, Lasso and Hyper-Lasso) are compared to polygenic score models for the top ranked SNPs, considering weighted vs un-weighted scores, univariate vs multivariate weights and internal vs external weights. Our methods are applied to the genetic prediction of LDL & HDL and tested using Whitehall II and British Women's Health & Heart cohort studies, having almost 50,000 SNPs from the HumanCVD BeadChip, for over 5,000 and 3,000 individuals, respectively. We found shrinkage approaches rarely improved on standard weighted risk score models. This is due to the small proportion of SNPs with significant effects, and the strict shrinkage on the SNP effect sizes, especially without an initial pre-filtering of the SNPs, despite Lasso and Hyper-Lasso already including variable selection. We demonstrate how the results for predictive accuracy and best prediction model are trait-specific, highlight the benefit of a large training sample size and emphasise the importance of variable selection. As a conclusion from these results we see no reason to replace existing gene score methods by more complex methods for prediction, at this stage.

144
A Weighted U statistic for Genetic Association Analyses of Sequencing Data

Changshuai Wei (1) Ming Li (2) Zihuai He (3) Qing Lu (1)

(1) Michigan State University
(2) University of Arkansas
(3) University of Michigan

Despite the recent success of genome-wide association studies, a large proportion of genetic variants predisposing to complex diseases remain uncovered. Evidence from genetic studies and evolutionary theory has suggested rare variants could play an important role in the biological pathways of complex diseases. The advance of next generation sequencing technology facilitate the generation of massive amount of genetic variants and offers great opportunity to investigate the role of millions rare variants in the genetic etiology of complex disease. Nevertheless, great challenge has also been posed to statistical analyses of high-dimensional sequencing data. The association analyses based on traditional statistical methods endure substantial power loss because of low frequency of genetic variants and extremely high dimensionality of the data. We developed a weighted U statistic, referred to as USEQ, for high-dimensional association analysis of next-generation sequencing data. Based on the non-parametric U statistic, USEQ makes no assumption of the underlying disease model and can be applied to various types of phenotypes (e.g., binary and continuous phenotypes). Through simulation studies and an empirical study, we found USEQ outperformed a commonly used SKAT method when the underlying assumption is violated (e.g., the phenotype follows a heavily skewed distribution) and attained comparable performance to SKAT when underlying assumption is satisfied. In an empirical study of Dallas Heart Study (DHS) sequencing data, USEQ was also able to detect the association of ANGPTL 4 with very low density lipoprotein cholesterol.

145
Modeling chemotherapeutic-induced toxicities through integration of cell line and clinical genome-wide analyses

Heather E. Wheeler (1) Eric R Gamazon (1) Cristina Rodriguez-Antona (2) M. Eileen Dolan (1) Nancy J. Cox (1)

(1) Department of Medicine, University of Chicago, Chicago, IL
(2) Spanish National Cancer Research Center, Madrid, Spain

Dose-limiting toxicities such as peripheral neuropathy and neutropenia limit the effectiveness of chemotherapy drugs. Our goal is to understand to what extent cell line models can capture the overall genetic architecture of patient chemotherapy toxicity susceptibility. When comparing modestly sized pharmacogenomic genome-wide association studies from patients and lymphoblastoid cell lines (LCLs) treated with the same drug, SNPs rarely overlap at stringent thresholds such as $P < 10{-}6$, but significant overlaps of SNPs at more relaxed thresholds determined by enrichment analysis through random sampling are possible. Under this cumulative hypothesis, large numbers of common variants with small effects account for substantial heritability. For example, we observed an enrichment of carboplatin-induced cytotoxicity SNPs from LCLs ($n = 608$) in the neutropenia-associated SNPs from ovarian and lung cancer patients ($n = 143$) treated with paclitaxel and carboplatin (empirical $P < 0.001$). While we did not observe an enrichment of carboplatin cytotoxicity SNPs in the neuropathy-associated SNPs from the same patients, we did observe an enrichment of paclitaxel-induced LCL ($n = 247$) cytotoxicity SNPs in the peripheral neuropathy SNPs (empirical $P = 0.034$). Interestingly, we also observed an enrichment of paclitaxel cytotoxicity SNPs in diabetic neuropathy associated SNPs from patients ($n = 1651$) in the GoKinD cohort (empirical $P < 0.001$). These enrichments demonstrate that susceptibilities to increased cytotoxicity in LCLs and increased drug- and diabetes-induced toxicities in patients likely have some genetic mechanisms in common and support the role of LCLs as a preclinical model for investigating such toxicities.

