Mini-Symposium

Genomics-based Personalized Medicine

M 1

Application of genomic tests in breast cancer management

Martin Filipits[1]
[1]Medical University of Vienna, Institute of Cancer Research, Vienna, Austria

Breast cancer is a heterogeneous disease at the clinical, biological and particularly at the molecular level. Gene expression profiling has improved the knowledge on the complex molecular background of this disease and allows a more accurate prognostication and patient stratification for therapy. Several genomic tests have been developed with the aim of improving prognostic information beyond that provided by classic clinicopathologic parameters. Some of these tests are currently available in the clinic and are used to determine prognosis and more importantly to assist in determining the optimal treatment in patients with hormone receptor-positive breast cancer.

Available data suggest that information generated from genomic tests has resulted in a change in decision making in approximately 25%-30% of cases. The clinical relevance of genomic tests and their ability to define prognosis and determine treatment benefit will be discussed.

M 2

Risk prediction models using family and genomic data

Joan E Bailey-Wilson[1]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, United States of America

Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, United States of America}{Advances in our ability to model personal risk of developing a disease have accelerated as large epidemiologic and genomic studies have increased our understanding of disease causation. Prediction of disease risk can be based on personal history of environmental exposures, family history of disease and personal genotypes at genetic susceptibility loci. Approaches to predicting risk of disease that utilize familial and genetic information will be discussed for a range of different causal models from simple Mendelian disorders that are caused by variants in a single gene to diseases caused by complex actions of multiple risk factors. The utility of adding family history and personal genotypes into disease risk models will be covered. Accurate disease risk prediction can be important to individual health since it can encourage individuals to have more frequent screening procedures, to undertake environmental risk reduction, and to undergo preventive medical procedures and treatments.

M 3

The importance of appropriate quality control in - omics studies as required for personalized and stratified medicine

Bertram Müller-Myhsok[1,2,3]
[1]Max Planck Institute of Psychiatry, Munich, Germany
[2]Munich Cluster for Systems Neurology (SyNergy), Munich, Germany
[3]Institute for Translational Medicine, University of Liverpool, Liverpool, United Kingdom

Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, United States of America}{Advances in our ability to model personal risk of developing a disease have accelerated as large epidemiologic and genomic studies have increased our understanding of disease causation. Prediction of disease risk can be based on personal history of environmental exposures, family history of disease and personal genotypes at genetic susceptibility loci. Approaches to predicting risk of disease that utilize familial and genetic information will be discussed for a range of different causal models from simple Mendelian disorders that are caused by variants in a single gene to diseases caused by complex actions of multiple risk factors. The utility of adding family history and personal genotypes into disease risk models will be covered. Accurate disease risk prediction can be important to individual health since it can encourage individuals to have more frequent screening procedures, to undertake environmental risk reduction, and to undergo preventive medical procedures and treatments.

M 4

Study Designs for Predictive Biomarkers

Andreas Ziegler[1,2]
[1]Institute of Medical Biometry and Statistics, University of Lübeck, University Medical Center Schleswig-Holstein, Campus Lübeck, Lübeck, Germany
[2]Center for Clinical Trials, University of Lübeck, Lübeck, Germany

Biomarkers are of increasing importance for personalized medicine, including diagnosis, prognosis and targeted therapy of a patient. Examples are provided for current use of biomarkers in applications. It is shown that their use is extremely diverse, and it varies from pharmacodynamics to treatment monitoring. The particular features of biomarkers are discussed. Before biomarkers are used in clinical routine, several phases of research need to be successfully passed, and important aspects of these phases are considered. Some biomarkers are intended to predict the likely response of a patient to a treatment in terms of efficacy and/or safety, and these biomarkers are termed predictive biomarkers or, more generally, companion diagnostic tests. Using examples from the literature, different clinical trial designs are introduced for these biomarkers, and their pros and cons are discussed in detail.

Educational Workshop

Pharmacogenomics: When Drug Response Gets Personal

E 1

Pharmacogenomics: Past, Present and Future

Brooke Fridley[1]
[1]University of Kansas Medical Center, USA

Pharmacogenetics is the study of the role of inheritance in individual genetic variation in response to drugs. In this post-genomic era, pharmacogenetics has evolved into pharmacogenomics, the study of the influence of genetic variation across the entire genome on drug response. Pharmacogenomics has been heralded as one of the first major clinical applications of the striking advances that have occurred and continue to occur in human genomic science. In this talk, I will provide an overview of pharmacogenomics and discuss the past, present and future of pharmacogenomics in the 21st century.

E 2

Assessing the genetic basis of drug response

John Witte[1]
[1]University of California, San Francisco, USA

By definition pharmacogenomic traits have an underlying genetic basis. Nevertheless, accurately estimating the heritability of drug response is important for designing studies and knowing how much genetic variation can or has been explained. Unlike most quantitative and qualitative traits, however, response to treatment has two unique, complicating factors: it is a gene-drug interaction and the outcome is often in terms of time-to-event. Here I will present and apply methods that address these two aspects when estimating the genetic basis (or heritability) of pharmacogenomic traits.

E 3

Clinical Utility in Pharmacogenomics: Getting Beyond Individual Variants

Hae Kyung Im[1]
[1]University of Chicago, USA

Studies in pharmacogenomics have identified many individual variants with sufficiently large effect sizes to have clinical utility, and many of these are now the subject of implementation studies at a variety of levels. Recent research on common diseases and complex traits have, however, raised the possibility that mixed models allowing separately for the contribution of variants with larger effect sizes and a polygenic background may yield improved prediction. As we medical centers routinely move to having large-scale genome data routinely available on patients, as opposed to one-off genotyping for the prescribing of specific drugs, the opportunity to build predictors of adverse events and efficacy using large scale genome data rather than individual (or small numbers of) variants becomes a real possibility. Using real examples from large-scale studies, we will contrast prediction based on individual or small numbers of variants with predictions based on large-scale information. We will also discuss efforts to implement these alternative approaches in EMR settings.

E 4

Smoking behavior and lung cancer risk related to nicotinic acetylcholine receptor variants and metabolic variants.

Christopher I Amos[1]
[1]Geisel School of Medicine, USA

In this presentation I contrast the discovery of genetic variants that influence smoking behavior including initiation, daily consumption and cessation. The most prominent associations are with the nicotinic acetylcholine receptor gene family on chromosome 15q25.1. These genes along with CYP2A6 strongly influence smoking behavior and also affect lung cancer risk. I will describe the striking impact that variation in these genes appears to have on the efficacy of pharmacological interventions to influence smoking cessation. Finally, I will describe studies of lung cancer risk and how these genes relate to it, along with a further discussion of the potential relevance of novel associations recently discovered for squamous lung cancer that may influence chemotherapeutic responses.

Invited Speakers

I 1

Enrichment designs for the development of personalized medicine

Martin Posch[1]
[1]Vienna Medical University, Austria

If the response to treatment depends on genetic biomarkers, it is important to identify (sub)populations where the treatment has a positive benefit risk balance. One approach to identify relevant subpopulations are subgroup analyses where the treatment effect is estimated in biomarker positive and biomarker negative groups. Subgroup analysis are challenging because different types of risks are associated with inference on subgroups: On the one hand, ignoring a relevant subpopulation one could miss a treatment option due to a dilution of the treatment effect in the full population. Even, if the diluted treatment effect can be demonstrated in an Overall population, it is not ethical to treat patients that do not benefit from the treatmemt, if they can be identified in advance. On the other hand selecting a spurious sub-population is not without risk either: it might increase the risk to approve an inefficient treatment (inflating the type 1 error rate), or may wrongly lead to restricting an efficient treatment to a too narrow fraction of a potential benefiting population. The latter can not only lead to reduced revenue from the drug, but is also unfavourable from a public health perspective. We investigate these risks for non-adaptive study designs that allow for inference on subgroups using multiple testing procedures as well as adaptive designs, where subgroups may be selected in an interim analysis. Quantifying the risks with utility functions the characteristics of such adaptive and non-adaptive designs are compared for a range of scenarios.

I 2

Causal association structures in -omics data: how far can we get with statistical modeling?

Krista Fischer[1]
[1]Tartu University, Estonia

This talk mainly concentrates on the setting where association of one genotype marker (typically SNP) with two correlated phenotypes is studied. In so-called "Mendelian Randomization" studies the main parameter of interest corresponds to a causal effect of one phenotypic trait on another trait, whereas a genetic marker is used as an instrument. Despite of the increasing number of publications using this methodological approach, the underlying assumptions are often overlooked. Therefore, many of the published effect estimates may actually be biased and misleading. One of the main untestable assumptions is the "no pleiotropy" assumption - the genotype has a direct causal effect on one phenotype only, whereas the effect on the second phenotype is fully mediated by the first one.

When this is not fulfilled, the genotype is said to have a pleiotropic effect on both phenotypes, whereas another class of models is been designed to estimate such effects. However, we will show that mathematically one cannot distinguish between the two models: the model underlying the Mendelian Randomization scenario and the model for pleiotropic effect. We will discuss whether some sensitivity analysis methods may help to draw a correct conclusion here.

In addition, we discuss another assumption underlying the Mendelian Randomization idea: the "no-treatment-effect heterogeneity" assumption. Here a parallel can be drawn with randomized clinical trials, where this assumption is crucial to allow for active treatment on the control arm. Using also simulation results, the effect of deviations from this assumption is studied.

I 3

The relevance of epigenomics for personalized medicine

Christoph Bock[1]
[1]Research Center for Molecular Medicine of the Austrian Academy of Sciences, Austria

In my presentation, I will summarize the role of next generation sequencing for personalized medicine and highlight the relevance of bioinformatic and biostatistical methods for interpreting the vast amount of genome, epigenome and transcriptome data that are being generated at CeMM and at many genomics institutes world-wide. The talk will also discuss our ongoing work with the European BLUEPRINT project consortium (http://blueprint-epigenome.eu/) aimed at establishing comprehensive epigenome maps of hematopoietic cell types and various types of leukemia cells. I will conclude by outlining an integrated computational/experimental approach toward rational design of epigenetic combination therapies (Bock and Lengauer 2012 Nature Reviews Cancer), which we pursue in collaboration between the CeMM Research Center for Molecular Medicine and the Medical University of Vienna.

I 4

Fine mapping of complex trait loci with coalescent methods in large case-control studies

Ziqian Geng[1], Paul Scheet[1], Sebastian Zöllner[1]
[1]University of Michigan, USA

Case-control studies are widely used to identify genomic regions containing disease variants. However, identifying the underlying risk variants for complex diseases is challenging due to the complicated genetic dependence structure caused by linkage disequilibrium (LD). By modeling the evolutionary process of a target region, coalescent-based approaches improve this identification by using all available haplotype information. Such methods estimate the genealogy at all sites in the region and thus model the probability of carrying risk variants at all loci jointly. From these probabilities we obtain Bayesian confidence intervals (CIs) where true risk variants are most likely to occur. Additionally, the genealogy at each position provides more information about the shared ancestry of neighboring sites. Indeed, such careful modeling of the shared ancestry of sequences is also beneficial in haplotyping and variant calling in regions of interests (ROI) where traditional hidden Markov approaches struggle. However, existing coalescent-based methods are computationally very challenging and can only be applied to samples below 200 individuals. Here, we propose a novel approach to overcome this difficulty, so that it can be applied to large-scale studies. First, we infer a set of clusters from the sampled haplotypes so that haplotypes within each cluster are inherited from a common ancestor. Then, we apply coalescent-based approaches to approximate the genealogy of ancient haplotypes at different positions across the ROI. Doing so, the dimension of external nodes in coalescent models is reduced from the total sample size to the number of clusters. Finally, we evaluate the position-specific cluster genealogy and their descendants' phenotype distribution, to integrate over all positions and establish CIs where risk variants are most likely to occur. In simulation studies, our method correctly localizes short segments around true risk positions for both rare (1%) and common (5%) risk variants in datasets with thousands of individuals. In summary, we have developed a novel approach to estimate the genealogy throughout sequenced regions. In fine mapping of complex trait loci, our method is applicable for large-scale case-control studies using sequencing data.

I 5

The interface hypothesis in explaining host-bacterial interactions in the human gut

Knut Rudi[1]
[1]Norwegian University for Life Sciences, Norway

Our gut microbiota is tremendously complex, outnumbering the host cells by a factor of ten and the number of genes by a factor of one hundred. The gut microbiota serves the main functions of extracting energy from the food, production of vitamins and other (essential) biomolecules, in addition to protection towards pathogens. However, despite major efforts we do still not know the basic mechanisms for host-bacterial interactions in the gut. We have therefore recently proposed the interface hypothesis, advocating the importance of positive host selection for mutualistic gut bacteria. I will present details about the hypothesis, and how it is supported from the current knowledge about the human gut microbiota.

Neel and Williams Award Candidates

A 1

A novel method using cross pedigree shared ancestry to map rare causal variants in the presence of locus heterogeneity

Haley J Abel[1], Michael A Province[1]
[1]Division of Statistical Genomics, Washingon University School of Medicine, St. Louis, MO

Currently, there is great interest in the use of family studies to identify rare variants underlying complex disease. However, attempts at fine mapping are confounded by locus heterogeneity, which results in noisy and poorly localized linkage signals, and an abundance of rare variants, which frequently segregate with phenotype by chance. As rare variants shared across pedigrees are likely to be of recent origin, we have developed an approach leveraging identity-by-descent (IBD) between pedigree founders to better localize linkage signals in the presence of heterogeneity. Our method relies on segments shared identically-by-state (IBS) across pedigrees: it optimizes over pedigree members and calculates a score based on the sum of maximal pairwise shared lengths at each locus. Use of unphased IBS makes it both computationally efficient, so that pedigree-based permutation tests assessing significance are tractable, and robust to genotyping and haplotype phase switch errors. Moreover, our method provides a cross-family metric to permit local clustering of families near IBD regions: this allows stratification by recent shared ancestry, and, in simulations, accurately recovers ancestral relationships. We have evaluated the performance of our method by coalescent simulation of founder individuals, followed by gene-dropping onto pedigrees. Under a variety of scenarios, with rare causal variants (MAF<0.01) and modest effect sizes (OR=5-7), our approach achieves 60-80% power, and is able to detect shared ancestral segments harboring rare causal variants where multipoint linkage and rare-variant burden tests fail.

Categories: *Coalescent Theory, Heterogeneity, Homogeneity, Linkage and Association*

A 2

Survival analysis with delayed entry in selected families with application to human longevity.

Mar Rodriguez Girondo[1], Jeanine Houwing-Duistermaat[1]
[1]LUMC, The Netherlands

Although there is evidence from several studies that longevity aggregates within families, identification of genetic factors has not been successful. Reasons for lack of progress might be the ad hoc definition of being older than a specific threshold (e.g. older than 90 years of age). As alternative we will consider survival models for the analysis of longevity in family studies. Challenges are to model the ascertainment of the families, to take into account correlation between family members and to deal with delayed entry. Methods for survival analysis with delayed entry in small clusters are available (e.g. Rondeau et al, 2012). These methods provide bias estimates for larger clusters (Jensen et al, 2004), because they do not adjust for ascertainment. We propose a Cox model with a frailty and with inverse probability weighting to account for the selection of the families and the delayed entry. The weights will be based on the latent frailties in a proportional hazards model. Via simulations we showed that our approach performs better (less bias) than existing methods for large families (>8 subjects) and large frailties (>0.5). This work is motivated by the Leiden Longevity study comprising 420 families with at least two nonagenarian siblings. The size of sibships with members who become older than 60 years varies from 2 to 13 siblings. The maximum observed age is 107 years and 13% of the nonagenarians is still alive. We estimated the effect of APOE E4 allele on survival. The estimate of the variance of the frailty was 0.082 and 0.230 for the standard approach and our approach respectively. The estimate of the log hazard ratio was -0.272 (s.e. 0.113) and -0.212 (s.e. 0.070) for the standard and our approach respectively.

Categories: *Ascertainment, Association: Family-based, Heritability*

A 3

Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees

Mohamad Saad[1], Ellen M Wijsman[1]
[1]Department of Biostatistics, University of Washington

In the last two decades, complex traits have become the main focus of genetic studies. The hypothesis that both rare and common variants are associated with complex traits is increasingly being discussed. Family-based association studies using relatively large pedigrees are suitable for both rare and common variant identification. Because of the high cost of sequencing technologies, imputation methods are important for increasing the amount of information at low cost. A recent family-based imputation method, GIGI, is able to handle large pedigrees and accurately impute rare variants, but does less well for common variants where population-based methods (e.g.; BEAGLE) perform better and can also be used. We propose a flexible approach to combine imputation data from family- and population-based methods. We select, for every SNP and every subject, the set of 3 genotype posterior probabilities from the method with the highest variance of these probabilities. We also extend the association test SKAT-RC, originally proposed for data from unrelated subjects, to family data with continuous trait in order to make use of such imputed data. We call this extension "famSKAT-RC". We compare the performance of famSKAT-RC and several other existing burden and kernel association tests. In simulated pedigree sequence data, our results show an increase of imputation accuracy from the combined approach. Also, the data show an increase of power of the association tests with this approach over the use of either family- or population-based imputation methods alone, in the context of rare and common variants in a single gene. Moreover, our results showed better performance of famSKAT-RC compared to the other considered tests, in most scenarios investigated.

Categories: *Association: Family-based, Multiple Marker Disequilibrium Analysis, Sequencing Data*

A 4

Mixed modeling for time-to-event outcomes with large-scale population cohorts and genome-wide data

Christian Benner[1,2], Matti Pirinen[1], Emmi Tikkanen[1,2], Samuli Ripatti[1,2,3]
[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
[2]Hjelt Institute, University of Helsinki, Helsinki, Finland
[3]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Recent development on linear mixed models has provided a common framework for heritability estimation, multi-locus association testing and genomic prediction of quantitative traits in population cohorts of unrelated individuals. The possibility to use population cohorts rather than family structures could open up new avenues also for time-to-event outcomes in genetic epidemiology. However, connecting time-to-event outcomes to big genomics data has so far not been computationally feasible with hitherto existing software. We introduce a novel survival analysis method for heritability estimation, multi-locus association testing and genomic risk prediction that scales to millions of genetic markers and health events in tens of thousands of individuals. Motivation for our work comes from a large and unique collection of Finnish population cohorts for which we have both detailed genomic and comprehensive health registry data. Our approach implements a very flexible piecewise constant hazard model that contains an individual-specific Gaussian random effect with an arbitrary covariance structure. Computationally, we transform the problem to a Poisson model, which we analyze by fitting a hierarchical generalized linear model. We demonstrate the runtime efficiency of our method and give an example of heritability estimation and multi-locus association testing for cardiovascular disease related events using up to 16,000 Finnish individuals. Our work extends the computational tractability of linear mixed models from quantitative traits to time-to-event outcomes and will prove useful, e.g., for combining information across individuals' genomes and their hospital records.

Categories: *Association: Genome-wide, Cardiovascular Disease and Hypertension, Heritability, Maximum Likelihood Methods*

A 5

The collapsed haplotype pattern method for linkage analysis of next-generation sequencing data

Gao T Wang[1], Di Zhang[1], Biao Li[1], Hang Dai[1], Suzanne M Leal[1]
[1]Baylor College of Medicine

Traditionally, linkage analysis was used to map Mendelian diseases. Next generation sequencing (NGS) makes it possible to directly sequence individuals with Mendelian diseases and identify causal variants by filtering. Linkage analysis of SNP data are sometimes used in conjunction with NGS to increase the success of identifying the causal variant. With the reduction in cost of NGS, DNA samples from multiple families can be sequenced and linkage analysis can be performed directly using NGS data. Inspired by "burden" tests for complex trait rare variant association studies, we developed the collapsed haplotype pattern (CHP) method to generate markers from sequence data for linkage analysis. To demonstrate the power of the CHP method we analyzed and performed power calculations using data from several deafness genes. Power analysis showed that the CHP method is substantially more powerful than analyzing individual SNVs. Specifically for an autosomal recessive model with allelic heterogeneity and locus heterogeneity of 50%, it requires 12 families for the CHP method to achieve a power of 90% for the SLC26A4 gene, while analyzing individual SNVs requires >50 families to achieve the same power. Unlike the commonly practiced filtering approaches used for NGS data, the CHP method provides statistical evidence of the involvement of a gene in Mendelian disease etiology. Additionally because it incorporates inheritance information and penetrance models it is less likely than filtering to exclude causal variants in the presents of phenocopies and/or reduced penetrance. We recommend the use of the CHP method in parallel to filtering methods to take full advantage of the power of NGS in families.

Categories: *Linkage Analysis, Sequencing Data*

A 6

Meta-analysis approach for haplotype association tests: a general framework for family and unrelated samples

Shuai Wang[1], Jing H Zhao[2], Mark O Goodarzi[3], Josée Dupuis[1]
[1]Department of Biostatistics, Boston University School of Public Health, Boston, MA
[2]MRC Epidemiology Unit, University of Cambridge, Institute of Metabolic Science, Addenbrooke's Hospital, Box 285 Hills Road, Cambridge, United Kingdom
[3]Division of Endocrinology, Diabetes and Metabolism, Cedars-Sinai Medical Center, Los Angeles, CA

Meta-analysis has been widely used to improve power to detect associated variants in genome-wide association studies. Several meta-analysis methods have been developed and successfully applied to combine association tests of single variant and gene-based tests from multiple cohorts. However, meta-analysis of haplotype association results remains a challenge, because different haplotypes may be observed across cohorts. We propose a two-stage meta-analysis approach to combine haplotype analysis results. Our approach allows each cohort to contribute association results from uniquely observed haplotypes, in addition to haplotypes observed in multiple cohorts. In the first stage, each cohort computes the expected haplotype effects in a regression framework, selecting the most frequent haplotype, which can vary across cohorts, as the reference haplotype and including a random familial effect to account for relatedness, if appropriate. For the second stage, we propose a multivariate generalized least square meta-analysis approach to combine haplotype effects from multiple cohorts. Association tests for each haplotype and a global test can be obtained within our framework. A simulation study shows that our approach has the correct type I error. We present an application to genotypes from Illumina HumanExome Beadchip array, where we assess the association between haplotypes formed by rare variants in a fasting glucose-associated locus (G6PC2). We then combined haplotype analysis results from 18 CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium) cohorts. The global haplotype association test is highly significant (p=1.1e-17), and more significant than any single-variant and gene-based tests.

Categories: *Association: Family-based, Diabetes, Haplotype Analysis, Quantitative Trait Analysis*

Contributed Platform Presentations

C 1

Identification of blood pressure (BP) related candidate genes by population-based transcriptome analyses within the MetaXpress Consortium

Christian Müller[1], Katharina Schramm[2], Claudia Schurmann[3], Soonil Kwon[4], Arne Schillert[5], Christian Herder[6], Georg Homuth[3], Simone Wahl[7], Harald Grallert[7], Andreas Ziegler[5]
[1]General and Interventional Cardiology, University Heart Center Hamburg, Germany
[2]Institute of Human Genetics, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany
[3]Interfaculty Institute for Genetics and Functional Genomics, University Medicine and Ernst-Moritz-Arndt-University Greifswald, Greifswald, Germany
[4]Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, USA
[5] Institute of Medical Biometry and Statistics, University of Lübeck, University Medical Center Schleswig-Holstein, Campus Lübeck, Lübeck, Germany
[6]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Germany
[7]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

High blood pressure (BP) is a global major risk factor for cardiovascular diseases. We analyzed associations between the transcriptome and BP traits in large cohorts of the MetaXpress Consortium. Transcriptomic data from the Illumina HumanHT-12 BeadChip array were available for 4533 individuals from the three German cohorts and one US cohort. Expression levels were measured in monocyte (n=2549, Gutenberg Health Study (GHS)) and whole blood cell (n=1984 Cooperative Health Research in the Region of Augsburg (KORA F4) and Study of Health in Pomerania (SHIP-TREND), Multi-Ethnic Study of Atherosclerosis (MESA)). Associations to systolic BP (SBP), diastolic BP (DBP) and pulse pressure (PP) were estimated by linear regression with adjustments for sex, age, body mass index (BMI), RNA storage time, amplification layout and RNA integrity number within each study. A pooled analysis was conducted within GHS and MESA using the inverse variance method. Significant associations (FDR≤0.05) were selected for replication in KORA F4 and SHIP-TREND. Genes with consistent effect directions and p≤0.05 in both initial studies were selected as candidates. In total, 8 unique genes were consistently associated with systolic blood pressure (SBP), diastolic blood pressure (DBP) or pulse pressure (PP) in both discovery and replication steps: CEBPA, CRIP1, F12, LMNA, MYADM, TIPARP, TPPP3 and TSC22D3. In total, the candidate genes explained between 4-13%, 4-6% and 2-8% of inter-individual variance of SBP, DBP and PP, respectively. This is the first study investigating the associations between BP traits and whole transcriptomes in more than 4000 individuals. The comprehensive analyses highlight eight genes which are associated with BP.

Categories: *Association: Candidate Genes, Association: Genome-wide, Cardiovascular Disease and Hypertension, Gene Expression Arrays, Gene Expression Patterns*

C 2

Mixed-model analysis of common variation reveals pathways explaining variance in AMD risk

Jacob B Hall[1], Margaret A Pericak-Vance[2], William K Scott[2], Jaclyn L Kovach[2], Stephen D Schwartz[2], Anita Agarwal[3], Milam A Brantley[3], Jonathan L Haines[1], William S Bush[1]
[1]Institute for Computational Biology, Case Western Reserve University, Cleveland, OH
[2]John P Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL
[3]Department of Ophthalmology and Visual Sciences, Vanderbilt University, Nashville, TN

Age-related macular degeneration (AMD) is the leading cause of irreversible blindness in the elderly in developed countries and can affect more than 10% of individuals over age 80. AMD has a large genetic component, with heritability estimated to be between 45% & 70%. Numerous loci have been identified and implicate various molecular mechanisms and pathways in AMD pathogenesis. Eight pathways, including angiogenesis, antioxidant activity, apoptosis, complement activation, inflammatory response, nicotine metabolism, oxidative phosphorylation, and the tricarboxylic acid cycle, were selected for our study based on an extensive literature review. While these pathways have been proposed in literature, the overall extent of the contribution to AMD heritability for each pathway is unknown. In a case-control dataset, we used Genome-wide Complex Trait Analysis (GCTA) to estimate the proportion of variance in AMD risk explained by all SNPs in each pathway. SNPs within a 50 kb region flanking each gene were assessed, as well as more distant, putatively regulatory SNPs, based on data from the ENCODE project. We found that 19 established AMD risk SNPs contributed to 13.3% of the variation in risk in our dataset, while the remaining 659,181 SNPs contributed to 36.7%. Adjusting for these 19 risk SNPs, the complement activation and inflammatory response pathways explained a statistically significant proportion of additional variance in AMD risk (9.8% and 17.9%, respectively), with other pathways showing no significant effects (0.3% – 4.4%). Our results show that additional variants associated with complement activation and inflammation genes contribute to AMD risk, and that these variants are likely in coding and nearby regulatory regions.

Categories: *Case-Control Studies, Heritability, Maximum Likelihood Methods, Multilocus Analysis, Pathways*

C 3

A Phenome-Wide Association Study of Numerous Laboratory Phenotypes in AIDS Clinical Trials Group (ACTG) Protocols

Anurag Verma[1], Sarah A Pendergrass[2], Eric S Daar[3], Roy M Gulick[4], Richard Haubrich[5], Gregory K Robbins[6], David W Hass[7], Marylyn D Ritchie[1]
[1]The Pennsylvania State University, University Park, Pennsylvaina, USA
[2]The Pennsylvania State University, University Park, PA, USA
[3]Department of Medicine, Los Angeles Biomedical Research Institute, Harbor-UCLA Medical Center, Torrance, California, USA
[4]Weill Medical College of Cornell University New York, New York, USA
[5]University of California San Diego, San Diego, California, USA
[6]Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[7]Vanderbilt University, Nashville, Tennessee, USA

Phenome-Wide Association Studies (PheWAS) have the potential to efficiently discover novel genetic associations across multiple phenotypes. Prospective clinical trials data offer a unique opportunity to apply PheWAS to pharmacogenomics. Here we describe the first PheWAS to explore associations between genotypic data and clinical trial data, both pre-treatment and following initiation of antiretroviral therapy. A "pre-treatment" PheWAS considered 27 laboratory variables from 2807 subjects who had participated in 4 ACTG protocols (ACTG384, A5142, A5095 and A5202), and analyzed ~ 5M imputed SNPs. Lowest p-values were for pre-treatment bilirubin, neutrophil counts, and HDL cholesterol levels. These and multiple other laboratory variables matched associations in the NHGRI GWAS Catalog. An "on-treatment" PheWAS considered data from 1181 subjects from A5202. We considered 838 phenotypes and sub-phenotypes derived from 6 variables: CD4 counts, HIV control, fasting LDL, fasting triglycerides, efavirenz pharmacokinetics (PK), and atazanavir PK. We considered 2,374 annotated drug-related SNPs from PharmGKB. Of 23 associations with the lowest p-values (by phenotype), 21 (91%) were with genes with matching biological plausibility: LDL with LPL and APOE; triglycerides with LPL; CD4 counts with innate immune response gene TNF, HIV control with adaptive immune response gene HLA-DRQA1, efavirenz PK with CYP2B6; atazanavir PK with drug transporter gene ABCC4. This analysis highlights the potential utility of PheWAS to evaluate clinical trials datasets for genetic associations.

Categories: *Association: Candidate Genes, Association: Genome-wide, Association: Unrelated Cases-Controls, Bioinformatics, Case-Control Studies, Epigenetics, Multivariate Phenotypes, Population Genetics, Population Stratification*

C 4

eMERGE Phenome-Wide Association Study (PheWAS) Identifies Clinical Associations and Pleiotropy for Functional Variants

Anurag Verma[1], Shefali S Verma[1], Sarah A Pendergrass[1], Dana C Crawford[2], David R Crosslin[3], Helena Kuivaniemi[4], William S Bush[2], Yuki Bradford[5], Iftikhar Kullo[6], Sue Bielinski[6]
[1]Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA
[2]Case Western University, Cleveland, OH, USA
[3]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA, USA
[4]Geisinger Health System, Danville, PA, USA
[5]Vanderbilt University, Nashville, TN
[6]Mayo Clinic, Rochester, MN, USA

We performed a phenome-wide association study (PheWAS) exploring the association between stop-gained genetic variants and a comprehensive group of phenotypes to identify novel associations and potential pleiotropy. Using multiple bioinformatics tools we selected 38 functionally relevant stop-gained/null genetic variants within the genotypic data of 37,972 unrelated patients from seven study sites in the Electronic Medical Records and Genomics (eMERGE) Network. We calculated comprehensive associations between these variants and case-control status for 3,518 ICD9 diagnosis codes (requiring ≥ 3 visits per individual to identify case status, ≥10 case subjects per ICD9 code). Associations were adjusted for age, sex, site, platform and the first 3 principal components. A total of 418 associations passed a liberal significance threshold of p< 0.01. The most significant association was between GLG1 rs9445 and "chronic non-alcoholic liver disease" (p=4.12x10-5, β=2.60). We identified many potentially pleiotropic associations at p< 0.01, 35 out of 38 SNPs demonstrated associations with more than one phenotype, and 17 SNPs were each associated with > 10 different ICD9 codes. For example, we found associations for IL34 rs4985556 with 25 diagnoses, such as "lupus erythematosus" (p=5.94 x10-3, β=0.98) and for GBE1 rs2229519 with 33 diagnoses, such as "hypertension" (p=1.2x10-3, β=0.067), "hyperlipidemia" (p=6.66 x10-3, β=0.058), and "ocular hypertension" (2.49 x10-3, β=0.21). We will seek replication of these results. In conclusion, our PheWAS shows stop-gained variants may have important pleiotropic effects, and that PheWAS are a powerful strategy to mine the full potential of the EMR for genome-phenome associations.

Categories: *Association: Genome-wide, Multilocus Analysis*

C 5

A novel G-BLUP-like phenotype predictor leveraging regional genetic similarity and its applications in predicting disease severity and drug response

Quan Long[1], Eli A Stahl[2], Jun Zhu[1]
[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai
[2]Department of Psychiatry, Icahn School of Medicine at Mount Sinai

Clinical uses of phenotype predictions based on genotype (e.g., Personalized Medicine) are emerging, empowered by high-throughput technology. It is well known that disease severity and drug response differ significantly across populations or individual patients with different genetic background. Predicting such phenotypes using seemingly unrelated samples, then stratifying patients based on these predictions, could be crucial for the design of clinic trails. There are two major active branches of genotype-based phenotype predictions based on whole genome regression. One is model selection, in which all genetic markers are modeled (usually in conjunction with Bayesian or other variable selection criteria), and which may suffer from overfitting due to astronomical number of combinations of variables/markers; the other is G-BLUP based on random effects regression, fitting phenotypic variance by kinship matrix of the sample estimated from genotypic similarity, which may run the risk of underfitting for complex traits for which infinitesimal model does not hold. We developed a G-BLUP like predictor that strikes the balance on the above trade-off. Based on GWAS signals or biological a priori knowledge, a few regions are selected and their phenotypic contributions estimated by G-BLUP. Then, model selection is applied to specify weights for the different regions. Using simulations, we demonstrate that the present predictor significantly improves prediction power in general and investigate conditions under which it performs best or not compared with pure model selection or standard G-BLUP. We apply this model to real data for various traits of multiple diseases, focusing on disease severity and drug response.

Categories: *Quantitative Trait Analysis*

C 6

Mitochondrial GWA analysis in several complex diseases using the KORA population

Antonia Flaquer[1], Karl-Heinz Ladwig[2], Rebecca Emeny[2], Melanie Waldenberger[3], Harald Grallert[3], Stephan Weidinger[4], Christa Meisinger[5], Thomas Meitinger[6], Annette Peters[2], Konstantin Strauch[7]
[1]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany.
[2]Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
[3]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
[4]Department of Dermatology, Allergology and Venerology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany
[5]Myocardial Infarction Registry, Augsburg, Germany
[6]Institute of Human Genetics, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
[7]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universitüt, Munich, Germany

Mutations of mitochondrial DNA (mtDNA) are under a growing scientific spotlight; scientists believe these mutations play a central role in many, if not most, human diseases. The small circular mtDNA has proven to be a Pandora's box of pathogenic mutations and rearrangements. Being extremely sensitive to environmental threats, mitochondria produce high-energy molecules – adenosine triphosphate (ATP). Mitochondria also generate reactive oxygen species (ROS), which participate in cell signaling and communication, particularly between nuclear and mitochondrial genes. Our main goal is to identify mitochondrial susceptibility genes for human complex diseases. The classical statistical techniques used to date to analyze the nuclear genome are not appropriate to directly be applied to the mitochondrial genome. Some adjustments and new methods need to be developed in the context of mapping mitochondrial polymorphisms. Using different genotyping platforms such as the Affymetrix 6.0 GeneChip array, Illumina MetaboChip 200K, Illumina Human Exome Beadchip array, and Affymetrix Axiom chip array we performed mitochondrial GWA analysus in the KORA population with several phenotypes: BMI, cholesterol, post-traumatic stress disorder, thyroid diseases, anxiety, depression, and asthma, among others. Our findings highlight the important role of the mtDNA among the factors that contribute to the risk of human complex diseases and suggest that variants in the mitochondrial genome may be more important than has previously been suspected.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Case-Control Studies, Causation, Psychiatric Diseases, Quantitative Trait Analysis*

C 7

A dramatic resurgence of the GIGO syndrome in the 21st century

Françoise Clerget-Darpoux[1], Emmanuelle Génin[2]
[1]IHU imagine- INSERM U781, Paris, France
[2]INSERM UMR1078, Brest, France

In the search of the genetic factors underlying multifactorial diseases, the way is paved by epidemics of GIGO (Garbage-In Garbage-Out) syndrome. A first outbreak took place in the late 1980's when geneticists building on the success of model-based linkage analysis in monogenic diseases started to use monogenic models to study multifactorial diseases. A second outbreak is ongoing, with genome-wide association study (GWAS) heritability estimates. Almost all GWAS on multifactorial diseases quantify the contribution of the identified genetic variants to disease susceptibility through heritability estimates. These estimates are compared to the ones obtained from familial disease segregation in order to determine how much of the heritability is missing and to prompt the search for other culprits such as rare variants. Heritability estimates are obtained under the additive polygenic model assuming that the genetic susceptibility in multifactorial disease is only explained by variants with moderate and additive effects. This simplistic model cannot be rejected based on the information provided by bi-allelic tag-SNPs, not because it is the true model, but because this information is extremely poor for modelling the effect of genetic risk factors. Several examples such as PTPN22 in rheumatoid arthritis clearly illustrate the fact that using the tag-SNP information alone may lead to a huge underestimation of the real effect and to an incorrect classification in terms of risk. GWAS has proven to be an efficient tool for susceptibility gene detection but not for their modelling. In this work, we show how heritability estimates could be biased when the disease model is misspecified.

Categories: *Association: Genome-wide, Heritability, Multifactorial Diseases, Prediction Modelling*

C 8

Large Scale Prediction and Dissection of Complex Traits

Hae Kyung Im[1], Eric R Gamazon[1], Keston Aquino-Michaels[1], Nancy J Cox[1]
[1]The University of Chicago

High accuracy prediction of disease susceptibility and drug response is necessary to make personalized or precision medicine a reality. Despite initial optimism at the completion of the human genome sequence, accurate predictive tests for many common conditions are still unavailable. The small portion of the total variability explained by genomewide significant genes have dampened the enthusiasm. However, studies of the total heritability explained by full set of genotyped variants show that there is ample room for improvement. Recent power calculations have showed that in order to achieve prediction R squares close to heritability estimates, we may need millions of individuals in our studies. However, given the rate of increase in sample sizes of large-scale meta-analysis studies (over a quarter Million for BMI), we are not too many years away from achieving these numbers. Also, advances in electronic medical records and scalable computing systems are allowing us to gather and handle these massive sample sizes. To take full advantage of the growing amount of information, we are building a publicly available catalog of prediction models --predictDB-- that hosts additive models for a range of phenotypes such as inflammation markers, disease risk, lipid traits, anthropomorphic traits, to name a few. Furthermore, we have built prediction models for gene expression levels in multiple tissues as well as microRNAs. In addition to prediction, we use these models to dissect the biology of complex traits. For example, we use the prediction models for gene expression to find genes that are differentially expressed in silico between cases and controls for a range of diseases. This is a novel gene based association test, termed PrediXcan, which directly tests the hypothesis that genetic variation alters disease risk through the regulation of gene expression levels. Application to the Wellcome Trust Case Control Consortium data yielded many genome-wide significant hits. Many of them are known disease genes but many are novel and replication efforts are under way.

Categories: *Association: Genome-wide, Gene Expression Patterns, Prediction Modelling*

C 9

Genetic predictors of longer telomeres are strongly associated with risk of melanoma

Jennifer H Barrett[1], David T Bishop[1], Nicholas K Hayward[2], Christopher I Amos[3], Paul DP Pharoah[4], Florence Demenais[5], Matthew H Law[2], Mark M Iles[1], The GenoMEL Consortium
[1]University of Leeds, UK
[2]QIMR Berghofer Medical Research Institute, Brisbane, Australia
[3]Dartmouth College, Hanover, USA
[4]University of Cambridge, UK
[5]INSERM, Paris, France

Telomeres protect the single-stranded chromosome ends from damage. Telomeres shorten with age and environmental exposures such as smoking. Telomere length (TL) has been related to a number of age-related diseases, usually through cross-sectional studies from which the direction of effect cannot be inferred. In contrast to most diseases, modest evidence has accumulated that longer TL is positively associated with the number of melanocytic nevi and with the risk of a few cancers, including melanoma. As more is discovered about the genetic basis of TL, Mendelian randomisation principles may be invoked to elucidate this. A recent genome-wide association study of TL identified 7 genome-wide significant SNPs[1]. Based on these SNPs and their estimated effect sizes a "telomere score" was created, and its relationship with melanoma risk was investigated using >11,000 cases and 13,000 controls. Four of the 7 SNPs showed nominal evidence of association with melanoma risk ($p<0.05$). There was strong evidence of an association between the score and melanoma risk ($p<10-8$); the estimated risk of melanoma to those with a telomere score in the highest quartile was almost 30% higher than to those with a score in the lowest quartile. Further analysis suggests that when the telomere score used here is refined, by using a denser imputation panel and by including more SNPs, the score is likely to be an even stronger predictor of melanoma risk. The genetic association suggests that, rather than reverse causation, the associations observed between TL and cancer risk are due either to a direct causal effect of longer telomeres or to the pleiotropic effect of a number of genes.

[1]Codd et al, Nat Genet 2013; 45:422-427

Categories: *Association: Unrelated Cases-Controls, Cancer, Mendelian Randomisation*

C 10

Detection of cis and trans eQTLs/mQTLs in purified primary immune cells

Silva Kasela[1,2], Liina Tserel[3], Tõnu Esko[4], Harm-Jan Westra[5], Lude Franke[5], Krista Fischer[4], Andres Metspalu[1,2], Pärt Peterson[3], Lili Milani[4]
[1]Estonian Genome Center, University of Tartu, Tartu, Estonia
[2]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia
[3]Institute of General and Molecular Pathology, University of Tartu, Tartu, Estonia
[4]Estonian Genome Center, University of Tartu, Tartu, Estonia
[5]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

A diverse repertoire of T cells is crucial for effective defense against infection with pathogens throughout life. CD4+ T cells are vital elements of the adaptive immune response, which have been associated with the pathogenesis of autoimmune and inflammatory diseases. CD8+ T cells are critically involved in defense against infections and can also contribute to the initiation and regulation of several organ-specific autoimmune diseases. In order to investigate the cell-type specific effects of nearby SNPs on gene expression (cis eQTLs) and DNA methylation (cis mQTLs), we purified CD4+ and CD8+ cells from the peripheral blood of over 600 healthy individuals. We determined the SNP genotypes (700K), expression levels of 47,000 transcripts from 300 subjects and methylation levels of 450,000 CpG sites from the 50 youngest and 50 oldest subjects. In total, we detected more cis eQTLs and mQTLs in CD4+ compared to CD8+ cells with a large overlap between the cell populations. Further, we selected a set of 9648 SNPs which have been associated with immune system related diseases from studies using the ImmunoChip and SNPs from the reports in the GWAS catalog. Despite the several fold smaller sample size, we were able to identify recently reported trans-acting expression master regulator SNPs on chromosome 12 and 16 (Fairfax et al. 2012, Westra et al. 2013). Moreover, our study revealed that some of the eQTLs identified in whole blood originate from CD4+ cells only, and we also identified downstream regulated genes that could not be detected in whole blood. For example, we found three SNPs associated with type 1 diabetes, Crohn's disease, and inflammatory bowel's disease to affect the expression of the STAT1 and IRF1 genes in trans in CD4+ cells.

Categories: *Epigenetic Data, Epigenetics, Gene Expression Patterns, Genomic Variation, Quantitative Trait Analysis*

C 11

Why Next-Generation Sequencing Studies May Fail: Challenges and Solutions for Gene Identification in the Presence of Familial Locus Heterogeneity

Suzanne M Leal[1], Regie Lyn P Santos-Cortez[1], Atteeq U Rehman[2], Meghan C Drummond[2], Saima Riazuddin[3], Deborah A Nickerson[4], Wasim Ahmad[5], Sheikh Riazuddin[6], Thomas B Friedman[2], Ellen S Wilch[7]

[1]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA

[2]Laboratory of Molecular Genetics, National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Rockville, Maryland 20850, USA

[3]Laboratory of Molecular Genetics, Division of Pediatric Otolaryngology Head and Neck Surgery, Cincinnati Childrens Hospital Medical Center, Cincinnati 45229, Ohio, USA

[4]University of Washington Center for Mendelian Genomics

[5]Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan

[6]National Center of Excellence in Molecular Biology, University of the Punjab, Lahore 54590, Pakistan

[7]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA

Next-generation sequencing (NGS) of exomes and genomes has accelerated the identification of genes involved in Mendelian phenotypes. However, many NGS studies fail to identify causal variants. An important reason for such failures is familial locus heterogeneity, where causal variants in two or more genes within a single pedigree underlie Mendelian trait etiology. As examples of intra- and inter-sibship familial locus heterogeneity, we present 10 consanguineous Pakistani families segregating hearing impairment (HI) due to homozygous mutations in two different HI genes and a large European-American pedigree in which HI is caused by pathogenic variants in three different genes. We have identified 41 additional pedigrees with syndromic and nonsyndromic HI for which a single known HI gene has been identified but only segregates with the phenotype in a subset of affected pedigree members. We estimate that locus heterogeneity occurs in 15.3% (95% confidence interval 11.9 to 19.9%) of the families in our collection where we have identified at least one variant in a previously published HI gene which only segregates with HI phenotype in a subset of affected pedigree members. We demonstrate novel approaches to apply linkage analysis and homozygosity mapping which can be used to detect locus heterogeneity using either NGS or SNP array data. Results from the analysis can also be used to group sibships or individuals most likely to be segregating the same causal variants and thereby aid in gene identification. The results can be used to aid in the selection of pedigree members for NGS. It is demonstrated how these methods can increase the success rate of gene identification for families with locus heterogeneity.

Categories: *Association: Family-based, Heterogeneity, Homogeneity, Linkage Analysis, Sequencing Data*

C 12

Variation in estimates of kinship observed between whole-genome and exome sequence data

Elizabeth E Blue[1]
[1]University of Washington

Genotypic variation may be used to estimate relationships between individuals. These relationships are clearly important when confirming pedigree structure and testing the co-segregation of a variant with a trait. It is also important when testing association of a genetic variant with case/control status in a set of "unrelated" subjects: ex., the reason why principal components are included as covariates to minimize the effects of population stratification. The popularity of exome sequencing for disease gene discovery suggests we need to know whether these data provide accurate estimates of relationships between subjects. The exome represents ~1% of the genome, and does not represent a random subsample of genomic variation. Here, we compare a method-of-moments and the KING-robust estimator of kinship applied to SNPchip data, whole exome, and whole genome sequence data for four subjects with known pedigree relationships. SNPchip-based estimates of kinship are similar to the pedigree-based expectation, with the KING-robust estimates deviating slightly more than the method-of-moments estimates. However, the exome-based estimates are much more variable: overestimating some relationships by as much as a third and underestimating others by nearly a quarter of the pedigree-based expectation. We explore the effects of allele frequency, linkage disequilibrium, and the number of markers on estimates of kinship drawn from whole genome sequence data. These results suggest we must account for the non-random distribution of variation in the exome when estimating relationships between subjects.

Categories: *Ascertainment, Genomic Variation, Linkage and Association, Population Stratification, Sequencing Data*

C 13

Robust genotype calling from very low depth whole genome sequencing data

Arthur L Gilly[1], Jeremy Schwartzentruber[1], Angela Matchan[1], Aliki-Eleni Farmaki[2], George Dedoussis[2], Petr Danecek[1], Lorraine Southam[1,3], Eleftheria Zeggini[1]
[1]Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK
[2]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens, Greece
[3]Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK

Low-depth whole-genome sequencing (WGS) has been proposed as a powerful approach to complex trait association study design, as it allows for a reduced per-sample cost and hence greater sample size. However variants and genotypes called at these depths tend to be less reliable than chip-typed ones, and robust guidelines for variant filtering, genotype refinement and imputation have not been established yet. In this study, we focus on 995 samples from a Greek isolated population (HELIC study), sequenced at very low depth (1x). We used GWAS and exome chip data (available for all samples) and high-depth exome sequencing (for a subset of samples) as truth sets for performance calculations. We find that the Variant Quality Score Recalibration tool typically used to filter low-confidence sites can react unpredictably to small changes in the underlying model's parameters for low depth WGS data calling. We show that these pitfalls can be avoided with a comprehensive exploration of the parameter space. We demonstrate that over 80% of true low-frequency (1%<MAF<5%) variants are found, compared to an average 60% for 0.1%<MAF<1% and 40% for MAF<0.1%. We perform extensive benchmarking of the BEAGLE, IMPUTE2 and MVNCall refinement tools and show that with the help of the 1000 Genomes reference panel, it is possible to reach a >95% genotype concordance and a >90% minor allele concordance across the whole MAF spectrum. We replicate known association hits, thereby providing a proof of concept for a robust processing pipeline for low-depth WGS variant calls.