146
Recurrent tissue-specific mtDNA mutations

Scott M Williams (1) David C Samuels (2) Bingshan Li (2) Zhuo Song (2) Eric Torstenson (2) Jason H Moore (1) Jonathan Haines (2) Deborah Murdock (2) Douglas Mortlock (1) Chun Li (2)

(1) Dartmouth College
(2) Vanderbilt University

Mitochondrial DNA (mtDNA) variation can affect phenotypic variation; therefore, knowing its distribution within and among individuals is of importance to understanding many human diseases. Intra-individual mtDNA variation (heteroplasmy) has been generally assumed to be random. We used massively parallel sequencing to assess heteroplasmy across ten tissues and demonstrate that in unrelated individuals there are tissue-specific, recurrent mutations. Certain tissues, notably kidney, liver and skeletal muscle, displayed the identical recurrent mutations absent from other tissues in the same individuals. Using RFLP analyses we validated one of the tissue-specific mutations in the two sequenced individuals and replicated the patterns in two additional individuals. These recurrent mutations all occur within or in very close proximity to sites that regulate mtDNA replication, strongly implying that these variations alter the replication dynamics of the mutated mtDNA genome. These recurrent variants are all independent of each other and do not occur in the mtDNA coding regions. The most parsimonious explanation of the data is that these frequently repeated mutations experience tissue-specific positive selection, probably through replication advantage.

147
A systematic evaluation of gene- and pathway-level methods for genome-wide association studies through simulations

Genevieve L Wojcik (1) WH Linda Kao (2) Priya Duggal (1)

(1) Johns Hopkins Bloomberg School of Public Health, Baltimore MD
(2) Johns Hopkins Bloomberg School of Public Health, Baltimore MD. Johns Hopkins School of Medicine, Baltimore MD.

Genome-wide association studies (GWAS) are ubiquitous in genetic epidemiology, identifying thousands of variants over the past decade. However, these GWAS do not explain a large portion of the estimated heritability. One reason for this "missing heritability" may be stringent significance thresholds correcting for multiple comparisons and minimizing type I error. Weaker association signals below this threshold may be biologically relevant if they are within a particular gene or pathway. To address this issue, methods to analyze variants in gene- and pathway-level units were developed, many as secondary analyses for GWAS data. Despite their application to various data sets, there is no consensus as to the best method. A simulation using data from the Wellcome Trust Case Control Consortium (WTCCC) was performed to systematically compare these gene- and pathway-level methods based on sensitivity, specificity, type I and II error using a polygenic additive model. The accuracy of these programs was also examined for association with gene/pathway size, minor allele frequencies of causal single nucleotide polymorphisms (SNPs), and the effect sizes of the causal SNP. The influence of sample size was also evaluated with traditionally underpowered ($n=500$) and large ($n=4500$) study sizes. Classical methods not designed explicitly to handle linkage disequilibrium had higher type I error, while newer methods that account for structure were underpowered for smaller effect sizes. All methods were able to find truly associated genes that would have been ignored by current GWAS significance thresholds. While the method depends on the user's priorities, gene- and pathway-level methods remain valuable tools that offer biological insight for GWAS.

148
Bayesian Dictionary Learning in Genetic Studies on Related Individuals

Xiaowei Wu (1) Hongxiao Zhu (1)

(1) Virginia Polytechnic Institute and State University

Dictionary learning is shown effective to characterize complex structure of data and has been widely used for data restoration in machine learning. Its application in genetic studies, however, has not been fully explored. We propose to analyze genetic data on related individuals using Bayesian dictionary learning (BDL). The proposed method relies on a probit model to link polychotomous genotypes with a set of latent variables. The mean of each latent variable is formulated as a sparse linear combination of common over-complete basis elements, which constitutes a "dictionary". We assume a beta-process prior for the coefficient sequence of the dictionary elements and develop a Markov chain Monte Carlo algorithm for posterior inference. BDL provides an integrated solution to different genetic problems including genotype imputation, allele frequency estimation and kinship coefficient estimation. For genotypes on related individuals with unknown pedigree structure, the method can effectively borrow strength from the common between-individual correlations along the genome. Simulation studies show that the proposed BDL method outperforms the best linear unbiased predictor (BLUP) on imputation, and provides estimations of allele frequencies and kinship coefficients with improved accuracy.