Categories: *Association: Genome-wide, Bioinformatics, Data Mining, Data Quality, Sequencing Data*

C 14

Insights into the genetic architecture of anthropometric traits using whole genome sequence data

Ioanna Tachmazidou[1], Graham RS Ritchie[1,2], Josine Min[3], Klaudia Walter[1], Jie Huang[1], John Perry[4], Thomas Keane[1], Shane McCarthy[1], Yasin Memari[1]
[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom
[3]MRC Integrative Epidemiology Unit, University of Bristol
[4]MRC Epidemiology Unit, University of Cambridge

Body weight and fat distribution measures are associated with increased risk of cardiometabolic disease. As part of the UK10K study, we investigated the genetic architecture of 12 anthropometric traits in 3,538 individuals with ~7x whole genome sequence (WGS) data from the ALSPAC and TwinsUK cohorts. Variants discovered through WGS, along with those from the 1000 Genomes Project, were imputed into additional individuals from the ALSPAC and TwinsUK cohorts with GWAS data, increasing the total sample size to 11,178. We investigated association between anthropometric traits and ~9 million variants with MAF≥0.01 and ~5 million variants with MAF 0.001-0.01. In silico replication was sought in 16 external cohorts for a total sample size of 15,000-40,000 depending on trait. We observe a significant excess of independent previously not reported variants with MAF>0.01 and p<10-5 in UK10K in all anthropometric traits. We find significant enrichment of variants associated with BMI in UK10K and established monogenic obesity genes. Further replication is ongoing, but interim analyses identify replicating signals, for example, variant chr5:105105444 (EAF 0.0084; UK10K p=4.69x10-5; replication p=2.53x10-4; overall p=5.53x10-8, sample size=27,687) is a novel signal associated with waist circumference adjusted for BMI. Waist to hip ratio is associated with variant chr9:23016057 (EAF 0.003; UK10K p=6.11x10-5; replication p=2.92x10-04; overall p=5.98x10-8, sample size=25,373). These replicating signals are at variants with MAF<0.01, have modest effect sizes and are not present in HapMap. Larger sample sizes are required for the identification and replication of further rare variant associations with anthropometric traits.

Categories: *Association: Genome-wide, Quantitative Trait Analysis, Sequencing Data*

C 15

Standard Imputation versus Generalizations of the Basic Coalescent to Estimate Genotypes

Maria Kabisch[1], Ute Hamann[1], Justo Lorenzo Bermejo[2]
[1]Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

Genotypes that have not been directly measured are often imputed in association studies to increase statistical power, to refine association mapping, and to detect genotyping errors. Most often applied imputation methods exploit the present linkage disequilibrium (LD) among genetic variants to infer genotypes. This causes a strong dependence of imputation accuracy on the similarity of LD patterns in the study population and the reference panel. Alternatively, coalescent theory assumes that haplotypes are related through the underlying population genealogy. Coalescent-based imputation relaxes the assumption of identical LD patterns and may thus result in an increased accuracy. To examine this hypothesis, we first assessed the imputation accuracy under the basic coalescent. Study and reference haplotypes were simulated using 'msms'[1]. Haplotypes were paired at random to mimic biallelic variants. Ten percent of the variants in the study were randomly selected and assumed to be directly measured, the rest was masked. 'BATWING' was used to estimate one thousand genealogical trees, which were subsequently summarized in a consensus three with 'SumTree'[2,3]. Expected coalescence times were used to identify haplotype templates for genotype imputation. Finally, masked genotypes were imputed and compared with the true genotypes to quantify the accuracy of imputation. After examining genotype imputation under the basic coalescent, population growth and population structure were incorporated. Imputation accuracies reached by standard methods, e.g. IMPUTE2, will be compared with coalescent-based results at the IGES 2014 conference.

[1] Ewing, Hermisson (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics, 26: 2064-5.
[2] Wilson, Weale, Balding (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. Journal of the Royal Statistical Society, Series A, 166: 155-88.
[3] Sukumaran, Holder (2010) DendroPy: A Python library for phylogenetic computing. Bioinformatics 26: 1569-1571.

Categories: *Coalescent Theory*

C 16

Improvement of genotype imputation accuracy through integration of sequence data from a subset of the study population

Barbara Peil[1], Maria Kabisch[2], Christine Fischer[3], Ute Hamann[2], Justo Lorenzo Bermejo[1]
[1]Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany
[2]Molecular Genetics of Breast Cancer, German Cancer Research Center (DFKZ), Heidelberg, Germany
[3]Institute of Human Genetics, University Hospital Heidelberg, Heidelberg, Germany

Unmeasured genotypes in genetic association studies can be estimated (imputed) using external data repositories, for example the HapMap, ideally complemented with sequence data from own study individuals. Several studies have evaluated which individuals are most helpful for genotype imputation. Initial efforts focused on a selection of reference individuals who best reflected recombination patterns in the study population. More recently, the advantage of genetic diversity in the reference panel has been recognized. We have compared different strategies to select study individuals for sequencing in order to maximize imputation accuracy. Five alternative strategies were examined in HapMap based simulations. The strategy “none” incorporated no additional sequence to the external reference panel. The strategy “random” incorporated the sequences of a random subset of 10% study individuals. The strategies “univariate depth”, “bivariate depth” and “trivariate depth” relied on a genomewide principal component analysis of the study population, followed by the identification of 10% of study individuals with the largest statistical depth based on the first one, first two and first three principal components. As expected, the inclusion of additional sequences from the own study population outperformed imputation exclusively relying on external reference panels. The selection of study individuals based on the univariate depth was the best strategy in simulations mimicking European association studies. Detailed results for additional investigated scenarios will be provided at the conference.

Categories: *Association: Genome-wide, Data Quality, Missing Data, Sequencing Data*

C 17

Learning Genetic Architecture of Complex Traits Across Populations

Marc Coram[1], Sophie I Candille[1], Hua Tang[1]
[1]Stanford University

Genome-wide association studies (GWAS) have successfully revealed many loci that influence complex traits and disease susceptibilities. An unanswered question is "to what extent does the genetic architecture underlying a trait overlap between human populations?" We explore this question using blood lipid concentrations as a model trait. We demonstrate striking similarities in genetic architecture of lipid traits across human populations. In particular, we found that a disproportionate fraction of lipid variation in African Americans and Hispanic Americans can be attributed to genomic loci exhibiting statistical evidence of association in Europeans, even though the precise genes and variants remain unknown. At the same time, we found substantial allelic heterogeneity within shared loci, characterized both by population-specific rare variants and variants shared among multiple populations that occur at disparate frequencies. Exploiting this overlapping genetic architecture, we develop a population-sensitive approach that substantially improves the efficiency of GWAS in non-European populations.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Population Stratification*

C 18

Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry

Daniel Shriner[1], Fasil Tekola-Ayele[1], Adebowale Adeyemo[1], Charles N Rotimi[1]
[1]National Human Genome Research Institute

We investigated ancestry of 3,528 modern humans from 163 ethno-linguistic groups. We identified 19 ancestral components, with 94.4% of individuals showing mixed ancestry. After using whole genome sequences to correct for ascertainment biases in genome-wide genotype data, we dated the most recent common ancestor to 140,000 years ago. We detected an Out-of-Africa migration 100,000–87,000 years ago, leading to peoples of the Americas, east and north Asia, and Oceania, followed by another migration 61,000–44,000 years ago, leading to peoples of the Caucasus, Europe, the Middle East, and south Asia. We dated eight divergence events to 33,000–20,000 years ago, coincident with the Last Glacial Maximum. We refined understanding of the ancestry of several ethno-linguistic groups, including African Americans, Ethiopians, the Kalash, Latin Americans, Mozabites, Pygmies, and Uygurs, as well as the CEU sample. Ubiquity of mixed ancestry emphasizes the importance of accounting for ancestry in history, forensics, and health.

Categories: *Ascertainment, Population Genetics*

C 19

Model Comparison and Selection for Count Data with Excess Zeros in Microbiome Studies

Wei Xu[1,2], Andrew D Paterson[2,3], Williams Turpin[4], Kenneth Croitoru[4], Lizhen Xu[3]
[1]Department of Biostatistics, Princess Margaret Hospital, Toronto, ON, Canada
[2]Program in Genetics and Genome Biology, the Hospital for Sick Children, Toronto, ON, Canada
[3]Dalla Lana School of Public Health, University of Toronto, Toronto,ON, Canada
[4]Division of Gastroenterology, Zane Cohen Centre for Digestive Diseases, Mount Sinai Hospital, Toronto, ON, M5T 3L9, Canada

In human microbiome studies, it is often of interest to identify clinical or genetic factors that are associated with different bacterial taxa. The microbiota sequence count data are complex with features such as high dimension, over-dispersion, and often excess zeros. In addition, the number of total reads varies among subjects. Zero inflated or hurdle models provide possible analytic approaches for this type of data and the variation in total reads can be adjusted as offsets. However, in practice, one part models which ignore zero inflation are often used. To determine the pattern of superiority of using zero inflated or hurdle models over the simplified one part models, we designed extensive simulation studies to compare the performance of different statistical methods under a variety of generating scenarios. These scenarios include: different levels of zero inflation; presence of dispersion; different magnitude and directions of the covariate effect on both the structural zero and count components. The results show that, compared to one-part models, the hurdle and zero inflated models have well controlled type I errors, higher power, better goodness of fit measures, and are more accurate and efficient in the parameter estimation. Besides that, the hurdle models have similar goodness of fit and parameter estimation for the count component as their corresponding zero inflated models. However, the estimation and interpretation for the parameters for the zero components can be different. In addition, we developed a comprehensive model selection and analysis strategy to analyze this type of data. This strategy was implemented in a gut microbiome study of >400 independent subjects.

Categories: *Microbiome Data, Quantitative Trait Analysis*

C 20

Bayesian Latent Variable Models for Hierarchical Clustered Taxa Counts in Microbiome Family Studies with Repeated Measures

Lizhen Xu[1], Andrew D Paterson[1,2], Wei Xu[2,3]
[1]Dalla Lana School of Public Health, University of Toronto, Toronto,Canada
[2]Program in Genetics and Genome Biology, the Hospital for Sick Children, Toronto, ON, Canada
[3]Department of Biostatistics, Princess Margaret Hospital, Toronto, ON, Canada

In microbiome studies, taxa count data are often over-dispersed and include excess zeros. Furthermore, different taxa belonging to the same taxonomic hierarchical cluster are often correlated due to their similar 16S rRNA sequences. Added characteristics of microbiome data is the repeated measures on related family members. Joint modeling of multiple taxa using family data with repeated measures is desirable but non-trivial due to the complex correlation and multi-dimensional outcome data. To overcome these challenges, we propose to use the latent variable (LV) methodology. The LV approach links the multiple taxa counts by introducing a latent random variable that represents the unobserved trait of their common taxonomy cluster. The latent variable formulation also provides a flexible way to allow for outcomes with discrete components, in our case, the negative binomial outcomes with or without zero inflation. LV also provides an effective way to detect pleiotropic genes, with effects on multiple taxa. We build our LV inference in a Bayesian framework. Samplings from the posterior distribution are obtained using MCMC algorithms. The parameter expansion technique is used to improve the mixing of chains and the Bayesian deviance information criteria (DIC) and Bayes factors are used for model selection. Extensive simulations show that our method performs well in capturing the correlations among the multiple taxa induced by shared host genetic factors. We then illustrate our method with a gut microbiome study of lean and obese twins.

Categories: *Bayesian Analysis, Markov Chain Monte Carlo Methods, Microbiome Data, Multivariate Phenotypes, Quantitative Trait Analysis*

C 21

A Retrospective Likelihood Approach for Efficient Integration of Multiple Omics and Non-Omics Factors in Case – Control Association Studies of Complex Diseases

Brunilda Balliu[1], Roula Tsonaka[1], Stefan Boehringer[1], Jeanine Houwing - Duistermaat[1]
[1]Leiden University Medical Center, The Netherlands

Integrative omics, the joint analysis of outcome and multiple types of omics data, such as gen-omic, epigen-omic and transcript-omic data, offers a promising alternative to genome-wide association studies, for more powerful and biologically relevant association studies [1,2]. These studies usually employ the case-control design, and they often include data on additional non-omic covariates, e.g. age or gender, that may modify the underlying omics risk of cases or controls. An unanswered question is how to best integrate multiple omics, and possibly non-omics information to maximize statistical power in studies that ascertain individuals on the basis of phenotype. Most publications on integrative omics have relied on some variant of the prospective logistic regression to model the association between outcome and risk factors [2]. However, while such an approach has improved power in studies with random ascertainment, relative to methods that analyze each data source separately, it often loses power under case-control ascertainment [3]. In this article, we propose a novel statistical method for integrating multiple omics, and possibly non-omics factors, in case-control association studies. Our method is based on a retrospective likelihood function that properly reflects the case-control sampling, by modeling the joint distribution of the omics and non-omics factors conditional on the case-control status. When possible, we explicitly impose the independence assumption between the omics and non-omics covariates. The new method provides accurate control of false-positive rates while maximizing statistical power. The method is illustrated using simulated and real data examples.

[1] H Li (2013), WIRE:SBM, 5(6): 677–686.
[2] Huang Y et al. (2014), Ann. Appl. Stat. , 8(1):352-376.
[3] Zaitlen et al. (2012), PLoS Genet 8(11): e1003032.

Categories: *Ascertainment, Case-Control Studies, Data Integration, Epigenetic Data, Epigenetics, Gene Expression Arrays, Maximum Likelihood Methods*

C 22

Inference for high-dimensional feature selection in genetic studies

Claus T Ekstrøm[1]
[1]Biostatistics, University of Copenhagen

Feature selection is a necessary step in many genetic applications because the biotechnological platforms provide a cheap and fast means for producing high-dimensional data. This need for dimension reduction is heightened further for example when data from different omics are combined into simultaneous integrated data analysis or when higher-level interactions among the available predictors are considered (which is the case for gene-gene or gene-environment interactions or in epigenetics). Penalized regression models such as the Lasso or the elastic net have proved useful for variable selection in many genetic applications - especially for situations with high-dimensional data where the numbers of predictors far exceeds the number of observations. These methods identify and rank variables of importance but do not generally provide any inference of the selected variables. Thus, the variables selected might be the most ''important'' but need not be significant. We propose a significance test for evaluating the number of significant selection(s) found by the Lasso. This method rephrazes the null hypothesis and uses a randomization approach which ensures that the error rate is controlled even for small samples. The ability of the algorithm to compute p-values of the expected magnitude is demonstrated with simulated data and the algorithm is applied to two dataset: one on prostate cancer and a full GWAS. The proposed method is found to provide a powerful way to evaluate the set of selections found by penalized regression when the number of predictors are several orders of magnitude larger than the number of observations.

Categories: *Association: Family-based, Association: Genome-wide, Bioinformatics, Gene - Environment Interaction, Gene - Gene Interaction*

Posters

P 1

Increased power for detection of parent-of-origin (imprinting) effects in genome-wide association studies using haplotype estimation

Richard Howey[1], Heather J Cordell[1]
[1]Newcastle University, UK

In genetic studies, parent-of-origin (imprinting) effects can be considered as the phenomenon whereby an individual's phenotype depends both on their own genotype and on the parental origin of the constituent alleles. Several methods have been proposed to detect such effects in the context of studies of case/parent trios with single nucleotide polymorphism (SNP) genotype data. For most case/parent trios, the genotype combinations are such that the parent-of-origin of the alleles in the child can be determined unambiguously, but this is not true when all three individuals are heterogenous at a single SNP under study. Existing methods for the detection of parent-of-origin effects in the context of genome-wide association studies (GWAS) thus either perform some sort of "averaging" over possible configurations or else discard these ambiguous case/parent trios. The power to detect parent-of-origin effects would be increased if the true parental origin of the alleles could be determined with a higher degree of certainty. We present here an extension to the GWAS method implemented in the PREMIM/EMIM software to detect parent-of-origin effects using external estimates of haplotypes provided by the program SHAPEIT2, thereby using surrounding SNP information to help better estimate the parental origin of alleles at a given test SNP. We show through simulations that our approach has increased power over previous versions of EMIM and achieves power near to that achieved if the parent-of-origin of alleles were known.

P 2

Epidemiological Profile of Cleft Palate in the State of Bahia-Brazil

Marcela MQL Leiro[1], Renata LLF de Lima[1], Luzia Poliana dos Anjos Silva [1]
[1]University Federal of Bahia, Brazil

Cleft lip and palate (FLPs) are a set of malformations of the face representing the most common congenital anomalies of the human species . Brazilian data on craniofacial anomalies are still considered scarce and scattered, due to the difficulty of reporting these cases in the public health system. Given the different population , environmental, social , lifestyle and issues of racial miscegenation in Brazil characteristics , the prevalence of this anomaly seems to vary in each state of the country . OBJECTIVE : To describe the epidemiology of patients with Cleft Lip and / or Palate residents of the State of Bahia . Study of quantitative trait runs through cross-sectional case series with sample group consisted of children aged 0-12 years, who are part of a program of care in Centrinho - BA . RESULTS : Of the 206 patients there was a slight prevalence of females (51 %) , and non - syndromic cases , 95 % . Of the total sample , 53 % had CLP and only 19 % FL , and 119 cases (58 %) were born in the state . The FLP was more prevalent in patients with a positive family history , 71 cases (34.5 %) . Regarding the etiology of PLF 9.3% (19 cases) reported having used alcohol during pregnancy . It was noted socioeconomic situation of vulnerability in patients with CLP where 60 % (124 cases) had an income of 1-3 minimum wages . CONCLUSION : It was observed through this study, a higher incidence of CLP in relation to FL associated with a higher prevalence in blacks , with the socioeconomic vulnerability exposed population.

P 3

Generalized Functional Linear Models for Gene-based Case-Control Association Studies

Ruzong Fan[1], Yifan Wang[1], James L Mills[1], Tonia C Carter[2], Iryna Lobach[3], Alexander F Wilson[4], Joan E Bailey-Wilson[4], Daniel E Weeks[5], Momiao Xiong[6]
[1]National Institute of Child Health and Human Development, National Institutes of Health
[2]Marshfield Clinic
[3]University of California, San Francisco
[4]National Human Genome Research Institute, National Institutes of Health
[5]University of Pittsburgh
[6]University of Texas - Houston

By using functional data analysis techniques, we developed generalized functional linear models for testing association between a dichotomous trait and multiple genetic variants in a genetic region while adjusting for covariates. Both fixed and mixed effect models are proposed and compared. Extensive simulations show that Rao's efficient score tests of the proposed fixed effect models are very conservative since they generate low type I errors, and global tests of the mixed effect models are very robust since they generate accurate type I errors. Furthermore, we found that the Rao's efficient score test statistics of the proposed fixed effect models have higher power than the sequence kernel association test (SKAT) and its optimal unified version (SKAT-O) in most cases when the causal variants are both rare and common. When the causal variants are all rare (i.e., minor allele frequencies less than 0.03), the Rao's efficient score test statistics and the global score tests have similar or slightly lower power than SKAT and SKAT-O. In practice, it is not known whether rare variants or common variants in a gene are disease-related. All we can assume is that a combination of rare and common variants influences disease susceptibility. Thus, the superior performance of the proposed models when the causal variants are both rare and common shows that the proposed models can be very useful in dissecting complex traits. SNP data related to neural tube defects and Hirschsprung's disease are analyzed by the proposed methods and SKAT and SKAT-O for a real application and comparison. The methods can be used in either gene-disease genome-wide/exome-wide association studies or candidate gene analyses.

Categories: *Association: Candidate Genes, Association: Genome-wide, Association: Unrelated Cases-Controls, Case-Control Studies, Linkage and Association, Multilocus Analysis, Multiple Marker Disequilibrium Analysis, Sequencing Data*

P 4

Genetic analysis of the chromosome 15q25.1 region identifies IREB2 variants associated with lung cancer

Christopher I Amos[1], Ivan P Gorlov[1], James D McKay[2], Loïc LeMarchand[3], Yafang Li[1], Gianluca Severi[4], David C Christiani[5], Paul Brennan[2], John K Field[6], Rayjean J Hung[7]
[1]Dartmouth College
[2]International Agency for Research on Cancer
[3]University of Hawaii
[4]Human Genetics Foundation, Torino, Italy and University of Melbourne
[5]Harvard University School of Public Health
[6]University of Liverpool
[7]University of Toronto

Genome-wide association studies of lung cancer identified the region of chromosome 15q25.1 that includes a nicotinic acetylcholine receptor cluster as being the most strongly associated with lung cancer risk. To characterize the impact that specific functional variants in this region have upon risk for lung cancer development we performed fine mapping selecting all currently known SNPs influencing lung cancer risk along with coding SNPs in the 200 megabase region surrounding CHRNA5, a gene known to influence smoking behavior in this region. Markers used in analysis were selected based upon the following criteria: known functional effect on activity, validation in African or European populations, position across the region, predicted effect on function, r-square with other markers less than 80%. We fine mapped the region by genotyping 1395 SNPs extending from the gene CRABP1 to ADAMTS7 from position 79103132 to position 79103132 using a custom Affymetrix Axiom array in 3063 cases and 2940 controls of European ancestry from 5 studies: MSH-PMH, EPIC, MEC, LLPC, HPFS & NHS. Odds ratios (OR) adjusted for age, sex, the first two principal components and population were estimated using logistic regression. Across this region, 101 SNPs met the multiple testing corrected threshold ($p<3.5\times10\text{-}5$). The most significant SNPs lie in a region of IREB2 with the most significantly associated variant being rs17483686 (OR=1.26, p=8.93 x 10-12). The previously well characterized SNP in CHRNA5, rs16969968, which causes reduced signaling, yielded a less significant association (OR=1.24, p=8 x 10-10). These findings suggest IREB2, a gene related to iron metabolism, plays a role in lung cancer development in addition to nearby nicotinic receptors.

Categories: *Association: Candidate Genes, Cancer, Fine Mapping, Gene - Environment Interaction*

P 5

A novel integrated framework for large scale omics association analysis

Ramouna Fouladi[1,2], Kyrylo Bessonov[1,2], Francois Van Lishout[1,2], Jason H Moore[3], Kristel Van Steen[1,2]
[1]Systems and Modeling unit, Montefiore Institute, University of Liege, Liege, Belgium
[2]Bioinformatics and Modeling, GIGA-R, University of Liege, Liege, Belgium
[3]Department of Genetics, Institute for Quantitative Biomedical sciences, Geisel School of Medicine at Dartmouth college, Lebanon, US

Genome-wide association studies (GWA studies) have been very successful in identifying numerous genetic loci associated with a wide range of complex traits. These discoveries have revealed new pathways that seem to play a significant role in common diseases. Single omics studies, such as GWAs, only provide limited information to disease-related biological or functional mechanisms. In an omics – disease trait association setting, ideally, a generic tool is created that can deal with different granularities of omics information (i.e., different architectures of common and rare variants, epigenetic markers, gene expression). Here, a novel omics association analysis technique is proposed that builds upon the Model-Based Multifactor Dimensionality Reduction (MB-MDR) framework. At the basis of the method lies a data organization step that involves clustering of individuals. In the first implementations of MB-MDR, these features were SNPs, and individuals were clustered according to their genotypes. In genomic MB-MDR, any feature (continuous or categorical) can be analyzed, and features mapped to genomic "regions of interest" (ROIs) are submitted to a clustering algorithm to find groups of similar individuals on the basis of selected ROIs. When applied to exome-sequencing data, we can identify a gene as a ROI, and can take both rare and common features mapped to these regions as input features. We then propose to cluster individuals according to their similarities based on rare and common variants, after which classic MB-MDR is applied. The performance of several feature selection methods, similarity measures, and clustering algorithms in genomic MB-MDR is investigated using synthetic and real-life exome sequencing data.