149
Natural and Orthogonal interaction framework for modeling gene-gene interactions applied to cutaneous melanoma

Feifei Xiao (1) Jianzhong Ma (2) Guoshuai Cai (3) Shenying Fang (4) Jeffrey E. Lee (4) Qingyi Wei (5) Christopher I. Amos (5)

(1) Department of Genetics, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA; Department of Biostatistics, Yale University School of Public Health, New Haven, CT, USA
(2) Department of Genetics, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA
(3) Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center, Houston, TX USA
(4) Department of Surgical Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA
(5) Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

Epistasis, or gene-gene interaction, is the departure from additive genetic effects of several genes on a trait; thus, the same alleles of one gene may display different genetic effects in different genetic backgrounds. In this study, we generalized the coding technique for a natural and orthogonal interactions (NOIA) model for association studies along with gene-gene interactions for dichotomous traits and human complex diseases. The NOIA model is an orthogonal model which has non-correlated estimators for genetic effects. This modeling scheme is important for estimating influence from multiple loci. We conducted simulations and real data analyses to evaluate the performance of the NOIA model. Both simulation and real data analyses revealed that the NOIA statistical model had greater power for detecting both main genetic effects and some interaction effects than the usual model. Although associated genes have been identified for predisposing people to melanoma risk: HERC2 at 15q13.1, MC1R at 16q24.3 and CDKN2A at 9p21.3, no gene-gene interaction has been fully explored for melanoma risk. By applying the NOIA statistical model to a genome-wide melanoma dataset, we confirmed the previously identified significantly associated genes. We also found potential regions at chromosome 5 and 4 that may interact with the HERC2 and MC1R genes, respectively. Our study provides not only statistical characteristics for the orthogonal NOIA model but also useful insights for understanding of the influence of interactions on melanoma risk.

150
Genetic Interaction Networks for Integrative Identification of Disease Risks in Signal Pathways Using a Nonparametric Bayes Model

Chuanhua Xing (1) Tsuyoshi Kunihama (2) Douglas Kiel (3) L. Adrienne Cupples (1) David Dunson (2)

(1) Boston University
(2) Duke University
(3) Hebrew Senior Life, Harvard Medical School

In genetic association studies, there is commonly interest in identifying genetic and environmental factors predicting a disease phenotype while accounting for important covariates. Due to the daunting dimensionality that arises in genome-wide association studies (GWAS) and next generation sequencing studies, typical analyses rely on screening of SNPs or perhaps genes one by one. However, it has been increasingly realized that higher order interactions among genetic variants and environmental factors play an important role in the development of complex diseases. Existing methods such as regression modeling and Bayes networks face problems with computational scaling and false positives, by ignoring or partially capturing such interactions. We propose a new nonparametric Bayesian approach for parsimoniously characterizing higher order interactions, while addressing the scaling and false positive concerns, using probabilistic tensor

factorizations. We extend a previous successful approach for categorical data to accommodate mixed categorical and continuous variables. We aim to derive genetic interaction networks for a better identification of disease risks by accounting for high-order interactions among genetic and environmental factors. The methods are evaluated through detailed simulation studies and applied to Osteoporosis data of fat intake by gene interaction on quantitative computed tomography derived bone density.

151
A method of differential methylation analysis of next-generation sequencing with covariates

Hongyan Xu (1) Robert H. Podolsky (1) Duchwan Ryu (1) Varghese George (1)

(1) Georgia Regents University

DNA methylation at CpG loci is an important biomedical process involved in many complex diseases including cancer. In recent years, the development of next-generation sequencing (NGS) yields large amount of DNA methylation data. We have developed a statistical approach for detecting differentially methylated CpG sites for NGS data based on clustered data analysis. However, our approach did not allow for covariates. Research has shown that DNA methylation is correlated with age, sex, and population. Therefore, it is important to account for the effect of such covariates in differential methylation analysis. In this study, we extend our method to allow for covariates based on generalized estimating equations. Simulations show that the extended test maintains correct type-I error rate and is robust under several distributions for the measured methylation levels. It improves power over our previous test. Finally, we apply the test to our NGS data on chronic lymphocytic leukemia. The results indicate that it is a promising and practical test for genome-wide methylation analysis.

152
Classification based on a permanental process

Jie Yang (1) Klaus Miescke (1) Peter McCullagh (2)

(1) University of Illinois at Chicago
(2) University of Chicago

In this talk we introduce a new classification method based on a permanental process for supervised classification problems. Regardless of the number of classes or the dimension of the feature space, the method requires only 2~3 parameters for the covariance function. The method is effective even if the feature region occupied by one class is a patchwork interlaced with regions occupied by other classes. An application to DNA microarray analysis indicates that the method is effective even for high-dimensional data. It can employ feature variables in an efficient way to reduce the prediction error significantly. This is critical when the true classification relies on non-reducible high-dimensional features.

153
Methods to compare trait-dependent sampling designs for rare-variant association analysis

Yildiz E Yilmaz (1) Jerald F Lawless (2) Shelley B Bull (3)

(1) Memorial University of Newfoundland
(2) University of Waterloo
(3) University of Toronto

Trait-dependent designs offer a means to achieve cost efficient studies of the genetic association of a quantitative trait (QT) with rare variants. Under a two-phase design, we assume a sample from an existing population-based cohort of individuals phenotyped in phase 1 will be sequenced in phase 2. We develop a framework to assess relative efficiency and power of alternative designs based on evaluation of the expected large sample non-centrality parameter (NCP) in a likelihood ratio statistic (LRS), including extreme-trait and stratified sampling designs as special cases. The latter, e.g., could over-sample the tails of the QT distribution and under-sample intermediate trait values. Given the QT distribution and a postulated rare variant score distribution with associated genetic effect, design considerations focus on specification of the number and size of strata, and on the allocation of the phase 2 sample to the specified strata (effectively the stratum-specific sampling fractions). The LRS is derived from a likelihood based on data (Yi, vi) for observation i, assuming that, in the population, Yi is normally distributed with mean b0 + b1vi, and Pr (vi = v) is the distribution of a burden-based rare variant score v for v = 0,1,…, J. For a particular strata specification, we can compare an array of potential sample allocations, identify the one that maximizes the power of the LRS for a test of b1 by maximizing the expected NCP, and evaluate consequences of choosing an allocation that may be more robust to model misspecification. In addition, we can assess sensitivity of the sample allocation and expected power to the postulated rare variant effect size.

154
Real Time Classification of Viruses in 12 Dimensions

Chenglong Yu (1) Troy Hernandez (1) Hui Zheng (1) Shek-Chung Yauz (2) Hsin-Hsiung Huang (1) Rong He (3) Jie Yang (1) Stephen Yau (4)

(1) Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago
(2) Information Technology Services Center, Hong Kong University of Science and Technology
(3) Department of Biological Sciences, Chicago State

University
(4) Department of Mathematical Sciences, Tsinghua University

The rapid development of sequencing technologies produces a large number of viral genome sequences. Characterizing genetic sequences and determining viral origins have always been important issues in virology. The study of sequence similarity at the interfamily level is especially crucial for revealing key aspects of evolutionary history. The International Committee on Taxonomy of Viruses authorizes and organizes the taxonomic classification of viruses. Thus far, the detailed classifications for all viruses are neither complete nor free from dispute. For example, the current missing label rates in GenBank are 12.1% for family label and 30.0% for genus label. Using the proposed Natural Vector representation, all 2,044 single-segment referenced viral genomes in GenBank can be embedded in R^12. Unlike other approaches, this allows us to determine phylogenetic relations for all viruses at any level (e.g., Baltimore class, family, subfamily, genus, and species) in real time. Additionally, the proposed graphical representation for virus phylogeny provides a visualization of the distribution of viruses in R^12. Unlike the commonly used tree visualization methods which suffer from uniqueness and existence problems, our representation always exists and is unique. This approach is successfully used to predict and correct viral classification information, as well as to identify viral origins; e.g. a recent public health threat, the West Nile virus, is closer to the Japanese encephalitis antigenic complex based on our visualization. Based on cross-validation results, the accuracy rates of our predictions are as high as 98.2% for Baltimore class labels, 96.6% for family labels, 99.7% for subfamily labels and 97.2% for genus labels.

155
A more powerful method for mixed-model case-control association analysis with covariates, related individuals and missing data

Sheng SZ Zhong (1) Mary Sara MSM McPeek (2)

(1) Department of Statistics, University of Chicago
(2) Department of Statistics and Department of Human Genetics, University of Chicago

We consider the problem of testing for association between a binary trait and a SNP, while adjusting for relevant covariates, in a sample with related individuals, where missing data are allowed. We propose a novel estimating equation approach that can be effectively viewed as a hybrid of logistic regression and linear mixed-effects model (LMM) approaches. In various simulated scenarios, the hybrid method outperforms the approaches based on either logistic regression or LMMs. Our method results in an even greater power increase over the other methods when the sample includes related individuals with missing data (genotype, phenotype or covariates), because we make full use of the relationship information by incorporating partially missing data in the analysis while correcting for dependence. Unlike another type of methods, built on generalized linear mixed models (GLMMs), our estimating equation approach is computationally feasible for genome-wide analysis in large complex pedigrees. Furthermore, our method is robust to the phenomenon of reduced power when non-confounding covariates are included in case-control association studies with low disease prevalence. The method can be extended to allow an empirical covariance matrix for analysis of unrelated individuals, to allow for population structure or cryptic relatedness.

156
A unified rare variant association approach for qualitative and quantitative traits using both family and unrelated samples

Xiaofeng Zhu (1) Tao Feng (1)

(1) Case Western Reserve University

Many statistical methods for analyzing rare variant association have been developed but most of them focus on unrelated samples. Incorporating family data in rare variant association analysis has become increasingly appreciated. Here we introduce a computation efficient rare variant association approach, which can be applied to qualitative and quantitative traits using family and unrelated samples. Our simulation study suggests the proposed method has more power than famSKAT when only rare variants are associated with a trait. We demonstrated that family data are useful in searching for very rare variants underlying complex traits.