Categories: *Association: Candidate Genes, Bioinformatics, Data Integration, Epigenetics*

P 6

Inclusive Composite Interval Mapping and Skew-Normal Distribution

Elisabete Fernandes[1]
[1]CEMAT-Center for Computacional and Stochastic Mathematics, Portugal

The composite interval mapping, CIM, (Jansen and Stam, 1994; Zeng, 1994)is the most commonly used method for QTL mapping with populations derived from biparental crosses. However, the CIM may not completely ensure all its advantageous properties. The modified algorithm, called as inclusive composite interval mapping, ICIM, (Wang et al., 2007) has a simpler form than that used in CIM, but a faster convergence speed. ICIM retains all advantages of CIM over IM and avoids the possible increase of sampling variance and the complicated background marker selection process in CIM. This approach makes use of the assumption that the quantitative phenotype follows a normal distribution (Kruglyak and Lander, 1995). Many phenotypes of interest, however, follow a highly skewed distribution, and in these cases the false detection of a major locus effect may occur (Morton, 1984). An interesting alternative is to consider a skew-normal mixture model in ICIM, and the resultingmethod is here denoted as skew-normal ICIM. This method, which is similar to ICIM, assumes that the quantitative phenotype follows a skew-normal distribution for each QTL genotype. The maximum likelihood estimates of parameters of the skew-normal distribution are obtained by the expectation-maximization (EM) algorithm. The proposed model is illustrated with real data from an intercross experiment that shows a significant departure from the normality assumption. The performance of the skew-normal ICIM is assessed via stochastic simulation. The results indicate that the skew-normal ICIM has higher power for QTL detection and better precision of QTL location as compared to ICIM.

Categories: *Association: Candidate Genes, Association: Genome-wide, Fine Mapping, Maximum Likelihood Methods, Quantitative Trait Analysis*

P 7

Transmission-based Tests For Genetic Association Using Sibship Data

Hemant Kulkarni[1], Saurabh Ghosh[1]
[1]Indian Statistical Institute, Kolkata

The classical Transmission Disequilibrium Test (TDT) for binary traits (Spielman et al. 1993) is a family-based alternative to population based case-control studies and is protected against population stratification, and hence, an association finding can be attributed to the presence of linkage. There have also been some extensions of the classical TDT for quantitative traits. However, these tests, which are based on the trio design (two parents and an offspring) do not remain valid as tests for association in the presence of sibship data since the marginal effect of linkage can result in transmission bias of alleles. In our study, we have modified the TDT test procedure for both binary as well as quantitative traits based on sibship data using a permutation based approach. We select one offspring at random from each family and compute the usual trio-based test statistic. We repeat this procedure and consider two test statistics based on the mean and the maximum value of the trio-based test statistics obtained over different replications. We obtain the exact distribution of the test statistic using permutations. We perform extensive simulations to evaluate the powers of the proposed tests under a wide spectrum of genetic models and different distributions of a quantitative trait. We find that the test statistic based on the mean yields more power compared to that based on the maximum.

Categories: *Association: Candidate Genes*

P 8

Identification of rare causal variants in sequence-based studies

Marinela Capanu[1], Iuliana Ionita-Laza[2]
[1]Memorial Sloan Kettering Cancer Center
[2]Columbia University

Pinpointing the small number of causal variants among the abundant naturally occurring genetic variation is a difficult challenge, but a crucial one for understanding precise molecular mechanisms of disease and follow-up functional studies. We propose and investigate two complementary statistical approaches for identification of rare causal variants in sequencing studies: a backward elimination procedure based on groupwise association tests, and a hierarchical approach that can integrate sequencing data with diverse functional and evolutionary annotations for individual variants. Using simulations, we show that incorporation of multiple bioinformatic predictors of deleteriousness, such as PolyPhen-2, SIFT and GERP++ scores, can improve the power to discover truly causal variants. As proof of principle, we apply the proposed methods to VPS13B, a gene mutated in the rare neurodevelopmental disorder called Cohen syndrome, and recently reported with recessive variants in autism. We identify a small set of promising candidates for causal variants, including a rare, homozygous probably-damaging variant that could contribute to autism risk.

Categories: *Association: Candidate Genes, Case-Control Studies, Genomic Variation, Sequencing Data*

P 9

Targeted resequencing of GWAS loci: insight into genetic etiology of cleft lip and palate through analysis of rare variants with focus on the 8q24 region

Margaret A Taub[1], Elizabeth J Leslie[2], The CleftSeq Consortium
[1]Johns Hopkins University
[2]University of Pittsburgh

Non-syndromic cleft lip with or without cleft palate (CL/P) is a common birth defect with complex inheritance. Despite considerable progress in identifying risk loci in several genome-wide association studies (GWAS), identification of the causal variants at each locus remains a challenge. To this end, we selected thirteen regions from earlier GWAS and candidate gene studies, totaling 6.3Mb, for targeted capture and deep sequencing in 1521 case-parent trios with CL/P from several populations. We performed statistical analyses on common, de novo and rare variants. Here, we focus on the latter, in particular in the 8q24 region. While many rare variant tests focus on coding variants, 8q24, as a gene desert, requires other approaches. We performed regulatory-region based burden tests to see if rare variants in a particular regulatory element were over- or under-transmitted. No results were significant after multiple testing correction. We used the likelihood-ratio based Scan-Trio method to find windows with over- or under-transmitted rare variants, restricting our analyses to variants with CADD score >10 and assessing significance by permuting transmitted and untransmitted haplotypes. This analysis revealed a promising cluster of variants near the GWAS hit in 8q24. We also did haplotype-based testing where haplotypes were grouped by allele carried at rs72728755, the SNP giving most significant signal in the transmission-disequilibrium test (TDT). We tested for differences in the presence of rare variants between deleterious and protective haplotypes by searching sliding windows for clusters of rare variants seen only on transmitted haplotypes. Significance was evaluated by permutation.Grants: U01-HG005925; R01-DE016148.

Categories: *Association: Candidate Genes, Association: Family-based, Haplotype Analysis, Linkage and Association, Sequencing Data*

P 10

A joint association model of effects of rare versus common variants on Age-related Macular Degeneration (AMD) using a Bayesian hierarchical generalized linear model

Wilmar M Igl[1], for the International AMD Genomics Consortium (IAMDGC)
[1]Department of Genetic Epidemiology, University of Regensburg, Germany

Purpose: AMD is a common cause of blindness in older people with a strong genetic contribution from common variants (CVs). Recently, several rare variants (RVs, MAF < 1%) were found. So far the contribution of RVs and CVs has not been examined in a comprehensive joint model. Methods. The IAMDGC data comprise 33,976 unrelated Europeans (16,144 Advanced AMD cases, 17,832 controls). 569,645 variants across the genome were genotyped on a custom-modified HumanCoreExome array by Illumina. The analyses focus on 18 known and 17 novel loci from single-variant analyses. The applied Bayesian hierarchical generalized linear model (here: logistic, Yi and Zhi, 2011) extends the generalized linear model framework by jointly estimating individual variant and group variant (here: rare vs. common) effects based on genetic risk scores. Weakly informative Bayesian priors (Hierarchical Cauchy) were used. All results were adjusted for ancestry principal components and DNA source as covariates and for multiple testing per locus. Results. The analysis of 225 rare versus 199 common variants (total 424, α=1E-4) in the CFI locus, showed independent group-level effects of rare (OR=2.06, CI95%=[1.92;2.22], p=1.09E-87) and common (OR=2.39, CI95%= [2.18,2.62], p=7.38E-77) variants. Significant single variant effects were only observed for the known rare variant rs141853578 (G119R, OR=3.24, CI95%=[2.06;5.10], p=3.34E-07) in this joint model. Results for other loci will be presented. Conclusions. Joint modeling of genetic effects give additional insights into the genetic architecture of disease compared to conventional single-variant tests. References Yi, N., & Zhi, D. (2011). Bayesian analysis of rare variants in genetic association studies. Genetic Epidemiology, 35(1), 57–69.

Categories: *Association: Candidate Genes, Association: Unrelated Cases-Controls, Bayesian Analysis, Case-Control Studies*

P 11

Association Between Blood Pressure Susceptibility Loci and Urinary Electrolytes

Bamidele O Tayo[1], Holly Kramer[1], Colin A McKenzie[2], Guichan Cao[1], Ramon Durazo-Arvizu[1], Amy Luke[1], Terrence Forrester[2], Richard S Cooper[1]
[1]Loyola University Chicago, Maywood, IL
[2]University of the West Indies, Kingston, Jamaica

BACKGROUND: Genome-wide association studies have led to identification and validation of about 40 susceptibility loci for blood pressure and hypertension especially among individuals of European ancestry. Even though these genetic variants collectively explain only a small fraction of the heritability for blood pressure phenotypes, similar associations with blood pressure phenotypes remain to be demonstrated in individuals of African ancestry. OBJECTIVE: As part of the study on genetics of hypertension in Blacks, we sought to identify possible associations between BP susceptibility loci and urinary sodium and potassium among individuals of African origin. METHOD: We obtained mean daily urinary sodium and potassium from three 24-hour samples collected from 613 adult Jamaicans that consisted of 140 males and 473 females. The subjects were genotyped using the Illumina MetaboChip genotyping array that contains selected variants for metabolic and atherosclerotic / cardiovascular disease traits. In the present study, we analyzed only the available quality controlled 25 blood pressure susceptibility loci previously reported by The International Consortium for Blood Pressure. Each of the 25 variants was tested for association with urinary sodium and potassium under an additive genetic mode of inheritance using multivariable linear regression model that adjusted for age, sex, body mass index and age-by-sex interaction covariates. To control for possible population stratification from admixture, we also included the first 10 principal components from the autosomal genotypes in the model. RESULTS & CONCLUSION: Our findings reveal association ($p<0.004$) between urinary potassium and variants in the TBX5-TBX3 (rs10850411), ADM (rs7129220) andPLCE1 (rs932764) loci. This study provides preliminary data that genetic variants associated with BP susceptibility may be associated with urinary sodium and potassium excretion; additional studies to confirm these findings are required.

Categories: *Association: Candidate Genes, Cardiovascular Disease and Hypertension*

P 12

A systematic evaluation of short tandem repeats in lipid candidate genes: riding on the SNP-wave

Claudia Lamina[1], Margot Haun[1], Stefan Coassin[1], Anita Kloss-Brandstätter[1], Christian Gieger[2], Annette Peters[3], Konstantin Strauch[2], Lyudmyla Kedenko[4], Bernhard Paulweber[4], Florian Kronenberg[1]

[1]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria

[2]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Neuherberg, Germany

[3]Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

[4]First Department of Internal Medicine, Paracelsus Private Medical University Salzburg, Austria

Structural genetic variants as short tandem repeats (STRs) are not targeted in SNP-based association studies and thus, their possible association signals are missed. We systematically searched for STRs in gene regions known to contribute to total cholesterol, HDL cholesterol, LDL cholesterol and triglyceride levels in two independent studies (KORA F4, n=2553 and SAPHIR, n=1648), resulting in 16 STRs that were finally evaluated. In a combined dataset of both studies, the sum of STR alleles was regressed on each phenotype, adjusted for age and sex. The association analyses were repeated for 1000G imputed SNPs in a 200kb region surrounding the respective STRs in the KORA F4 Study. Three STRs were significantly associated with total cholesterol (within LDLR, the APOA1/C3/A4/A5/BUD13 gene region and ABCG5/8), five with HDL cholesterol (3 within CETP, one in LPL and one in APOA1/C3/A4/A5/BUD13), three with LDL cholesterol (LDLR, ABCG5/8 and CETP) and two with triglycerides (APOA1/C3/A4/A5/BUD13 and LPL). None of the investigated STRs, however, showed a significant association after adjusting for the lead or adjacent SNPs within that gene region. The evaluated STRs were found to be well tagged by the lead SNP within the respective gene regions. Therefore, the STRs reflect the association signals based on surrounding SNPs. In conclusion, none of the STRs contributed additionally to the SNP-based association signals identified in GWAS on lipid traits.

Categories: *Association: Candidate Genes, Quantitative Trait Analysis*

P 13

Linkage disequilibrium mapping of multiple functional loci in case-control studies

Yen-Feng Chiu[1], Li-Chu Chien[1], Kung-Yee Liang[2], Lee-Ming Chuang[3]
[1]National Health Research Institutes, Taiwan, ROC
[2]National Yang Ming University, Taiwan, ROC
[3]National Taiwan University Hospital, Taiwan, ROC

Most complex diseases are multifactorial, involving multiple genetic factors and their joint effects. For such diseases, methods accounting for multiple loci may be more powerful than single-locus analyses and may offer improved precision of disease-locus localization. We propose a semiparametric multipoint linkage disequilibrium (LD) mapping approach to estimate simultaneously the disease loci, the genetic effects of disease loci, and the joint effects and interactions of two adjacent loci, and to construct corresponding CIs for these parameters. This proposed method builds upon large sample properties, which is useful for a high-density genome-wide association study (GWAS) with common variants. Chromosomal regions can be divided by LD blocks or genes to localize functional loci in each subregion. We apply the proposed approach to a data example of case-control studies. Results of the simulations and data example suggest that the developed method performs well in terms of bias, variance, and coverage probability under scenarios with up to three disease loci.

Categories: *Association: Candidate Genes, Association: Genome-wide, Case-Control Studies, Gene - Gene Interaction, Population Genetics*

P 14

Genetic variants in transporter and metabolizing genes and survival in colorectal cancer patients treated with oxaliplatin combination chemotherapy

Elisabeth J Kap[1], Petra Seibold[1], Yesilda Balavarca[2], Lina Jansen[1], Natalia Becker[1], Michael Hoffmeister[1], Cornelia M Ulrich[2], Barbara Burwinkel[3], Hermann Brenner[1], Jenny Chang-Claude[1]
[1]German Cancer Research Center
[2]National Center for Tumor Diseases
[3]University of Heidelberg

Oxaliplatin has become one of the main chemotherapeutic agents for the treatment of colorectal cancer (CRC). Metabolic and transporter enzymes are involved in the clearance of chemotherapeutic agents. Variants in genes encoding these enzymes may cause variation in response to oxaliplatin and could therefore be potential predictive markers. Therefore we comprehensively assessed differential effects of 931 genetic variants in transporter and metabolizing genes and overall survival (OS) in CRC patients who received oxaliplatin chemotherapy compared to patients treated with other chemotherapeutics. We included 623 CRC patients diagnosed between 01.01.2003 and 31.12.2007 and recruited in a German population-based study (DACHS), who received adjuvant chemotherapy (201 patients received oxaliplatin). Survival analysis was performed using a Cox regression model, adjusted for age, sex, UICC stage, cancer site and BMI. Effect modification by oxaliplatin treatment was assessed using a multiplicative interaction term. Median follow-up time in patients receiving oxaliplatin was 4.9 years after which 96 patients were deceased. Rs11203943 (NAT1), rs7017402 (NAT1) and rs4148872 (TAP2) showed differential association with OS according to oxaliplatin treatment (Unadjusted p-values <0.001), although results were not significant after FDR correction (FDR p <0.05). Our data suggest that genetic variants in NAT1 and TAP2 may be predictive markers for oxaliplatin treatment. We plan to use additional SNPs (imputed to the 1000 genome reference panel) to identify further potential predictive markers.

Categories: *Association: Candidate Genes, Cancer*

P 15

Post-Genome-Wide Association Study Using Generalized Structured Component Analysis

Hela Romdhani[1], Aurélie Labbe[1], Heungsun Hwang[1]
[1]McGill University

We are interested in developing a statistical framework for the joint analysis of multiple correlated traits and multiple genotype measures from candidate regions in genetic studies. We propose to use structural equation modeling with latent variables for the association structure between the observed variables and some components mediating the relationships between genotypes and phenotypes. The model is constructed on the basis of prior biological knowledge of both clinical and genetic pathways. We use the Generalized Structured Component Analysis (GSCA) to estimate the model's parameters. Test procedures for different kinds of directed effects measured by GSCA have been developed and powers have been assessed by simulations. Finally, an analysis of the QCAHS survey data is performed using this new approach.

Categories: *Association: Candidate Genes, Multivariate Phenotypes, Pathways*

P 16

Detecting Maternal-Fetal Genotype Interactions Associated with Conotruncal Heart Defects: A Haplotype-based Analysis with Penalized Logistic Regression

Mario A Cleves[1], Ming Li[1], Steve W Erickson[1], Charlotte A Hobbs[1], Jingyun Li[1], Xinyu Tang[1], Todd G Nick[1], Stewart L Macleod[1]
[1]University of Arkansas for Medical Sciences

Non-syndromic congenital heart defects (CHDs) develop during embryogenesis as a result of a complex interplay between environmental exposures, genetics and epigenetic causes. Genetic factors associated with CHDs may be attributed to either independent effects of maternal or fetal genes, or the inter-generational interactions between maternal and fetal genes. Detecting gene-by-gene interactions underlying complex diseases is a major challenge in genetic research. Detecting maternal-fetal genotype (MFG) interactions and differentiating them from the maternal/fetal main effects has presented additional statistical challenges due to correlations between maternal and fetal genomes. Traditionally, genetic variants are tested separately for maternal/fetal main effects and MFG interactions on a single-locus basis. We conducted a haplotype-based analysis with a penalized logistic regression framework to dissect the genetic effect associated with the development of non-syndromic conotruncal heart defects (CTD). Our method allows simultaneous model selection and effect estimation, providing a unified framework to differentiate maternal/fetal main effect from the MFG interaction effect. In addition, the method is able to test multiple highly linked SNPs simultaneously with a configuration of haplotypes, which reduces the data dimensionality and the burden of multiple testing. By analyzing a dataset from the National Birth Defects Prevention Study (NBDPS), we identified seven genes (GSTA1, SOD2, MTRR, AHCYL2, GCLC, GSTM3 and RFC1) associated with the development of CTDs. Our findings suggest that MFG interactions between haplotypes in 3 of 7 genes, GCLC, GSTM3 and RFC1, are associated with non-syndromic conotruncal heart defects.

Categories: *Association: Candidate Genes, Association: Family-based, Association: Unrelated Cases-Controls, Gene - Gene Interaction*

P 17

Mutations screening of exons 7 and 13 of TMC1 gene (DFNB7/11) in Iranian autosomal recessive non-syndromic hearing loss (NSHL) probands using molecular techniques

Payam Ghasemi-Dehkordi[1], Negar Moradipour[1], Fatemeh Heibati[2], Shahrbanuo Parchami-Barjui[1], Ahmad Rashki[3], Morteza Hashemzadeh-Chaleshtori[1]
[1]Cellular and Molecular Research Center, Shahrekord University of Medical Sciences, Shahrekord, Iran
[2]Clinical Biochemistry Research Center, Shahrekord University of Medical Sciences, Sharekord, Iran
[3]Faculty of Veterinary Medicine, Department of Physiopathology, Zabol University, Zabol, Iran

Non-syndromic hearing loss (NSHL) is the most common birth defect which occur in approximately 1/1000 newborns. NSHL is a very heterogeneous trait and could be caused due to both genetic and environmental factors. Mutations of transmembrane channel-like 1 (TMC1) gene cause non-syndromic deafness in humans and mice. The aim of present study was to investigate the association of TMC1 gene mutations of locus DFNB7/11 in exons 7 and 13 in a cohort of 100 patients with hearing loss in Iran using polymerase chain reaction-single stranded conformation polymorphism (PCR-SSCP), heteroduplex analysis (HA), and DNA sequencing. The blood samples of hearing loss patients were collected from 10 provinces of Iran. DNA was extracted from specimens and mutations of exons 7 and 13 of TMC1 gene were investigated using PCR-SSCP. In addition, all samples were checked by heteroduplex analysis (HA) reaction and suspected specimens with shift bands were subjected to DNA sequencing for investigate the presence of any gene variation. In this study, no mutation was found in these two exons of TMC1 gene. These results concluded that TMC1 gene mutations have a very low contribution in patients and were not great clinical importance in these provinces of Iran. However, more studies are need to investigate the relationship between other parts of this gene with hearing loss in different population through the country. More research could clarify the role of this gene and its relation with deafness and provide essential information for the prevention and management of auditory disorder caused by this gene in Iranian population.
Keywords: TMC1 gene, Hearing loss, PCR-SSCP, Heteroduplex analysis, Iran

Categories: *Association: Candidate Genes, Genomic Variation*

P 18

Conotruncal Heart Defects and Common Variants in Maternal and Fetal Genes in Folate, Homocysteine and Transsulfuration Pathways

Mario A Cleves[1], Charlotte A Hobbs[1], Stewart L MacLeod[1], Stephen W Erickson[1], Xinyu Tang[1], Ming LI[1], Jingyun Li[1], Nick Todd[1], Sadia Malik[1]
[1]University of Arkansas for Medical Sciences

Congenital heart defects (CHDs) are the most prevalent structural birth defect, occurring in 8 to 11 of every 1,000 live births. Conotruncal heart defects (CTDs) comprise a subgroup of CHDs that are malformations of cardiac outflow tracts and great arteries. We investigated the association between CTDs and maternal and fetal single nucleotide polymorphisms (SNPs) in 60 genes in the folate, homocysteine and transulfuration pathways. We also examined whether periconceptional maternal folic acid supplementation modified these associations. Participants were enrolled in the National Birth Defects Prevention Study between 1997 and 2007. DNA samples from 616 case-parental triads affected by CTDs and 1,645 control-parental triads were genotyped using a custom Illumina® Golden Gate SNP array. Log-linear hybrid models, optimizing data from case and control triads, were used to identify maternal and fetal SNPs associated with CTDs. Wakefield's Bayesian false-discovery probability method (BFDP) was used to identifying noteworthy associations. Among 921 SNPs, 17 maternal and 17 fetal SNPs had a BFDP <0.8. Ten of the 17 maternal SNPs and 2 of the 17 fetal SNPs were found within the glutamate-cysteine ligase, catalytic subunit (GCLC) gene. Fetal SNPs with the lowest BFDP were found within the thymidylate synthetase (TYMS) gene. Additionally, the genetic risk of CTDs for 19 maternal and 9 fetal SNPs was found to be modified by periconceptional folic acid use. These results support previous studies suggesting that maternal and fetal SNPs within folate, homocysteine and transsulfuration pathways are associated with CTD risk. Maternal use of supplements containing folic acid may modify the impact of SNPs on the developing heart.

Categories: *Association: Candidate Genes, Association: Family-based, Association: Unrelated Cases-Controls*

P 19

Genetic Predisposition of XRCC1 in Schizophrenia Patients of South Indian Population

Sujitha S P[1], Lakshmanan S[2], Harshavaradhan S[3], Gunasekaran S[1], Anilkumar G[1]
[1]School of Biosciences and Technology, VIT University, Vellore 632014 Tamil Nadu, India
[2]Government Vellore Medical College, Vellore, Tamil Nadu, India
[3]Sri Narayani Hospital and Research Centre, Vellore, Tamil Nadu, India

Schizophrenia is a debilitating neuropsychiatric disorder. Several of the previous studies carried out to explore the etiology of this chronic disease suggest for its association with the SNPs (including the non-synonymous ones) at various gene loci; and these investigations produced varying results depending on ethnicity. Role of XRCC1 as a repair gene has been extensively studied on a wide variety of carcinomas. We have compelling reasons to consider this as a candidate gene that could influence schizophrenia. However, barring a few instances, the association studies on schizophrenia and the SNP at XRCC1 are quite meager. The present study, performed on a total of 523 subjects including 260 cases and 263 controls, depicts the association of rs25487 (Arg399Gln) polymorphism of XRCC1 with schizophrenia. The analysis revealed the strong genotypic ('AA'/Gln399Gln; p= 0.006) and allelic ('A'/Gln399; p=0.003, OR=1.448; 95% CI= 1.132 to 1.851) association of the SNP with schizophrenia. We are further encouraged to analyze the association of nicotine (if any) with schizophrenia, inasmuch as the individuals with schizophrenia have shown higher susceptibility to nicotine addiction. This study was performed in two cohorts with 260 case subjects (101 nicotine substance addicts and 159 nicotine substance naïve subjects) and 263 control subjects (with 90 subjects with addiction and 173 subjects without addiction).The study did not show any association of nicotine addiction with this non-synonymous mutation. To conclude, the present study clearly demonstrated the association of Gln399Gln with schizophrenia in Tamil population, and has ruled out the role of nicotine in the polymorphism as an epigenetic factor influencing the disease.

Categories: *Association: Candidate Genes, Epigenetics, Psychiatric Diseases*

P 20

A stochastic search through smoking images in movies, genetic and psycho-social factors associated with smoking initiation in Mexican American youths

Michael D Swartz[1], Matthew D Koslovsky[1], Elizabeth A Vandewater[2], Anna V Wilkinson[3]
[1]University of Texas School of Public Health, Division of Biostatistics
[2]University of Texas School of Public Health, Division of Health Promotion and Behavioral Science
[3]University of Texas School of Public Health, Division of Epidemiology, Human Genetics and Environmental Science

Since smoking is one of the strongest risk factors for lung cancer, identifying factors related to smoking initiation can have a high impact on reducing lung cancer rates. Ethnic differences in initiation rates have been observed, and Mexican American youths have been under studied. Recent independent studies have identified multiple factors associated with smoking initiation in Mexican American youths: exposure to smoking images in movies, genetic, and psycho-social factors. Here we simultaneously investigate all these factors and their potential interactions. Using a prospective cohort of 1,328 Mexican Ameican youths, we investigated single nucleotide polymorphisms (SNPs) from the opioid receptor and dopamine pathways, psycho-social factors and exposure to smoking related images in movies. We measured psycho-social factors using previously validated questionnaires and exposure to smoking images in movies using the Beach method. We used stochastic search variable selection methodology to jointly assess these associations with smoking initiation in Mexican American youths. We used priors that both imposed hierarchical models for interactions and controlled the false positive rate. Our preliminary findings identified smoking images in movies, age, gender, positive outcome expectations from smoking, risk taking tendencies, living with a smoker, peer influence, and serving detention in school, and a SNP on gene SNAP25 and another on OPRM1 related to smoking initiation. We did not identify any interactions.

Categories: *Association: Candidate Genes, Bayesian Analysis, Gene - Environment Interaction, Markov Chain Monte Carlo Methods*

P 21

Association between Apolipoprotein E genotype and cancer susceptibility: a meta-analysis

Anand R[1], Prakash SS[1], Veeramanikandan R[1], Richard Kirubakaran[2]
[1]Christian Medical College, Vellore, India
[2]South Asian Cochrane Center, Christian Medical College, Vellore, Tamilnadu, India-632002.

Apolipoprotein E (ApoE), a protein primarily involved in lipoprotein metabolism occurs in 3 isoforms (E2, E3 and E4). While studies evaluating the association between ApoE genotype and incidence of malignancies are available, the results are inconsistent. The objective of the present study was to analyze the association between APOE genotype and incidence of cancer by a meta-analysis. We conducted a literature search in the electronic databases for studies with information on APOE polymorphisms in malignancies. Sixteen studies (14 case-control/2 cohort; 77970 controls and 12010 cases) were included for the present meta-analysis. Pooled odds ratios (OR) with 95% confidence intervals (CI) were calculated assuming a random-effect model for all the genotypes and alleles. Subgroup analyses based on study design, ethnicity of populations, and site of cancer and source of controls were performed as a post-hoc measure. Appropriate tests to detect heterogeneity, publication bias and sensitivity were done at all stages. The pooled effect measure for the comparisons did not reveal an association in primary analyses. In the subgroup analyses we observed a significant negative association between APOE4+ genotypes and overall risk of cancer in the cohort study subgroup. There was also a weak positive association between APOE4+ genotypes and breast cancer. We observed a moderate inter-study heterogeneity for several of the comparisons (I2<40%). Sensitivity analyses did not alter the overall pooled effect measure in the major comparisons. There were no evidences to suggest a publication bias. Overall, the present meta-analysis did not show any association between APOE alleles or genotypes with incidence of cancer in general.

Categories: *Association: Candidate Genes, Cancer, Case-Control Studies*

P 22

Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of origin effect on body mass index

Clive J Hoggart[1], Giulia Venturini[2], Massimo Mangino[3], Felicia Gomez[4], George Davey-Smith[5], Valentin Rousson[6], Joel N Hirschhorn[7], Carlo Rivolta[1], Ruth JF Loos[8], Zoltan Kutalik[6]
[1]Department of Genomics of Common Disease, Imperial College London, London W12 ONN, UK
[2]Department of Medical Genetics, University of Lausanne, Lausanne 1005, Switzerland
[3]Department of Twin Research & Genetic Epidemiology, King's College London, London SE1 7EH, UK
[4]Department of Genetics, Division of Statistical Genomics, Washington University School of Medicine in St. Louis , St. Louis 63108, USA
[5]MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS82BN, UK
[6]Institute of Social and Preventive Medicine (IUMSP), Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne 1010, Switzerland
[7]Center for Basic and Translational Obesity Research and Divisions of Endocrinology and Genetics, Boston Childrenâ€™s Hospital, Boston 2115, USA
[8]MRC-Epidemiology Unit, University of Cambridge, Cambridge CB2 0QQ, UK

The phenotypic effect of some single nucleotide polymorphisms (SNPs) depends on their parental origin. We present a novel approach to detect parent-of-origin effects (POE) in genome-wide genotype data of unrelated individuals. The method exploits increased phenotypic variance in the heterozygous genotype group relative to the homozygous groups. We applied the method to >56,000 unrelated individuals to search for POEs influencing body mass index (BMI). Six lead SNPs were carried forward for replication in five family-based studies (of ~4,000 trios). Two SNPs replicated: the paternal rs2471083-C allele (located near the imprinted KCNK9 gene) and the paternal rs3091869-T allele (located near the SLC2A10 gene) increased BMI equally (beta=0.11 (SD), P<0.0027) compared to the respective maternal alleles. Real-time PCR experiments of lymphoblastoid cell lines from the CEPH families showed that expression of both genes was dependent on parental origin of the SNPs alleles (P<0.01). Our scheme opens new opportunities to exploit GWAS data of unrelated individuals to identify POEs and demonstrates that they play an important role in adult obesity.

Categories: *Association: Family-based, Association: Genome-wide, Association: Unrelated Cases-Controls, Epigenetics, Quantitative Trait Analysis, Transmission and Imprinting*

P 23

Interactive effect between DNAH9 gene and early-life tobacco smoke exposure in bronchial hyper-responsiveness

Marie-Hélène Dizier[1], Rachel Nadif[2], Patricia Margaritte-Jeannin[1], Sheila J Barton[3], Valérie Gagné-Ouellet[4], Chloé Sarnowski[1], Myriam Brossard[1], Nolwenn Lavielle[1], Jocelyne Just[5], Mark Lathrop[6]
[1]INSERM, U946, Université Paris Diderot, Paris, France
[2]INSERM, U1018, Villejuif, Université Paris Sud, France
[3]Faculty of Medicine, University of Southampton, Southampton, UK
[4]Université du Québec, Chicoutimi, Canada
[5]Centre de l'Asthme et des Allergies, INSERM, UMR_S 1136, Equipe EPAR, France
[6]Mc Gill University, Montréal, Canada

We previously performed a genome-wide linkage analysis of bronchial hyper-responsiveness (BHR) testing interaction with early life environmental tobacco smoke (ETS) exposure in the French Epidemiological study on the Genetics and Environment of Asthma (EGEA) Our goal was to conduct fine-scale mapping of the detected 17p11 region that showed linkage in ETS unexposed siblings only, to identify genetic variants interacting with ETS exposure that influence BHR. Analyses were first performed in the 388 French EGEA asthmatic families, using family-based association test (FBAT). To search for SNP x ETS interaction, we used a two-step strategy: 1) selection of SNPs showing FBAT association signals with BHR ($P<0.01$) in unexposed siblings; 2) FBAT homogeneity test between exposed and unexposed siblings of selected SNPs. For SNPs showing significant interaction, a log-linear modeling approach for testing interaction, as proposed by Umbach and Weinberg (2000), was applied for validation. Replication analyses were then conducted in two independent asthmatic family samples: 253 French-Canadian families (SLSJ) and 341 UK families. In EGEA families, 17 SNPs showed association signals with BHR in unexposed siblings. A single SNP showed significant interaction with ETS exposure using both methods ($P\leq10\text{-}3$). This result was replicated in the SLSJ families and meta-analysis of the two samples provided a strong improvement in the detection of interaction ($P=7.10\text{-}5$). There was however no replication in the UK families. The SNP showing significant interactive effect with ETS exposure in BHR is in a promising candidate gene, DNAH9, a gene well known to be associated with Primary Ciliary Dyskinesia. Funded: ANR-GWIS-AM-2011, Région IdF

Categories: *Association: Family-based, Gene - Environment Interaction, Multifactorial Diseases*

P 24

Detection of rare highly penetrant recessive variants using GWAS data

Steven Gazal[1], Mourad Sahbatou[2], Marie-Claude Babron[1], Jean-Charles Lambert[3], Philippe Amouyel[3], Emmanuelle Génin[4], Anne-Louise Leutenegger[1]
[1]INSERM U946, Paris, France
[2]CEPH, Paris, France
[3]INSERM U744, Lille, France
[4]INSERM U1078, Brest, France

Genome-wide association studies (GWAS) have identified several common genetic variants in multifactorial diseases. However, taken together, these variants only explain a small part of the heritability. Different candidates have been suggested to explain this missing heritability, and among them are variants with recessive effects that could play a role but have not been detected so far. Recessive variants are easy to detect when they are rare, fully penetrant, and involved in rare monogenic diseases. The strategy of choice to detect them is homozygosity mapping (HM), a powerful approach that consists in focusing on inbred families and searching for a region of the genome of shared homozygosity in the inbred cases. With the help of genome-wide genetic data, it is now possible to determine if an individual is inbred based on the observed genome homozygosity patterns. HM can then be performed without any knowledge of the genealogy. This could be used not only to detect rare recessive variants involved in monogenic diseases, but also to identify recessive Mendelian subentities of multifactorial disease. Several software have been developed to study inbreeding. However, none of them provide an integrative solution to estimate inbreeding, identify and visualize runs of homozygosity by descent and perform HM. We have recently developed the FSuite pipeline to open up the possibility to easily detect inbred cases in GWAS dataset, and to focus on them to perform HM allowing for heterogeneity. We will illustrate the possibilities offered by FSuite on a French GWAS dataset including 1,886 affected individuals with Alzheimer's disease. About 5% of the cases were found inbred and were eligible for HM, allowing the detection of 3 candidate genomic regions.

Categories: *Association: Family-based, Association: Genome-wide, Inbreeding, Isolate Populations, Multifactorial Diseases*

P 25

Copy Number Variation (CNV) detection in whole exome sequencing data for Mendelian disorders

Peng Zhang[1], Hua Ling[1], Elizabeth Pugh[1], Kurt Hetrick[1], Dane Witmer[1], Nara Sobreira[2], David Valle[2], Kim Doheny[1]
[1]Center for Inherited Disease Research, Institute of Genetic Medicine, The Johns Hopkins School of Medicine
[2]Institute of Genetic Medicine, The Johns Hopkins School of Medicine

The Centers for Mendelian Genomics (CMG) project uses next-generation sequencing and computational approaches to discover the genes and variants that underlie Mendelian conditions. While SNVs and INDELs explain some Mendelian conditions, many remain unresolved. We are interested to know to what extent unrecognized CNVs would resolve some of these. Compared to whole genome sequencing (WGS), whole-exome sequencing (WES) is a cost-effective alternative for finding disease genes harboring variants with relatively large effect size. However, identifying CNVs from WES has been a challenge because of the sparseness of the target regions and the non-uniform distribution of reads across genome. As part of the CMG project, we applied four prevailing CNV calling methods (XHMM, CoNIFER, ExomeDepth, and EXCAVATOR) on 677 WES samples (including 41 HapMap controls) to search for rare exonic CNVs that might be causal for the disease of interest. In our preliminary analysis, CoNIFER, ExomeDepth, XHMM, and EXCAVATOR detected an average of 3.5, 208, 13.3, and 58 CNVs (for sizes larger than 300 bp) per sample, respectively. Our initial analyses of three unsolved consanguineous pedigrees with the same phenotype revealed a homozygous two exon deletion (~ 2.45 kb) in a known causal gene in two of the families. We will compare the results between methods, examine the impact of controls used, and review a subset of findings in IGV.

Categories: *Association: Family-based, Association: Genome-wide, Bioinformatics, Case-Control Studies, Causation, Copy Number Variation, Data Integration, Data Mining, Fine Mapping, Genomic Variation, Linkage and Association, Sequencing Data*

P 26

Combining genetic and epigenetic information identified imprinted 4q35 variant associated with the combined asthma-plus-rhinitis phenotype

Chloé Sarnowski[1], Catherine Laprise[2], Miriam Moffatt[3], Giovanni Malerba[4], Andréanne Morin[2], Quentin Vincent[5], Klaus Rohde[6], Marie-Hélène Dizier[1], Jorge Esparza-Gordillo[6], Emmanuelle Bouzigon[1]

[1]1) U946, INSERM, PARIS, France; 2) Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, France

[2]3) Université du Quábec à Chicoutimi, Canada

[3]4) National Heart Lung Institute, Imperial College, UK

[4]5) Section of Biology and Genetics, Department of Life and Reproduction Sciences, University of Verona, Italy

[5]6) U1163, INSERM, PARIS, France

[6]7) Max-Delbrück-Center for Molecular Medicine (MDC), Berlin, Germany

We previously detected a linkage signal in the 4q35 region with the combined asthma-plus-rhinitis phenotype (AST+AR) in 615 European families when accounting for maternal imprinting (p=7x10-5). To further investigate this region, we tested the association between 1,300 SNPs (spanning 6 Mb) and AST+AR in 162 French EGEA families ascertained through asthma using the Parent-of-Origin-Likelihood Ratio Test. Replication analysis was performed in 152 asthmatic French Canadian SLSJ families for 18 SNPs detected at p<0.005. The top-replicated SNP (rs10009104) lying at 1.6 Mb from the linkage peak was detected under a best-fitting maternal imprinting model (pmeta=4x10-5) and accounted for most of the linkage signal. Many cis-regulatory elements are described in a 50 kb surrounding region of this SNP. Using the Quantitative Transmission Disequilibrium Test (QTDT), we tested for association between rs10009104 and 26 DNA methylation probes of that region, measured in white blood cells of 159 individuals (40 SLSJ families), while accounting for parent-of-origin effect and adjusting for AST+AR. Maternally inherited risk allele of rs10009104 was associated with increased methylation of the top-ranked probe (p<10-5 after permutations). This probe lies at 529 bp from the SNP and within regulatory elements that include a predicted active promoter in lung fibroblasts, DNase I hypersensitive clusters, and binding sites of two transcription factors involved in inflammatory response initiation (RelA and NF-κB) . This study identified a maternally imprinted SNP that affects AST+AR through an epigenetic mechanism. Funded: Conseil Régional Ile de France, ANR GWIS-AM, EC-FP6

Categories: *Association: Family-based, Epigenetic Data, Linkage and Association, Multifactorial Diseases, Transmission and Imprinting*

P 27

BAYESIAN LATENT VARIABLE COLLAPSING MODEL FOR DETECTING RARE VARIANT INTERACTION EFFECT IN TWIN STUDY

Liang He[1], Mikko J Sillanpää[2], Samuli Ripatti[3], Janne Pitkäniemi[4]
[1]Department of Public Health, Hjelt Institute, University of Helsinki, Finland
[2]Department of Mathematical Sciences, University of Oulu, Oulu FIN-90014, Finland; Department of Biology and Biocenter Oulu, University of Oulu, Oulu FIN-90014, Finland
[3]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Finland; Wellcome Trust Sanger Institute, UK
[4]Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland; Department of Public Health, Hjelt Institute, University of Helsinki, Finland

By analysing more next-generation sequence data than before, researchers have affirmed that rare genetic variants are widespread among populations and likely play an important role in complex phenotypes. Recently, a handful of statistical models have been developed to analyse rare variant association in different study designs. However, due to the scarce occurrence of minor alleles in data, appropriate statistical methods for detecting rare variant interaction effects are still difficult to develop. We propose a hierarchical Bayesian latent variable collapsing method (BLVCM), which circumvents the obstacles by parameterizing the signals of rare variants with latent variables in a Bayesian framework and is parameterised for twin data. The BLVCM manages to tackle non-associated variants, allow both protective and deleterious effects, capture SNP-SNP synergistic effect, provide estimates for the gene level and individual SNP contributions, and can be applied to both independent and various twin designs. We assess the statistical properties of the BLVCM using simulated data, and find that it achieves better performance in terms of power for interaction effect detection compared to the Granvil and the SKAT. As proof of practical application, the BLVCM is then applied to a twin study analysis of more than 20,000 gene regions to identify significant rare variants associated with low-density lipoprotein cholesterol (LDL-C) level. The results show that some of the findings are consistent with other previous studies, and some novel gene regions with significant SNP-SNP synergistic effects are identified. Key words: rare variant; Bayesian collapsing model; genetic association; LDL-C; twin study

Categories: *Association: Family-based, Association: Genome-wide, Bayesian Analysis, Gene - Gene Interaction, Genomic Variation, Markov Chain Monte Carlo Methods, Multilocus Analysis, Pathways, Population Genetics, Quantitative Trait Analysis, Sequencing Data*

P 28

Rare Variant Association Test for Nuclear Families

Zong-Xiao He[1], Niklas Krumm[2], Gao T Wang[1], Brian J O'Roak[3], Simons Simplex Sequencing Consortium, Evan E Eichler[3], Suzanne M Leal[3]
[1]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine
[2]Department of Genome Sciences, University of Washington
[3]Department of Molecular and Medical Genetics, Oregon Health and Science University

Population-based complex trait association studies of rare variants (RVs) are vulnerable to spurious associations due to population stratification. Analyzing trio data using the RV-transmission disequilibrium test [RV-TDT (He et al. 2014)] can avoid this problem. The TDT analyses only employ information on an affected offspring and their parents. When there are siblings, including them in analysis can provide additional association information. We extended the RV-TDT to analyze all types of independent nuclear families (NF) with at least one affected offspring (RV-NF). For all RV-NF tests type I error is well controlled even when there is a high level of population stratification or admixture. The power of the RV-NF test was evaluated using a number of disease models and nuclear pedigree configurations. The RV-NF is considerably more powerful than the RV-TDT to detect associations. For the RV-TDT and RV-NF power was evaluated by generating data for a 1,500bp gene for which the causal RVs have an odds ratio of 2. The power to detect and association is: 0.49 for 1,000 trios; 0.58 for 1,000 NF with one affected child and an unaffected child; and 0.65 for 1,000 NF with two affected children. In order to illustrate the application of the RV-NF methods, the exome data from 600 autism spectrum disorder NF with one affected child and one unaffected child were analyzed. RV associations with autism were found for several genes. Given the problem of adequately controlling for population stratification and admixture in RV association studies, the capability of analyzing all types of NFs and the growing number of NF studies with sequence data, the RV-NF method is extremely beneficial to elucidate the involvement of RVs in disease etiology.

Categories: *Association: Family-based, Sequencing Data*

P 29

Sample size and power determination for association tests in case-parent trio studies

Holger Schwender[1], Christoph Neumann[2], Margaret A Taub[3], Samuel G Younkin[4], Terri H Beaty[3], Ingo Ruczinski[3]
[1]Heinrich Heine University
[2]TU Dortmund University
[3]Johns Hopkins University
[4]University of Wisconsin

Transmission/disequilibrium tests (TDTs) are the most popular statistical tests for detecting single nucleotide polymorphisms (SNPs) associated with disease in case-parent trio studies considering genotype data from children affected by a disease and from their parents. Since several types of these TDTs have been devised, e.g., approaches based on alleles or on genotypes, it is of interest to evaluate which of these TDTs have the highest power in the detection of SNPs associated with disease. Since the test statistic of the genotypic TDT – which is equivalent to a Wald test in a conditional logistic regression model – had to be computed numerically, comparisons of other TDTs with the genotypic TDT have so far been based on simulation studies. Recently, we, however, have derived a closed-form solution for the genotypic TDT so that this analytic solution can be used to derive equations for power and sample size calculation for the genotypic TDT. In this presentation, we show how these equations can be derived and compare the power of the genotypic TDT with the one of the corresponding score test assuming the same underlying genetic mode of inheritance as well as the allelic TDT based on a multiplicative mode of inheritance.

Categories: *Association: Family-based, Association: Genome-wide, Linkage and Association, Maximum Likelihood Methods, Sample Size and Power*

P 30

Integration of DNA sequence variation and functional genomics data to infer causal variants underlying chemotherapeutic induced cytotoxicity response

Ruowang Li[1], Dokyoon Kim[1], Scott M Dudek[1], Marylyn D Ritchie[1]
[1]Center for Systems Genomics, The Pennsylvania State University, State College, PA

Carboplatin is a widely used chemotherapeutic drug for ovarian and lung cancer. Despite its broad usage, some patients experience severe side effects including myelosuppression and mucositis. Understanding the drug-induced cytotoxicity could potentially lead to personalized treatment. However, finding the causal genetic variants that influence the drug's cytotoxicity has been challenging. To identify variants that are key for carboplatin response, we performed an analysis that jointly analyzed DNA sequence variation and functional genomics data in CEU and YRI HapMap populations. Carboplatin response was measured on the CEU and YRI lmphoblastoid cell lines in terms of IC50, concentration required to stop 50% of cell growth. Using whole genome sequencing data from the 1000 Genomes Project and RNA sequencing data from the GEUVADIS project, we identified candidate genetic variants and gene expression variables that are associated with carboplatin IC50. To uncover potential interactions between candidate variants and gene expression factors, we integrated the candidates using grammatical evolution neural network implemented in ATHENA. The integration analysis identified unique sets of genetic variants and gene expression factors in interaction models in both CEU and YRI population with high predictive power (R2 >60%). To avoid selection bias, we also identified variants that are in linkage disequilibrium with the candidate variants. We then prioritized all the variants based on hundreds of functional genomic annotations from the ENCODE project, including genes, enhancers, and DNase-I sites. Based on the consistency and enrichment of functional annotations, we found potential causal variants for carboplatin response.

Categories: *Association: Genome-wide, Cancer, Causation, Data Integration, Genomic Variation*

P 32

Imputation for SNPs using summary statistics and correlation between genotype data

Sina Rüeger[1,2], Zoltán Kutalik[1,2]
[1]Institute of Social and Preventive Medicine, University Hospital and University of Lausanne, Lausanne
[2]Switzerland Swiss Institute of Bioinformatics, Lausanne, Switzerland

Genome-wide association studies use microarrays to measure SNPs that are often designed to tag many untyped variants, which can be imputed via the linkage disequilibrium (LD) between measured and untyped markers. The imputation methods, while making most of the available data, are computationally very expensive when it comes to imputing ~30-40M variants of the 1000 Genomes panel. These imputed variants are subsequently subjected to association with various traits.

We propose an approach that performs imputation directly on the association summary statistics (such as t-statistics) of typed SNPs. This allows a fast inference of the association strength of non-genotyped markers using that of the tagging SNPs. This approach bears similarities with the pioneering work of Pasaniuc et al. (2013). The novelty of our method lies in the optimized regularization of the pair-wise marker correlation matrix, a modified conditional expectation. It also allows for associations derived from different sample sizes. We reached further improvements by selecting the most relevant reference haplotype sets in order to impute summary statistics.

For testing we used the lipid association meta-analyses summary statistics from Willer et al. (2013). Using the association statistics from HapMap SNPs only, we imputed the effect size of non-HapMap SNPs and compared to the "true" effect size estimates resulting from genotype imputation and association. The results suggest that our test statistics agree closer ($r^2=0.87$) with the true values than the estimates provided by previous methods ($r^2=0.82$).

Such fast and accurate imputation methods will become increasingly important as reference panels grow in size and genotype imputation turns out to be less feasible.

Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, Hirschhorn J, Strachan DP, Patterson N, Price AL (2013) Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. ArXiv:1309.3258v1 [q-bio.QM]

Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, et al.; Global lipids genetics consortium (2013) Discovery and refinement of loci associated with lipid levels. Nature Genetics 45 (11), 1274-1283.

Categories: *Association: Genome-wide, Linkage and Association, Missing Data*

P 33

Evaluation of population stratification in a large biobank linked to Electronic Health Records

Mariza de Andrade[1], Gerard Thromp[2], Amber Burt[3], Daniel S Kim[4], Shefali S Verma[3], Anastasia M Lucas[3], Sebastian M Armasu[1], John A. Heit[1], Geoffrey M Hayes[5], Helena Kuivaniemi[2]
[1]Mayo Clinic, Rochester, MN, USA
[2]Geisinger Health System, Danville, PA, USA
[3]Pennsylvania State University, University Park, PA, USA
[4]University of Washington, Seattle, WA, USA
[5]Northwestern University, Chicago, IL, USA

For genomic association studies, combining samples across multiple studies in Networks or "Big Science" is standard practice. Increasing the number of subjects allows for power needed to assess association. Controlling for genomic ancestry is common, but there is a need to standardize the approach when calculating principal components (PCs) across cohorts such as elimination of SNPs with linkage disequilibrium (LD) pruning at r = 0.5 and a MAF < 0.03. Due to heterogeneity between sites, adjusting for PCs only, does not remove the site and platform bias. Therefore, we propose an alternative approach of generating PCs for our cohort to control for site and platform bias in addition to ancestry difference. Our approach consists on deriving the PCs using the loadings calculated from reference samples, much like generating PCs using the founders of families. We applied our approach using the electronic Medical Records and Genomics (eMERGE) Venous Thromboembolism African ancestry cohort that consists of four adult sites and four genotyping platforms that had previously been analysed controlling for site, platform and ancestry. Our results showed that our approach provided similar association results while both controlling for inflation (λ = 1.01 and 1.02 for standard and loadings, respectively) with the advantages of controlling for fewer covariates, thus less degrees of freedom. Therefore, we expect this approach will serve as a "Best Practices" for similar projects, and as a reference for assessing and controlling for confounders in addition to ancestry in genetic association studies.

Categories: *Association: Genome-wide, Population Stratification*

P 34

Estimating genetic effects on susceptibility and infectivity for infectious diseases

Floor Biemans[1], Piter Bijma[2], Mart CM De Jong[3]
[1]Quantitative Veterinary Epidemiology Group, Wageningen University; Animal Breeding and Genomics Centre, Wageningen University
[2]Animal Breeding and Genomics Centre, Wageningen University
[3]Quantitative Veterinary Epidemiology Group, Wageningen University

Transmission of infectious diseases is determined by susceptibility and infectivity of the individuals involved. An individual's genes for susceptibility affect the disease status of the individual itself, and thus represent a direct genetic effect. An individual's genes for infectivity, on the other hand, affect the disease status of other individuals, and thus represent a so-called indirect genetic effect (IGE). An IGE is a genetic effect of an individual on the phenotype of another individual. IGEs have been studied extensively in evolutionary biology, and can have profound effects on the rate and direction of evolution by natural selection. In genetic studies on infectious diseases, the current focus is largely on susceptibility, whereas genetics of infectivity can have major effects on disease transmission. However, little is known about the genetic background of infectivity. We show how genetic effects on susceptibility and infectivity can be estimated simultaneously from time-series data on disease status of individuals. An endemic disease was simulated, and the disease status (0/1) and genotype of individuals were recorded at several points in time. These data were analysed using a generalized linear model (GLM) with a complementary log-log link function. The model included two genetic terms: i) the genotype of the focal individual, representing susceptibility, and ii) the average genotype of its infected social partners (contacts), representing infectivity. First results showed that estimated genetic effects were almost unbiased. This work, therefore, provides a tool for genome-wide association studies aiming to identify genomic regions affecting susceptibility and infectivity of individuals to endemic diseases.

Categories: *Association: Genome-wide, Genomic Variation, Haplotype Analysis, Prediction Modelling*

P 35

Combined Methods to Explore Genetic Etiology of Related Complex Diseases

Shefali Setia Verma[1], Anurag Verma[1], Anastasia Lucas[1], Jim Linneman[2], Peggy Peissig[2], Murray Brilliant[2], Catherine A McCarty[3], Jonathan L Haines[4], Tamara R Vrabec[5], Gerard Tromp[5]
[1]Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA
[2]Marshfield Clinic, Marshfield, WI, USA
[3]Essentia Rural Health, Duluth, MN, USA
[4]Case Western University, Cleveland, OH, USA
[5]Geisinger Health System, Danville, PA, USA

Genome-wide association studies (GWAS) have identified several SNPs associated with either glaucoma or ocular hypertension (OHT). However, these susceptibility loci explain a small fraction of the genetic risk. Gene-gene interaction (GxG) studies are considered a potential avenue to identify this missing heritability. Using a dataset from the eMERGE (electronic Medical Records and Genomics) Network, which included GWAS data imputed using the 1000 Genomes, we were able to explore the genetic etiology of two very related common eye- diseases: glaucoma and OHT. OHT is one of the leading risk factor for glaucoma, thus we explored the relationships between these two traits at the molecular level. A total of 3,253 (glaucoma) and 3,154 (OHT) unrelated samples of ages 40-90 were extracted from the eMERGE study biorepositories. First, we performed GWAS and GxG studies for each trait using the imputed dataset and identified several main effects and GxG models that meet Bonferroni significance. Secondly, from the obtained GWAS with main effect $p< 0.01$, we also performed a pathway-enrichment analysis using KEGG database on both of these traits combined. Interestingly, we observed that genes in ABC transporter pathway are found to be associated with both glaucoma and OHT. The ABCA4 gene is highly associated with glaucoma and also shows significant interaction with GAD2 gene in OHT ($p= 2.71x10-11$). Lastly, out of 10 pathways shared between the two traits, ABC transporter genes are found to be highly associated with both the traits. In conclusion, we were able to identify novel SNP associations and GxG interactions for these traits and demonstrate the relationship between these two traits at the molecular level with the guidance of pathway analysis.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Bioinformatics, Gene - Gene Interaction, Pathways*

P 36

Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations

Yijuan Hu[1], Yun Li[2], Paul L Auer[3], Danyu Lin[4]
[1]Department of Biostatistics and Bioinformatics, Emory University, USA
[2]Department of Biostatistics, Department of Genetics, University of North Carolina, Chapel Hill, USA
[3]Joseph J. Zilber School of Public Health, University of Wisconsin, Milwaukee, USA
[4]Department of Biostatistics, University of North Carolina, Chapel Hill, USA

In the large cohorts typically used for genome-wide association studies (GWAS), it is prohibitively expensive to sequence all cohort members. A cost-effective strategy is to sequence subjects with extreme values of quantitative traits or those with specific diseases. By imputing the sequencing data from the GWAS data for the cohort members who are not selected for sequencing, one can dramatically increase the number of subjects with information on rare variants. However, treating the imputed rare variants as observed quantities in downstream association analysis may inflate the type I error, especially when the sequenced subjects are not a random subset of the whole cohort. Although the problem can be alleviated by restricting the analysis to variants that are accurately imputed, a large number of rare variants will be excluded as a result. In this article, we provide a valid and efficient approach to combining observed and imputed data on rare variants. We consider all commonly used gene-level association tests, including the burden test, variable threshold (VT) test, and sequence-kernel association test (SKAT), all of which are based on the score statistic for assessing the effects of individual variants on the trait of interest. We show that the score statistic based on the observed genotypes for sequenced subjects and the imputed genotypes for non-sequenced subjects is unbiased. We construct a robust variance estimator that reflects the true variability of the score statistic regardless of the sampling scheme and imputation quality, such that the corresponding association tests always have correct type I error. We demonstrate through extensive simulation studies that the proposed tests are substantially more powerful than the use of accurately imputed variants only and the use of sequencing data alone. We provide an application to the Women's Health Initiative (WHI). The relevant software is freely available.

Categories: *Association: Genome-wide, Data Integration*

P 37

A method for fast computation of the proportion of variants affecting a complex disease and of the additive genetic variance explained in GWAS SNP studies.

Luigi Palla[1], Frank Dudbridge[1]
[1]Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine

Recent research has addressed the estimation of variance explained by large sets of SNPs from a genomewide panel. A method based on polygenic scoring was proposed by Stahl et al (Nat Genet 2012) to estimate both variance explained and number of SNPs affecting the trait, via computationally intensive Bayesian methodology. We propose a fast analytic method based on the formula for the noncentrality parameter of the association test of a polygenic score with the trait of interest (Dudbridge, PLoS Genet 2013). We show how model parameters can be estimated from the results of multiple polygenic score tests based on SNPs with P-values falling in different intervals. We estimate model parameters using maximum likelihood and use a profile likelihood approach that allows rapid computation of reliable confidence intervals. We illustrate our method on several examples of complex diseases. We compare various choices for constructing polygenic scores, based on nested or disjoint intervals of p-values and on weighted or unweighted SNP effect sizes, in estimating variance explained (vg), fraction of genes affecting the trait (nf) and covariance between effects in training and replication samples. We find that for estimation of vg and nf only, the estimates are nearly unbiased and confidence intervals narrow, with less bias for disjoint intervals. When estimating all 3 parameters the estimates present even smaller bias, larger confidence intervals, but incur a larger bias for vg in the case of nested intervals. Overall we recommend use of this method based on the results derived from disjoint intervals.

Categories: *Association: Genome-wide, Case-Control Studies, Maximum Likelihood Methods, Quantitative Trait Analysis*

P 38

Correcting for sample overlap in cross-trait analysis of GWAS

Marissa LeBlanc[1], Verena Zuber[2], Arnoldo Frigessi[3], Bettina Kulle Andreassen[1]
[1]Epi-Gen, Institute of Clinical Medicine, Akershus University Hospital, University of Oslo, Oslo, Norway and Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Norway
[2]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway and Prostate Cancer Research Group, Centre for Molecular Me
[3]Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Norway and Statistics for Innovation, Norwegian Computing Center, Oslo, Norway

There is a growing interest in integrating genomic data over different traits, at the summary statistics level. This is of biological interest due to the partially shared genetic basis of many traits, termed pleiotropy. Using, for example, meta-analysis or a conditional false discovery rate (FDR) framework, pleiotropy can be leveraged to improve detection of common genetic variants involved in disease. This requires only summary statistics, not individual-level data. Summary statistics from genome-wide association studies (GWAS) conducted by global consortia are becoming easier to obtain, however these GWAS summary statistics are often not independent across traits due to partially overlapping samples. Our aims are twofold. First, we show the impact of sample overlap on cross-trait analysis of GWAS, and demonstrate with simulations that it can induce spurious correlation and an increased proportion of false positive findings. Second, we propose a correction that removes the spurious effects due to sample overlap. This correction involves first estimating the correlation of the summary statistics from the two studies (for all possible combinations of quantitative and binary outcomes), and then second, correcting for this spurious correlation via the Mahalanobis transformation. We present results from simulation studies and from actual GWAS data that show that the proposed correction for sample overlap properly controls for false positive findings while still allowing for the detection of true pleiotropic findings.

Categories: *Association: Genome-wide, Data Integration*

P 39

Epigenome-wide association study of centralized adiposity in 2,083 African Americans: The Atherosclerosis Risk in Communities (ARIC) Study

Lindsay Fernández-Rhodes[1], Yun Li[1], Mariaelisa Graff[1], Weihua Guan[2], Megan L Grove[3], Qing Duan[1], Guosheng Zhang[1], Myriam Fornage[3], James Pankow[2], Ellen W Demearath[2]
[1]University of North Carolina at Chapel Hill, North Carolina, USA
[2]University of Minnesota, Minnesota, USA
[3]University of Texas Health Science Center at Houston, Texas, USA

Central obesity is a leading predictor of cardiometabolic risk and its prevalence in the United States (US) has more than doubled since the 1980s, especially in US minorities. Evidence suggests that genetic factors contribute to central adiposity, measured as waist to hip ratio adjusted for body mass index (WHRa). DNA methylation patterns, a well-studied form of epigenetic modification, may also associate with WHRa. This study aims to examine the cross-sectional association between genome-wide CpG site methylation and WHRa in African Americans.

The Infinium Human Methylation 450K BeadChip was used to measure methylation in bisulphite-converted peripheral blood DNA from 2,083 African Americans (mean age 56.6 years) in the Atherosclerosis Risk in Communities study. Linear mixed effects models were used to test for association between methylation beta values and WHRa accounting for random effects for batch and fixed effects for age, sex, center, education, concurrent white blood cell count, household income, smoking, alcohol consumption, physical activity, five leukocyte cell type proportions, and principal components derived from genome-wide exonic genotype data.

We observed one significant negative association with WHRa at cg00574958 ($p=7x10-12$), which lies in the 5'UTR of CPT1A, a gene previously implicated with metabolic related traits. Weaker associations ($p<1x10-6$) were also observed at several autosomal sites requiring future independent replication.

Our observed CpG site association at a known metabolic locus suggests that epigenetic signatures of central adiposity may account for some of the missing heritability and inform our understanding of metabolic dysregulation.

Categories: *Association: Genome-wide, Cardiovascular Disease and Hypertension, Epigenetic Data, Epigenetics*

P 40

Can low-frequency variants be rescued in genome-wide association studies using sparse data methods?

Ji-Hyung Shin[1], Shelley B Bull[1]
[1]Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital & Dalla Lana School of Public Health, University of Toronto

For many complex traits, genetic variants that occur with low frequency (MAF<5%) are thought to be important. However, genome-wide scans of binary traits usually exclude low frequency variants even when the sample size is moderate, because conventional logistic regression inference can fail due to low counts of the observed low-frequency variants. Alternatively, sparse data methods such as Firth's penalized logistic regression likelihood ratio test or a small-sample-adjusted score test implemented in SKAT can provide valid results for a single variant. We investigate the performance of the standard logistic regression and the sparse data methods, using analytic derivations and finite-sample simulations across various scenarios. In the analytic investigation, we examine the simple case of a 2-by-2 contingency table to gain insight into differences among the methods. The analytic calculations show how the test statistics depend on the observed numbers of affected and unaffected individuals with the low-frequency variant, and on the disease prevalence. In the simulation study, we consider an additively coded genotype and a quantitative covariate, and vary disease prevalence, minor allele frequency and counts, and effect size of the genetic covariate to examine a range of settings. We find that no one test is uniformly better than the others. Overall, type 1 error rates are closest to the nominal level for the penalized likelihood ratio test and the small-sample-adjusted score test, while type 1 error rates for the other tests can be greatly inflated or deflated. The power for the small-sample-adjusted score test tends to be slightly higher than the penalized likelihood ratio test, but the difference may be insignificant in practice.

Categories: *Association: Genome-wide, Maximum Likelihood Methods*

P 41

A novel kernel-based statistical approach to testing association in longitudinal genetic studies with an application of alcohol use disorder in a veteran cohort

Zuoheng Wang[1], Zhong Wang[1], Joseph L. Goulet[1], John H. Krystal[1], Amy C. Justice[1], Ke Xu[1]
[1]Yale University

Alcohol dependence (AD) is a major public health concern in the United States and contributes to the pathogenesis of many diseases. The risk for AD is multifactorial including both genetic and environmental factors. Currently, the confirmed associations account for a small proportion of overall genetic risks for AD. Multiple measurements in longitudinal genetic studies provide a route to reduce noise and correspondingly increase the strength of signals in genome-wide association studies (GWAS). In this study, we developed a powerful kernel-based statistical method for testing the joint effect of gene variants with a gene region on disease outcomes measured over multiple time points. We applied the new method to a longitudinal study of veteran cohort (N=960) with both HIV-infected and HIV-uninfected patients to understand the genetic risk underlying AD. We found an interesting gene that may involve the interaction of HIV replication, suggestive of potential gene by environment effect in alcohol use and HIV. We also conducted simulation studies to access the performance of the new statistical methods and demonstrated a power gain by taking advantage of repeated measurements and aggregating information across a biological region. This study not only contributes to the statistical toolbox in the current GWAS, but also potentially advances our understanding of the etiology of AD. Acknowledgment: The authors thank the Veterans Aging Cohort Study and VA National Center for PTSD for generous support. The study is supported by NIH grant R21AA022870.

Categories: *Association: Genome-wide, Multiple Marker Disequilibrium Analysis, Psychiatric Diseases*

P 42

A Gene-Environment Interaction Between Copy Number Burden and Ozone Exposure in Relation to Risk of Autism

Dokyoon Kim[1], Heather Volk[2], Sarah A Pendergrass[1], Molly A Hall[1], Shefali S Verma[1], Santhosh Girirajan[1], Irva Hertz-Picciotto[3], Marylyn D Ritchie[1], Scott B Selleck[1]
[1]Department of Biochemistry & Molecular Biology, the Pennsylvania State University, University Park, PA
[2]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA; Department of Pediatrics, Childrenâ€™s Hospital Los Angeles, University of Southern California, Los Angeles, CA
[3]Department of Public Health Sciences, University of California, Davis, Davis, CA

Autism is a neurodevelopmental disorder characterized as a complex trait with a high degree of heritability as well as a documented susceptibility from environmental factors.The relative contributions of genetic factors, environmental factors and their interactions as they relate to risk of autism are poorly understood.While most autism related copy number variations (CNV) identified to date, each with a substantial risk, are highly penetrant for this disorder, they constitute large rare events contributing modestly to the overall heritability.Genome-wide analysis of CNVs have demonstrated a continuous risk of autism associated with the global level of copy number burden, measured as total base pairs of duplication or deletion across the genome.In addition, environmental exposure to air pollutants has been identified as a risk factor for developing autism.We have examined the relative contribution of CNV (measured as total base pairs of copy number burden), exposure to air pollution, and the interaction between air pollutant levels and copy number burden in a population based case-control study, Childhood Autism Risks from Genetics and Environment (CHARGE).A significant and sizable interaction was identified between duplication burden and ozone exposure (OR 2.78, P<0.005), greater than the main effect for either copy number duplication (OR 2.41, 95% CI: 1.36~4.82) or ozone alone (OR 1.19, 95% CI: 0.75~1.89).The overall implication of our finding is that significant gene-environment interactions associated with autism exist and could account for a considerable level of heritability not detected by evaluating DNA variation or environment alone.

Categories: *Association: Genome-wide, Copy Number Variation, Gene - Environment Interaction*

P 43

Choosing a case-control association test statistic for low-count variants in the UK Biobank Lung Exome Variant Evaluation Study

Nick Shrine[1], Louise V Wain[1], Ioanna Ntalla[1], James P Cook[1], Andrew P Morris[2], Eleftheria Zeggini[3], Jonathan Marchini[4], David P Strachan[5], Ian P Hall[6], Martin D Tobin[1]
[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom
[2]Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom
[3]Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom
[4]Department of Statistics, University of Oxford, Oxford, United Kingdom
[5]Population Health Research Institute, St George's University of London, London, United Kingdom
[6]Division of Therapeutics and Molecular Medicine, University of Nottingham, Nottingham, United Kingdom

The UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study is a nested case-control study to evaluate genetic susceptibility to chronic obstructive pulmonary disease (COPD), genetic variants associated with lung function and genetic resistance to tobacco smoke. 50K UK Biobank individuals were sampled from the extremes and middle of the lung function distribution in smoking and non-smoking strata. In order to identify rare, putative functional genetic variants, genome-wide genotyping was undertaken using a custom designed Affymetrix array that included 130K rare missense and loss of function variants, 642K variants selected for optimal imputation of common variation and improved imputation of low frequency variation (MAF 1-5%) and 9000 variants selected for improved coverage of known and candidate respiratory regions. Simulations have shown that logistic regression with the usual Wald test statistic at low minor allele count (MAC) is highly conservative. Alternative test statistics with more power at lower MACs can be anti-conservative, having markedly different type I error rates depending on the MAC and balance of cases and controls. Of the 782K variants passing QC in UK BiLEVE, around 57K have MAC < 20 with 11K singletons; phenotypic comparison groups have case-control ratios of either approximately 1:1 or 1:2. We compare inflation of test statistics, number of associated loci detected and computational efficiency of the score and Firth tests for association testing of rare variants in balanced and unbalanced case-control comparisons in UK BiLEVE. This research has been conducted using the UK Biobank Resource.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Case-Control Studies, Population Genetics, Sample Size and Power*

P 44

SNP CHARACTERISTICS PREDICT REPLICATION SUCCESS IN ASSOCIATION STUDIES

Ivan P Gorlov[1], Jason H. Moore[1], Olga Y Gorlova[1], Christipher I Amos, The Geisel School of Medicine, Dartmouth College
[1]The Geisel School of Medicine, Dartmouth College

The only way to distinguish true from false discoveries derived from Genome Wide Association Studies (GWAS) is replication. An independent replication of a SNP/disease association suggests that the association is real. Selecting SNPs for replication stage is based on p-vales from the discovery stage. Reproducibility of the top finding from discovery phase is low making identification of predictors of SNP reproducibility is important. We used disease-associated SNPs from more than 2,000 published GWASs to develop a model of SNP reproducibility. Reproducibility was defined as a proportion of successful replications among all replication attempts. The study reporting SNP/disease association for the first time was considered to be discovery and all consequent GWASs targeting the same phenotype replications. We found that -Log(P), where P is a p-value from the discovery study, was the strongest predictor of the SNP reproducibility. Other significant predictors include type of the SNP (e.g. missense vs intronic SNPs), minor allele frequency and eQTL status of the SNP. Features of the genes linked to the GWAS-detected SNP were also associated with the SNP reproducibility. Based on empirically defined rules, we developed a simplified reproducibility score (RS) model to predict SNP reproducibility. Both -Log(P) and RS independently predicted SNP reproducibility in a multiple regression analysis. We used data from 2 lung cancer GWAS studies as well as recently reported disease-associated SNPs to validate the model. Minus Log(P) outperforms RS when very top SNPs are selected, while RS works better with relaxed selection criteria. In conclusion, we developed an empirical model for prediction of the SNP reproducibility. The model can be used for selection SNPs for validation as well as for SNP prioritizing to be causal.

Categories: *Association: Genome-wide, Bioinformatics, Cancer*

P 45

Data-Driven Weighted Encoding: A Novel Approach to Biallelic Marker Encoding for Epistatic Models

John R Wallace[1], Molly A Hall[1], Shefali S Verma[1], Kristel van Steen[2], Elena S Gusareva[2], Jason H Moore[3], Brendan J Keating[4], Catherine A McCarthy[5], Sarah A Pendergrass[1], Marylyn D Ritchie[1]
[1]Center for Systems Genomics, The Pennsylvania State University, State College, PA
[2]Department of Electrical Engineering and Computer Science, University of LiÃ¨ge, LiÃ¨ge, Belgium
[3]Department of Genetics, Geisel School of Medicine at Dartmouth College, Lebanon, NH
[4]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA
[5]Essentia Institute of Rural Health, Duluth, MN

With Genome Wide Association Studies (GWAS), biallelic markers are typically encoded using an additive model, assigning values by the number of minor alleles each individual possesses. In detecting main effects, this encoding has been shown to be an adequate compromise; however, choosing one encoding makes an assumption about the biological action of every marker in the dataset, which can introduce artifacts. This is particularly an issue when interaction terms are added, as these artifacts can lead to spurious results. An alternative is the use of codominant encoding, which makes no assumption about the biological action of a marker, but the number of degrees of freedom required can dramatically reduce the power and introduce colinearity, particularly for interaction models.

To address these challenges, we have developed a novel and effective approach for encoding that is entirely data driven and requires no assumptions about the biological action of any particular marker, called “Data-Driven Weighted Encoding” (DaDWE) Using two real-world datasets: body-mass index data from 15,737 individuals across five different diverse cohorts and age-related cataract data from 3,377 samples (2,192 cases; 1,185 controls) from the Marshfield Clinic, we show that the choice of encoding can have a large impact. For a model with only main effects, we show that our method has identical results compared to codominant encoding, and when interaction terms are introduced, we show DaDWE has a distinct advantage due to reduced degrees of freedom. Further, using simulation data, we show that DaDWE is robust to multiple types of biological actions underlying potential predictive models, and is an appropriate choice for epistatic model discovery.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Case-Control Studies, Epigenetic Data, Epigenetics, Gene - Gene Interaction*

P 46

A One-Degree-of-Freedom Test for Supra-Multiplicativity of SNP Effects

Christine Herold[1], Vitalia Schüller[1], Alfredo Ramirez[2], Tatsiana Vaitsiakhovich[3], Tim Becker[1]
[1]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[2]Department of Psychiatry and Psychotherapy, University of Bonn, Bonn, Germany; Institute of Human Genetics, University of Bonn, Bonn, Germany
[3]Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany

Deviation from multiplicativity of genetic risk factors is biologically plausible and might explain why Genome-wide association studies (GWAS) so far could unravel only a portion of disease heritability. Still, evidence for SNP-SNP epistasis has rarely been reported, suggesting that 2-SNP models are overly simplistic. In this context, it was recently proposed that the genetic architecture of complex diseases could follow limiting pathway models. These models are defined by a critical risk allele load and imply multiple high-dimensional interactions. Here, we present a computationally efficient one-degree-of-freedom "supra-multiplicativity-test" (SMT) for SNP sets of size 2 to 500 that is designed to detect risk alleles whose joint effect is fortified when they occur together in the same individual. Via a simulation study we show that our original SMT is powerful in the presence of threshold models, even when only about 30–45% of the model SNPs are available. We can also demonstrate that the SMT outperforms standard interaction analysis under recessive models involving just a few SNPs. Nevertheless, in a second step we try to modify the indicator function to limit the multiple testing issue and improve power. In addition, we apply our test to 10 consensus Alzheimer's disease (AD) susceptibility SNPs that were previously identified by GWAS.

Categories: *Association: Genome-wide, Gene - Gene Interaction*

P 47

Fine-mapping eGFR susceptibility loci through trans-ethnic meta-analysis

Anubha Mahajan[1], Jeffrey Haessler[2], Nora Franceschini[3], Andrew Morris[4]
[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[2]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
[3]University of North Carolina, Chapel Hill, NC, USA
[4]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; Department of Biostatistics, University of Liverpool, Liverpool, UK; Estonian Genome Center, University of Tartu, Tartu, Estonia

Reduced estimated glomerular filtration rate (eGFR), is used to define chronic kidney disease (CKD). We performed trans-ethnic meta-analysis to fine-map known eGFR loci by leveraging differences in distribution of linkage disequilibrium between diverse populations. We considered six genome-wide association studies (GWAS) comprising of 23,568 individuals of European, African American, and Hispanic ancestry, each supplemented by imputation up to the 1000 Genomes Project reference panel (March 2012 release). Within each study, association with eGFR (MDRD equation) was tested under an additive model. We then combined association summary statistics across studies with MANTRA, 500kb up and down of the lead SNP at known eGFR loci, and constructed "credible sets" of SNPs that encompass 99% of the posterior probability of being causal. We resolved fine-mapping of potential causal variants to less than 20 variants at three loci: GCKR (3 SNPs, 144.5kb), UMOD/PDILT (4 SNPs, 39.3kb), and SHROOM3 (19 SNPs, 74kb). At GCKR, the credible set covers three SNPs including GCKR P446L, which is predicted to be the functional variant at this locus. Variants in the 99% credible set for SHROOM3, include intronic variants in the gene and overlap regulatory elements from ENCODE, thereby highlighting a potential mechanism for the action of this locus on eGFR. These findings provide evidence that trans-ethnic GWAS can be used to fine-map potentially causal variants at complex traits loci that can be taken forward for experimental validation and could help to further our understanding of the biological mechanisms underlying disease.

Categories: *Association: Genome-wide, Fine Mapping*

P 48

Are lipid risk alleles identified in genome-wide association studies ready for translation to clinical studies?

Alexander M Kulminski[1], Irina Culminskaya[1], Konstantin G Arbeev[1], Liubov S Arbeeva[1], Svetlana V Ukraintseva[1], Eric Stallard[1], Anatoli I Yashin[1]
[1]Duke University

Insights into genetic origin of diseases and related traits could substantially impact strategies for improving human health. The results of genome-wide association studies (GWAS) are often positioned as discoveries of unconditional risk alleles of complex health traits. We re-analyzed the associations of SNPs discovered as correlates of total cholesterol (TC) in a large-scale GWAS meta-analysis. We focused on three generations of 9,167 participants of the Framingham Heart Study (FHS) which was a part of that meta-analysis. We showed that none of SNPs available in the FHS has unconditional risk alleles for TC. Instead, the effects of these SNPs were clustered in different FHS generations in sex-specific or sex-unspecific fashion. Sensitivity of the effects to generations implies the role of the environment and/or the age-related processes. A striking result was predominant clustering of significant associations with the strongest effects in the youngest 3rd Generation cohort. This clustering was not explained by the sample size or procedure-therapeutic issues. The effect clustering in specific population groups may strongly affect sample sizes needed to detect genome-wide significance. As an example, the effect size for rs1800562 in the 3rd Generation cohort required as little as about 13,000 subjects to achieve genome significance whereas that in comparable sample of the FHS original and offspring cohorts required more than 106 subjects. The results on clustering of the effects of lipid risk alleles are in line with experimental evidence at phenotypic levels from prior studies. Our results suggest that standard GWAS strategies need to be greatly expanded to efficiently translate genetic discoveries into clinical studies.

Categories: *Association: Genome-wide*

P 49

Genome-wide meta-analysis of smoking-dependent genetic effects on obesity traits: the GIANT (Genetic Investigation of ANthropometric Traits) Consortium

Anne E Justice[1], Thomas W Winkler[2], Kristin L Young[1], Jacek Czajkowski[3], Nancy Heard-Costa[4,5], Mariaelisa Graff[1], Xuan Deng[6], Virginia Fisher[6], Tuomas *Kilpeläinen*[7], L Adrienne Cupples[4,6]

[1]University of North Carolina at Chapel Hill, Department of Epidemiology, Chapel Hill, NC, USA

[2]Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

[3]Department of Genetics Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA

[4]NHLBI Framingham Heart Study, Framingham, MA , USA

[5]Boston University, School of Medicine, Boston, MA, USA

[6]Department of Biostatistics, Boston University School of Public Health, Boston University, Boston, MA, USA

[7]The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Obesity and cigarette smoking (SMK) are important risk factors for cardiovascular disease. Yet, smokers often exhibit lower body mass index (BMI) and higher waist circumference (WC), and smoking cessation leads to weight gain. Genome-wide association (GWA) studies have identified loci that are associated with risk of overall and central obesity; yet little is known about how SMK influences genetic susceptibility to obesity. This study aims to identify loci associated with obesity measured by BMI, WC adjusted for BMI (WCa), and waist to hip ratio adjusted for BMI (WHRa), and the influence of SMK on those genetic associations. We analyzed study specific association results from 88 studies including up to 210,153 subjects with GWA or Metabochip data. Each study employed two association models: 1) SNP effects adjusted for SMK (βadj), 2) SNP effects stratified by SMK. Study specific results were combined by inverse-variance weighted fixed-effects meta-analyses. To detect SMK-dependent genetic effects on obesity, the SMK-stratified meta-analysis results were used to calculate (i) the difference in SNP associations between current and non-smokers (βdiff), and (ii) the joint estimates (βj) of the main effect and βdiff. We found genome-wide significant (GWS) ($p<5E-8$) evidence for non-zero βdiff for two loci associated with WCa, three with WHRa and two with BMI. A total of 81 loci for WCa (14 are novel), 68 loci for BMI (10 are novel) and 50 loci for WHRa (nine are novel) reached GWS for βj and/or βadj. Our results highlight the importance of appropriately modeling genetic associations by considering known biological relationships between phenotypes and environment.

Categories: *Association: Genome-wide, Cardiovascular Disease and Hypertension, Gene - Environment Interaction*

P 50

A Binomial Regression Model for Association Mapping of Multivariate Phenotypes

Saurabh Ghosh [1], Arunabha Majumdar[1]
[1]INDIAN STATISTICAL INSTITUTE, KOLKATA, INDIA

Most clinical end-point traits are governed by a set of quantitative and qualitative precursors and hence, it may be a prudent strategy to analyze a multivariate phenotype vector comprising these precursor variables for association mapping of the end-point trait. The major statistical challenge in the analyses of multivariate phenotypes lies in the modelling of the vector of phenotypes, particularly in the presence of both quantitative and binary precursors. Likelihood based approaches such as variance components as well as data reduction techniques such as principal components become infeasible or biologically difficult to interpret if some of the components of the phenotype vector are qualitative in nature. We propose a Binomial regression approach that models the likelihood of the number of minor alleles at a SNP conditional on the vector of multivariate phenotype using a logistic link function. This framework allows for the integration of quantitative as well as binary phenotypes and does not require any distributional assumptions on the phenotype vector. The test for association is based on all the regression coefficients corresponding to the constituent phenotypes. The method can be easily adopted for analyzing longitudinal data. We carry out extensive simulations under a wide spectrum of genetic models of a multivariate phenotype vector and show that the proposed test is more powerful compared to analyzing a reduced phenotype based on the first principal component of the constituent phenotypes as well as separate univariate analyses of the different phenotypes. We apply our method to analyze a multivariate phenotype comprising homocysteine levels, Vitamin B12 levels and folate levels in a study on coronary artery disease.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Multivariate Phenotypes, Quantitative Trait Analysis*

P 51

How to include chromosome X in your genome-wide association study

Christina Loley[1], Inke R König[1,2], Jeanette Erdmann[2,3], Andreas Ziegler[1,2,4]
[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany
[2]DZHK (German Centre for Cardiovascular Research), Lübeck, Germany
[3]Institut für Integrative und Experimentelle Genomik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany
[4]Zentrum für Klinische Studien, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

In current genome-wide association studies (GWAS), the analysis is usually focused on autosomal variants only, and the sex chromosomes are often neglected. Recently, a number of technical hurdles have been described that add to a reluctance of including chromosome X in a GWAS, including complications in genotype calling, imputation, and selection of test statistics. To overcome this, we provide a "how to" guide for analyzing X chromosomal data within a standard GWAS. Following a general pipeline for GWAS, we highlight the steps in which the X chromosome requires specific attention, and we give tentative advice for each of these. Through this, we show that by selection of sensible algorithms and parameter settings, the inclusion of chromosome X in GWAS is manageable. Closing this gap is expected to further elucidate the genetic background of complex diseases, especially of those with sex-specific features.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls*

P 52

Exome chip meta-analysis to identify rare coding variants associated with pulse pressure

James P Cook[1], Evelin Mihailov[2], Nicholas GD Masca[3], Fotios Drenos[4], Helen Warren[5], Martin D Tobin[1], Louise V Wain[1], Patricia B Munroe[5], ExomeBP Consortium
[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom
[2]Estonian Genome Center, University of Tartu, Tartu, Estonia
[3]Cardiovascular Biomedical Research Unit, University of Leicester, Leicester, United Kingdom
[4]Centre for Cardiovascular Genetics, University College London, London, United Kingdom
[5]William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom

Pulse pressure (PP) is a measure of arterial stiffness (calculated as the difference between systolic and diastolic blood pressure (BP)) which is a strong risk factor for cardiovascular disease and stroke. Large European genome-wide association studies have already identified multiple common variants associated with PP, however common variants do not explain all of the heritability of BP traits. It has been hypothesised that some of the remaining heritability is explained by rare variants. The exome chip was designed to act as an intermediate step between cost-effective whole genome SNP arrays, which predominantly measure common variation, and exome re-sequencing approaches, which measure rare coding variation. The array includes ~250,000 mainly low frequency exonic variants. The ExomeBP consortium has been formed to analyse the exome chip for four BP traits: SBP, DBP, PP and hypertension, and comprises ~83,000 individuals from 31 different studies.We report a large scale single variant meta-analysis of PP, including >150,000 polymorphic SNPs with minor allele frequency <1%. Results demonstrate replication of known pulse pressure loci as well as identification of novel loci not previously associated with blood pressure. Gene-based analyses are also being performed. I will describe the methodological challenges in undertaking single variant and gene-based meta-analyses of exome chip data, such as distinguishing between monomorphic and missing variants across studies, the effect of transforming the phenotype and the advantages of different gene based methods, and outline our plans to boost sample size to ~400,000 through collaboration with other consortia.

Categories: *Association: Genome-wide, Cardiovascular Disease and Hypertension, Quantitative Trait Analysis*

P 53

Genome-wide search for age- and sex-dependent genetic effects for obesity traits: Methods and results from the GIANT Consortium

Thomas W Winkler[1], Mariaelisa Graff[2], Anne Justice[2], Llilda Barata[3], Mary Feitosa[3], Iris M Heid[1], Ingrid Borecki[3], Kari E North[2], Zoltán Kutalik[4], Ruth JF Loos[5]

[1]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany

[2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514, USA

[3]Department of Genetics, Washington University School of Medicine, St Louis, Missouri 63110, USA

[4]Department of Medical Genetics, University of Lausanne, 1005 Lausanne, Switzerland

[5]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

Obesity differs between men and women and changes over time. Previous genome-wide association meta-analyses (GWAMAs) revealed sexually dimorphic loci for waist-hip ratio (WHR), but little is known whether genetic effects on obesity traits change with age. We thus conducted GWAMAs stratified by age (cut-off at 50 years) and by sex, involving 110 studies (N>310,000) of European ancestry. Each study tested up to 2.8M HapMap imputed SNPs for association with BMI and WHR in four strata (men≤50, women≤50, men>50, women>50). Using the stratum-specific estimates, we tested for age-specific effects (G x AGE), sex-specific effects (G x SEX), and for age-specific effects that differ between men and women (G x AGE x SEX). Each of the three interaction tests was conducted with and without a-priori filtering for the overall association. For BMI, our analysis yielded 15 loci with significant age-difference, of which 11 showed a stronger effect in the younger group. For WHR, our analysis yielded 44 sexually dimorphic loci, of which 11 showed opposite effects and 28 showed an effect in women only. We did not identify any 3-way G x AGE x SEX effects. Analytical power computations showed that our strategy (i) was well-powered for any kind of 2-way interaction (G x AGE, G x SEX) and for the most extreme 3-way interaction (involving opposite effects across the four strata), but (ii) lacks power to find the most plausible 3-way interactions (effects that are only present or only lacking in one of the four strata). Our results underscore the importance of age- and sex-stratified analyses to further investigate the genetic underpinning for obesity traits and demonstrate that more refined methods will be needed to establish most plausible 3-way interaction effects.

Categories: *Association: Genome-wide, Gene - Environment Interaction, Heterogeneity, Homogeneity, Sample Size and Power*

P 54

Meta-analysis of gene-set analyses based on genome wide association studies

Albert Rosenberger[1], Heike Bickeböller[1], Christopher I Amos[2], Rayjean J Hung[3], Paul Brennan[4]
[1]Universitätsmedizin Göttingen, Germany
[2]Geisel School of Medicine, US
[3]Lunenfeld-Tanenbaum Research Institute, Canada
[4]International Agency for Research on Cancer, Lyon, France

Gene-set analysis (GSA) methods are used as complementing approaches to genome-wide association studies (GWAS). The single marker association estimates of a predefined set of genes are either contrasted to those of all remaining genes or to a null non-associated background. To pool p-values of several GSAs, it is important to take into account the concordance in the observed patterns of single marker association estimates. We propose an enhanced version of Fisher's inverse χ^2-method META-GSA, but weighting each study to account for imperfect correlation between patterns. We investigated the performance of META-GSA by simulating 500 GWAS with 500 cases and 500 controls at 100 SNPs. Wilcoxon's rank sum test was applied as GSA for each study. We could demonstrate that META-GSA has greater power to discover truly associated genes sets compared to simply pooling the p-values. Under the H0, i.e. if there is no difference in the true pattern between the gene set of interest and the set of remaining genes, the results of both approaches are found to be almost without correlation. Thus, we recommend not relying on p-values alone when combining the results of independent GSAs. Applying META-GSA to pool results of four case-control GWAS of lung cancer risk (Central European Study and the Toronto/SLRI Study; German Lung Cancer Study and the MDACC Study) revealed the pathway GO0015291 ("transmembrane transporter activity") as significantly enriched with associated genes (GSA-method: EASE, p=0.0315 corrected for multiple testing).

Categories: *Association: Genome-wide, Cancer, Case-Control Studies, Pathways*

P 55

Meta-analysis of correlated traits using summary statistics from GWAS

Xiaofeng Zhu[1], Tao Feng[1]
[1]Department of Epidemiology and Biostatistics, Case Western Reserve University

Genome wide association study (GWAS) has identified many genetic variants underlying complex traits. Many detected genetic loci harbor variants that associate with multiple, even distinct traits.Most current analysis approaches focus on single traits, even though the final results from multiple traits are evaluated together. Such approaches miss the opportunity to systemically integrate the phenome-wide data available for genetic association analysis. In this study, we propose a general approach that can integrate association evidence from summary statistics of multiple traits, either correlated, unrelated, continuous or binary traits, which may come from the same or different studies. We allow for trait heterogeneity effects. Population structure and cryptic relatedness can also be controlled. Our simulations suggest that the proposed method has improved statistical power over single trait analysis in the most of cases we studied. We also applied our method to a large genome-wide association study and identified multiple variants which were missed by a single trait analysis. Our method also provides a way to study a pleotropic effect.

Categories: *Association: Genome-wide, Multivariate Phenotypes*

P 56

Studying the Ethnic Differences in the Genetics of Type 2 Diabetes using the Population Specific Human Phenotype Networks

Jingya Qiu[1], Christian Darabos[1], Jason H Moore[1]
[1]Dartmouth College

GWAS led to the discovery of 200+ SNPs at 150+ loci associated with type 2 diabetes mellitus (T2DM). It was also observed that East Asians develop T2DM at a higher rate, younger age, and lower BMI than their European ancestry counterparts. The reason behind this occurrence remains elusive. We constructed human phenotype subnetworks (HPSNs) based on ethnicity-specific data to quantitatively analyze and visualize the disparities in genetic variants between different ethnic groups. Our identification of interethnic differences in the genetic variants associated with T2DM suggests the possibility of different pathways involved in the pathogenesis of T2DM amongst different populations.

With comprehensive searches through the NHGRI GWAS catalog literature, we manually curated over 2,500 ethnicity-specific SNPs associated with T2DM and 48 other related traits. The GWAS catalog usually reports the data combined over the initial and replication samples, across the different ancestries. Analysis of all-inclusive data can be misleading, as not all variants are transferable across diverse populations. The extraction of ethnicity data allowed us to construct population-specific HPSNs.

We identified 99 SNPs highly significant to T2DM, most initially discovered in Europeans and replicated in East Asians, suggesting shared biological pathways. Of the 99 SNPs, however, 21 were specific to East Asian populations but impossible to replicate in other cohorts. Furthermore, many SNPs showed significant differences in studies of comparable size. For example rs2237892 in locus KCNQ1, a critical gene in insulin-secreting INS-1 cells, proved to be highly significant in East Asian population (p-Value=2.5E-40) but not in Europeans (p=7.2E-04).

Categories: *Association: Genome-wide, Bioinformatics, Data Mining, Diabetes, Gene - Gene Interaction, Pathways, Population Genetics, Prediction Modelling*

P 57

Hierarchical Bayesian Model integrating sequencing and imputation uncertainty using MCMC method for rare variant association detection

Liang He[1], Janne Pitkäniemi[1,2], Mikko J Sillanpää[3,4], Antti P Sarin[5], Samuli Ripatti[5,6]
[1]Department of Public Health, Hjelt Institute, University of Helsinki, Finland
[2]Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland
[3]Department of Mathematical Sciences, University of Oulu, Oulu FIN-90014, Finland
[4]Department of Biology and Biocenter Oulu, University of Oulu, Oulu FIN-90014, Finland
[5]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Finland
[6]Wellcome Trust Sanger Institute, UK

Next generation sequencing has led to the studies of rare genetic variants, which are thought to explain the missing heritability for complex diseases. Most existing statistical methods for RV association detection do not account for the presence of sequencing errors and imputation uncertainty, which can largely affect the power and perturb the accuracy of association tests due to rare observations of minor alleles. Some proposed methods that assign different weights based on genotype quality leads to the reduction of observations, and thus statistical power. We develop a hierarchical Bayesian approach to powerfully estimate the association between rare variants and complex diseases and account for genotype uncertainty from both whole-genome sequencing and imputation data using MCMC method. Our integrated framework, which combines the misclassification model with shrinkage-based Bayesian variable selection, estimates the association and predicts the low-quality genotype simultaneously by borrowing the strength from priors and the rest of high-quality data, and allows for dealing with sequencing and imputation data simultaneously. Sequencing quality information or imputation uncertainty is incorporated into the integrated framework to achieve the optimal power. We test the proposed method on simulated data and demonstrate that it outperforms other existing methods under various scenarios. Then we apply our model to a Finnish low-density lipid cholesterol study, which includes both whole-genome deep sequencing and imputation genotypic data, and both well-known and novel gene regions with RVs significantly related to low density lipoprotein cholesterol level are identified.

Categories: *Association: Genome-wide, Bayesian Analysis, Genomic Variation, Markov Chain Monte Carlo Methods, Missing Data, Multilocus Analysis, Population Genetics, Quantitative Trait Analysis, Sequencing Data*

P 58

Sex-specific association of MYLIP with mortality-optimized healthy aging index

Mary F Feitosa[1], Ryan L. Minster[2], Mary K Wojczynski[1], Jason L Sanders[3], Amy M Matteini[4], Richard Mayeux[5], Nicole Schupf[6], Thomas T Perls[7], Kaare Christensen[8], Anne B Newman[3]

[1]Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, MO

[2]Department of Human Genetics, University of Pittsburgh, PA

[3]Department of Epidemiology Graduate School of Public Health, University of Pittsburgh, PA

[4]Division of Geriatric Medicine and Gerontology, School of Medicine, Johns Hopkins University, Baltimore, MD

[5]Department of Neurology, Columbia University, New York, NY

[6]Taub Institute, College of Physicians and Surgeons, Columbia University, New York, NY

[7]Section of Geriatrics, Department of Medicine, Boston University, Boston School of Medicine and Boston Medical Center, MA

[8]The Danish Aging Research Center, Epidemiology, University of Southern Denmark, and Department of Clinical Genetics and Department of Clinical Biochemistry and Pharmacology, Odense University Hospital, Odense, Denmark

Elevated low-density lipoprotein (LDL) cholesterol is associated with increased risk of coronary artery disease, cognitive decline and dementia. Although these diseases predict mortality, knowledge of the relationship between dyslipidemia and its genetic contributors to mortality is limited. A mortality-optimized healthy aging index (HAI-M) demonstrated accuracy to predict mortality. We hypothesized that SNPs from the GLGC Consortium (N=30) associated with LDL contribute to HAI-M variability in 3,534 subjects from the Long Life Family Study. To create HAI, systolic blood pressure, pulmonary vital capacity, creatinine, fasting glucose, and modified-mini-mental-status-examination-score, were scored as 0 (healthiest tertile), 1 (middle tertile), or 2 (unhealthiest tertile, and clinical cutoffs for glucose), and the sum produced an index ranging from 0 (healthiest) to 10 (unhealthiest). The HAI-M was generated by applying regression coefficients from Cox proportional hazards models for death from the Cardiovascular Health Study to each component of the HAI. MYLIP-rs3757354 (p=0.0001, beta=-0.16±0.04) and APOH-rs1801689 (p=0.03, beta=0.23±0.10) were associated with HAI-M in a stepwise regression model. Accounting for family structure using a mixed model, MYLIP-rs3757354 was significantly associated with HAI-M (p=0.001, beta=-0.14±0.04). There were sex-specific effects. MYLIP-rs3757354 was significantly associated with HAI-M in men (p=0.0003, beta=-0.24±0.06), but not in women (p=0.23, beta=-0.07±0.06). MYLIP (6p23-p22) encodes the E3 ubiquitin ligase myosin regulatory light chain-interacting protein and promotes degradation of the LDLR, a process that may be relevant to healthy aging.

Categories: *Association: Genome-wide, Genomic Variation, Multiple Marker Disequilibrium Analysis*

P 59

Genetic determinants of liver function and their relationship to cardio-metabolic health

Niletthi De Silva[1], Debbie Lawlor[1], Thomas Gaunt[1], Abigail Fraser[1]
[1]University of Bristol

Introduction: Genome-wide association studies have identified several common variants robustly associated with liver function tests, primarily ALT, AST, ALP, GGT, Bilirubin and Albumin. These phenotypes have been used as markers of liver damage, and there is evidence from observational studies that these are related to future adverse cardiometabolic health. However, it is unclear to what extent these associations are causal or confounded (in particular by alcohol consumption and general greater adiposity). Aims: To examine the association of metabochip variants with ALP, ALT, AST, GGT, Bilirubin and Albumin to determine whether these replicate published genome-wide association (GWAS) findings and to identify any new variants robustly associated with these traits. To use Mendelian randomization study to test whether ALT, AST, ALP, GGT, Bilirubin and Albumin (markers of liver damage) causally influence CHD, stroke, type 2 diabetes and related continuos outcomes - fasting glucose, fasting insulin, LDL, HDL, triglycerides, total cholesterol, SBP and DBP. Methods: We carried out metabochip-wide meta-analyses of ALT, AST, ALP, GGT, Bilirubin and Albumin to identify any novel variants associated with these traits. We then tested multiple common variants robustly associated with ALT, AST, ALP GGT, Bilirubin and Albumin (3, 2, 11, 17, 5, 5 SNPs respectively) against incident and prevalent diabetes, CHD, stroke events, and the related continuous outcomes in 5437 individuals from four prospective cohorts under the UCLEB consortium. Results: We replicated several previously established loci robustly associated with ALT, AST, ALP, GGT, Bilirubin and Albumin. In addition we identified two novel loci associated with ALP and AST in the ABO and PNPLA3 locus respectively at $p<5x10-8$. We now aim to replicate these two novel loci in an independent data set from the discovery cohort. In multivarable analyses adjusted for several potential confounders (i.e: smoking status, social class, alcohol, BMI and waist circumference) we replicated several observational associations reported previously. Indviduals carrying greater number of ALT, AST, ALP, GGT, Bilirubin and Albumin raising allles had increased levels of ALT, AST, ALP, GGT, Bilirubin and Albumin ($p<0.001$). There was evidence from instrumental variables analyses that ALT, AST GGT and Albumin causally reduce the risk of stroke: OR per log10 increase in ALT, AST, GGT was 0.04 [95%CI: 0.01, 0.11), 0.00 [95%CI: 00, 0.03], 0.21 [95%CI:0.10, 0.44] respectively and OR per one mg/dl increase in albumin was 0.45 [95%CI:0.35,0.58]. Conclusion: Markers of liver damage in particualr ALT, AST GGT and Albumin may causally influence the risk of stroke.

Categories: *Association: Genome-wide, Cardiovascular Disease and Hypertension, Causation, Mendelian Randomisation*

P 60

Variable selection method for complex genetic effect models using Random Forests

Emily R Holzinger[1], Silke Szymczak[1], James Malley[2], Joan E Bailey-Wilson[1]
[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health
[2]Center for Information Technology, National Institutes of Health

Standard analysis methods for genome wide association studies (GWAS) are not robust to complex disease models, such as interactions between variables with small main effects. These types of effects could, in part, contribute to the heritability of complex human traits. Machine learning methods that are capable of identifying interactions, such as Random Forests (RF), are an alternative analysis approach. One caveat to RF is that there is no clear way to distinguish between probable true hits and noise variables based on the importance metric calculated. To this end, we have developed a novel variable selection method for RF that has three components: 1. A permutation procedure to calculate the RF importance score. 2. Null variance estimation method to create more meaningful thresholds for variable selection. 3. Recurrency to address noise in the results due to randomness of the method. First, we simulated datasets with various genetic models, including different levels of main and interaction effects. Next, we assessed the Type I error and power of the RF method and compared it to regression based methods. We further tested the performance of the variable selection method using a biological GWAS dataset. Our simulated data findings indicate that optimizing the selection threshold can greatly reduce the number of false positives in the selected variables. However, the optimal threshold is highly dependent on the underlying simulated genetic model. The recurrency aspect of the method assists in selecting the appropriate threshold. Additionally, the power to identify main effects is comparable to linear regression analyses with the correct main effect terms explicitly modeled. In the biological dataset, our method identifies a similar set of SNPs as linear regression. Future directions will involve testing and comparing methods for modeling the selected variables in a more interpretable fashion.

Categories: *Association: Genome-wide, Bioinformatics, Data Mining, Gene - Gene Interaction, Machine Learning Tools*

P 61

Identification of shared genetic aetiology between epidemiologically linked disorders with an application to obesity and osteoarthritis

Jennifer L Asimit[1], Kalliope Panoutsopoulou[1], Eleanor Wheeler[1], Sonja Berndt[2], the GIANT consortium, the arcOGEN consortium, Andrew P Morris[3,4], Inés Barroso[1], Eleftheria Zeggini[1]
[1]Wellcome Trust Sanger Institute
[2]National Cancer Institute, US National Institutes of Health
[3]Wellcome Trust Centre for Human Genetics, University of Oxford
[4]Department of Biostatistics, University of Liverpool

A common approach to a genetic overlap analysis of two traits involves comparing p-values from the genome-wide association study (GWAS) of each trait. However, p-values do not account for differences in power, whereas Bayes' factors do, and may be approximated using summary statistics. We use simulation studies to compare the power of frequentist and Bayesian approaches to overlap analyses, and to decide on thresholds for comparison between the two methods. It is empirically illustrated in single-disease associations that BFs have a decreasing proportion of false positives (PFP) as study size increases. For a log10(BF) threshold Lq of 1.69 (R=type II error cost/type I error cost=2, p0 = Pr(no association at SNP)=0.99), the PFP decreases from 7.38×10-4 (N=2,000 each cases/controls) to 3.37×10-4 (N=20,000), while for p-values the PFP fluctuates near the p-value threshold a regardless of study size. In a preliminary overlap analysis of obesity (GIANT consortium) with OA (arcOGEN consortium), the number of signals is similar at comparable threshold levels between BFs and p-values, though not always overlapping. For Lq=0.91 (R=type II error cost/type I error cost=12, p0=Pr(no association at variant)=0.99), there are 18 identified shared variants, and the comparable a levels of 0.003 and 0.004 result in 15 and 28 hits, respectively. The most notable difference is that the Bayesian list contains rs13107325 (in SLC39A8/ZIP8), a variant previously associated with obesity-related phenotypes such as BMI and blood pressure, and animal studies have shown that the zinc-ZIP8-MTF1 axis regulates OA pathogenesis. We are pursuing replication of this finding.

Categories: *Association: Genome-wide*

P 62

Investigation of genetic risk factors of very low birth weight infants within the German Neonatal Network

Michael Preuß [1,3], Andreas Ziegler[1,2], Egbert Herting[3], Wolfgang Göpel[3]
[1]Institute of Medical Biometry and Statistics, University at Lübeck, University Hospital Schleswig-Holstein - Campus Lübeck, Germany;
[2]Center for Clinical Trials, University at Lübeck, Lübeck, Germany
[3]Department of Pediatrics, University at Lübeck, Lübeck, Germany

Very Low Birth Weight (VLBW) infants have substantially increased mortality and morbidity rates, but the factors influencing long-term development are not well understood. The German Neonatal Network (GNN) was founded in 2009 to identify genetic, clinical and social factors influencing etiology and long-term development of VLBW. Clinical information includes oxygen demand, administration of surfactant, catecholamine, steroid hormones and bronchopulmonary dysplasia (BPD), brain haemorrhage (IVH), sepsis and death among others. The cohort size is 20,000, and the recruitment includes more than one quarter of all German VLBW per year. DNA samples from more than 9000 VLBW as well as buccal swabs from mothers have been collected from a total of 54 participating German hospitals. Approximately 2600 VLBW from GNN were genotyped on the Axiom™ Genome-Wide CEU 1 Array, and replication was performed in another 4400 GNN VLBW. Results of the initial genome-wide association study revealed genome-wide significance ($p < 5E-08$) for several traits. An interesting finding is for the use of surfactant during hospital stay with an association to LINGO2 (lead SNP rs4878404, initial $p = 5E-06$, replication one-sided $p = 2.3E-03$). These results demonstrate that GNN is a unique resource for genetic and pharmacogenetic studies in VLBW.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls*

P 63

Artificial intelligence analysis of epistasis in a genome-wide association study of glaucoma

Jason H Moore[1], Casey S Greene[1], Doug Hill[1]
[1]Dartmouth College

The genetic basis of primary open-angle glaucoma (POAG) is not yet understood but is likely the result of many interacting genetic variants that influence risk in the context of our local ecology. We introduce here the Exploratory Modeling for Extracting Relationships using Genetic and Evolutionary Navigation Techniques (EMERGENT) algorithm as an artificial intelligence approach to the genetic analysis of common human diseases. EMERGENT builds models of genetic variation from lists of mathematical functions using a form of genetic programming called computational evolution. A key feature of the system is the ability to utilize pre-processed expert knowledge giving it the ability to explore model space much as a human would. We describe this system in detail and then apply it to the genetic analysis of POAG in the Glaucoma Gene Environment Initiative (GLAUGEN) study that included approximately 1272 cases and 1057 controls. A total of 657,366 single-nucleotide polymorphisms (SNPs) from across the human genome were measured in these subjects. Analysis using the EMERGENT framework revealed a best model consisting of six SNPs that map to at least six different genes. Two of these genes have previously been associated with POAG in several studies. The others represent new hypotheses about the genetic basis of POAG. All of the SNPs are involved in non-additive gene-gene interactions. Further, the six genes are all directly or indirectly related through biological interactions to the vascular endothelial growth factor (VEGF) gene that is an actively investigated drug target for POAG. This study demonstrates the routine application of an artificial intelligence-based system for the genetic analysis of complex human diseases.

Categories: *Association: Genome-wide, Bioinformatics, Data Mining, Gene - Gene Interaction, Machine Learning Tools*

P 64

Mutations causing complex disease may under certain circumstances be protective in an epidemiological sense

Sabine Siegert[1], Andreas Wolf[2], David N Cooper[3], Michael Krawczak[2], Michael Nothnagel[1]
[1]Cologne Center for Genomics, University of Cologne, Cologne, Germany
[2]Institute of Medical Informatics and Statistics, Christian-Albrechts University, Kiel, Germany
[3]Institute of Medical Genetics, Cardiff University, Cardiff, United Kingdom

Guided by the practice of classical epidemiology, research into the genetic basis of complex disease usually takes for granted the dictum that causative mutations are invariably over-represented among affected as compared to unaffected individuals. However, employing various models of population history and penetrance, we show that this supposition is not true and that a mutation involved in the etiology of a complex disease can under certain circumstances be depleted rather than enriched in the affected portion of the population. Such mutations are 'protective' in an epidemiological sense and would often tend to be erroneously excluded from further studies. Our apparently paradoxical finding is due to the possibility of a negative correlation between complementary causative mutations that may arise as a consequence of the specifics of the population genealogy. This phenomenon also has the potential to hamper efforts to identify rare causative mutations through whole-genome sequencing.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls*

P 65

Genome-wide Association Study Identifies SNP rs17180299 and Multiple Haplotypes on CYP2B6, SPON1 and GSG1L Associated with Plasma Concentrations of the Methadone R- and S-enantiomer in Heroin-dependent Patients under Methadone Maintenance Treatment

Hsin-Chou Yang[1,2,3], Shih-Kai Chu[1,2,4], Sheng-Chang Wang[5], Sheng-Wen Liu[5], Ing-Kang Ho[5], Hsiang-Wei Kuo[5], Yu-Li Liu[5,6]

[1]Institute of Statistical Science, Academia Sinica

[2]Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica

[3]School of Public Health, National Defense Medical Center

[4]Institute of Biomedical Informatics, National Yang-Ming University

[5]Center for the Neuropsychiatric Research, National Health Research Institutes

[6]Department of Psychiatry, National Taiwan University Hospital and National Taiwan University College of Medicine

Although methadone metabolic pathway has been partially revealed there is still no report regarding genome-wide association studies to characterize genetic mechanisms of the plasma concentrations of methadone R- and S-enantiomer. We conducted the first genome-wide association study to identify genes associated with the plasma concentrations of methadone R- and S-enantiomer and their metabolites in a methadone maintenance cohort. We made a series of rigorous examinations in data quality control to remove poor samples and SNPs. We carried out genome-wide single-locus and haplotype-based association tests for four quantitative traits, the plasma concentrations of methadone R- and S-enantiomer and their metabolites, of 344 heroin-dependent patients who were treated with methadone maintenance treatment in the Han Chinese population of Taiwan. We identified a significant SNP rs17180299 ($p = 2.24\times10\text{-}8$) which can explain 9.541% of the variation of the plasma concentration of methadone R-enantiomer. We also identified 17 haplotypes on SPON1, GSG1L, and CYP450 genes associated with the plasma concentration of methadone S-enantiomer. They can explain about one-fourth of variation of the plasma concentration of S-methadone as a whole, where two significant haplotypes on CYP2B6 already explained 10.72% of the variation. In conclusion, we identified important SNP and haplotypes which contribute to genetic variation of plasma concentration.The results shed light on the genetic mechanism concerned with the metabolism of methadone maintenance treatment in heroin-dependent patients. Moreover, the results are also potentially applicable to prediction of methadone dose and methadone-related death.

Categories: *Association: Genome-wide, Haplotype Analysis, Quantitative Trait Analysis*

P 66

A nonparametric regression approach to the analysis of genomewide association studies

Pianpool Kirdwichai[1], M Fazil Baksh[1]
[1]University of Reading

Recently there has been a move towards development of regression inspired methods for analysis of genomewide association studies of complex diseases. This is because multiple testing methods, such as Bonferroni correction, tend to impose stringent significance thresholds and consequently, unless the study is very large, can reliably identify only those genomic regions with very strong association signals. However many complex diseases are suspected result from the cumulative action of many loci each having a small effect, there is a high probability the association signals in such studies will in fact be moderate and extremely strong signals will be very rare. Although methods with higher power than the Bonferroni correction have been proposed, these tend to produce more false positive findings. This challenging problem of methodology that is more efficient than existing approaches but with false positive findings comparable with Bonferroni is addressed in this talk. A novel method based on nonparametric regression, capable of reliably identifying candidate regions of disease-gene association in GWAS is developed and evaluated. The method is model-free and establishes significance thresholds that inherently account for the LD structure in the data through a tuning parameter and assigned weights. A theoretically supported, computationally efficient method for obtaining the optimal tuning parameter is proposed and evaluated. Results of extensive evaluations and comparisons with existing methods show that the proposed approach is not only powerful but also lead to substantial reduction in false positive findings. The method is illustrated using data from the Wellcome Trust Case Control Consortium study of Crohn's disease.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls*

P 67

Genetic insights into primary biliary cirrhosis – an international collaborative meta-analysis and replication study

Heather J Cordell[1], George Mells[2], Gideon M Hirschfield[3], Canadian/US/Italian/UK PBC Consortia, Carl Anderson[4], Mike Seldin[5], Richard Sandford[2], Katherine A Siminovitch[6]
[1]Newcastle University
[2]University of Cambridge
[3]University of Birmingham
[4]Wellcome Trust Sanger Institute
[5]UC Davis
[6]University of Toronto

Previous genome-wide association studies (GWAS) of primary biliary cirrhosis (PBC) have confirmed associations at the human leukocyte antigen (HLA)-region and identified 27 non-HLA susceptibility loci. We undertook genome-wide imputation and meta-analysis of discovery datasets from the North American, the Italian and the UK GWAS of PBC, with a combined, post-QC sample size of 2745 cases and 9802 controls. Following meta-analysis, index single nucleotide polymorphisms (SNPs) at selected loci with PGWMA<2×10-5 were genotyped in a validation cohort consisting of 3716 cases and 4261 controls. To prioritise candidate variants and genes at confirmed risk loci, we used the ENCODE and the 1000Genomes datasets to identify SNPs within regulatory elements and non-synonymous SNPs in LD with the index variant (r2>0.8). We identified seven previously unknown risk loci for PBC. Functional annotation of these loci revealed SNPs within regulatory elements that are predicted to affect expression of DGKQ (4p16), PAM (5q14) and IL21R (16p12), that are strongly-correlated to the index variant. Other candidate genes include IL12B (5q31), which forms part of the IL-12 signalling cascade, and CCL20 (2q36), which is involved in chemo-attraction of lymphocytes and dendritic cells towards epithelia and is expressed by TH17 cells originating from Foxp3+ T cells. Pathway analysis identified several highly plausible gene sets associated with PBC, including the IL-12 and JAK-STAT signalling pathways, and implicated several other immune processes in the pathogenesis of PBC, including innate immune processes (e.g. IFN-α,β signaling).

Categories: *Association: Genome-wide*

P 68

Genes Associated with Lung Cancer, Chronic Obstructive Pulmonary Disease, or Both

Jun She[1,2], Bo Deng[1,3], Jie Na[1], Julie M Cunningham[1], Zhifu Sun[1], Jason A Wampfler[1], Tanya M Petterson[1], Paul D Scanlon[1], Shuo Zhang[1,4], Christine Wendt[5]

[1]Mayo Clinic, MN, U.S.A.

[2]Zhongshan Hospital, Fudan University, Shanghai, People's Republic of China

[3] Institute of Surgery Research, Daping Hospital, Third Military Medical University, Chongqing, People's Republic of China

[4] Tulane University, New Orleans, LA

[5]University of Minnesota and Veterans Administration Medical Center, Minneapolis, MN, U.S.A.

Background Genetic contribution to lung cancer (LC) or chronic obstructive pulmonary disease (COPD) remains unclear; COPD is considered an important LC precursor independent of tobacco smoke exposure. Over 300 candidate genes have been associated with COPD and/or LC. We conducted a comprehensive validation study to tease apart these candidate genes using genome-wide single nucleotide polymorphism based analysis (SNP-GWA) in a Caucasian population. Methods We tested 4491 SNPs in 304 candidate genes after redundancy analysis of linkage disequilibrium. The SNP-GWA data that tested the association of these genes with LC and/or COPD consisted of 2484 subjects including LC only (n=612), LC and COPD (573), COPD only (537), and controls (762). The biological roles were elucidated by transcript expression quantitative trait loci (eQTL), differential mRNA expression between tumor and normal lung, and pathway analyses, along with allele-specific risks, assessed by odds ratio (OR) and 95% confidence interval (CI). Results We validated 11 SNPs of 8 candidate genes (G1-G8): 4 for LC with COPD (G1-G4), 4 for LC from COPD (G1,2,5,6), 2 for COPD only (G1,7) and 1 for LC only (G8). A SNP in G1 was inversely associated with COPD without LC (OR=0.47; 95% CI, 0.31-0.72) or with LC (OR=0.40; 0.27-0.60), supported by eQTL of SNP-alleletypes with mRNA levels in germline tissues (P=0.02) and differential expression of G1 (P<10-5). A SNP in G2 was inversely associated with LC that developed from COPD patients (OR=1.65; 1.17-1.78), with significant difference of G2 transcript levels in tumor and normal lung tissues (P=0.01). Conclusion We found 2 genes to be strongly associated with the risk of COPD and/or LC, indicating potential targets to intervene COPD and LC.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Cancer, Genomic Variation, Multifactorial Diseases, Pathways, Quantitative Trait Analysis*

P 69

A general approach for combining diverse rare variant association tests provides improved power across a wider range of genetic architecture

Nathan L Tintle[1], Brian Greco[2], Allison Hainline[3], Jaron Arbet[4], Kelsey Grinde[5], Alejandra Benitez[6]
[1]Dordt College
[2]University of Michigan
[3]Vanderbilt University
[4]Winona State University
[5]St. Olaf College
[6]Brown University

In the wake of the widespread availability of genome sequencing data made possible by way of next-generation technologies, a flood of gene-based rare variant tests have been proposed. Most methods claim superior power against particular genetic architectures. However, an important practical issue remains for the applied researcher—namely, which test should be used for a particular association study which may consider multiple genes and/or multiple phenotypes. Recently, tests have been proposed which combine individual tests to minimize power loss while improving the robustness to a wide range of genetic architectures. In our analysis, we propose an expansion of these approaches, by providing a general method that works for combining an arbitrarily large number of any gene-based rare variant test—a flexibility typically not available in other combined testing methods. We provide a theoretical framework for evaluating our combined test to provide direct insights into the relationship between test-test correlation, test power and the combined test power relative to individual testing approaches and other combined testing approaches. We demonstrate that our flexible combined testing method can provide improved power and robustness against a wide range of genetic architectures. We further demonstrate the performance of our combined test on simulated genotypes, as well as on a dataset of real genotypes with simulated phenotypes. We support the increased use of flexible combined tests in practice to maximize robustness of rare-variant testing strategies against a wide-range of genetic architectures.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Case-Control Studies, Genomic Variation*

P 70

A Methodological Comparison of Epistasis Modeling of High Order Gene-Gene Interactions with Application to Genetic Profiling of PA Infection among Cystic Fibrosis Patients

Wenjiang Fu[1], Mengtian Shen[1], Shunjie Guan[1]
[1]Michigan State University

Recent studies of epistasis have been focusing on high order gene-gene interactions, including the classification and regression trees (CART)-based methods, the Mann-Whitney U-statistic methods, Bayesian epistasis association mapping (BEAM), gene-based gene-gene interaction tests, and gene-based Multifactor dimensionality reduction (MDR). These methods have been developed to identify gene-gene interactions in GWA studies of complex diseases and have been demonstrated to identify potential high order interactions. However, comparison of these methods and their computational capacity has not been fully studied. In this paper, we will compare these methods and apply them to an exome sequencing study of cystic fibrosis (CF). Although CF is a recessive Mendelian disease with a mutation in the CFTR gene, the disease manifestation is complex with the potential dysfunction of a number of organs and high mortality rate in early ages. About 80% CF patients develop pseudomonas aeruginosa (PA) infection, which leads to failure of the lung, liver, pancreas, intestine or other organs, resulting in breathing difficulty, CF associated liver diseases, diabetes, male infertility, and other disorders, and ultimate death in early age. It has been recently reported that genes (eg. DCTN4) other than the CFTR may also be associated with the PA infection among CF patients. We apply a number of methods to identify high order gene-gene interactions for genetic profiling of PA infection using exome sequencing data of a case-control study. We compare these methods in terms of the power, the profiling robustness and accuracy. We conclude that PA infection among CF patients can be profiled using a small number of genes with high accuracy.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Bioinformatics, Case-Control Studies, Gene - Gene Interaction, Sequencing Data*

P 71

eQTL and pathway analysis on expression profiles of a cattle cross

Markus O Scheinhardt[1], Bodo Brand[2], Daisy Zimmer[3], Norbert Reinsch[3], Manfred Schwerin[1,4], Andreas Ziegler[1,5]

[1]Institute of Medical Biometry and Statistics, University Lübeck, Germany

[2]Institute for Genome Biology, Leibniz Institute for Farm Animal Biology, Dummerstorf, Germany

[3]Institute for Genetics and Biometry, Leibniz Institute for Farm Animal Biology, Dummerstorf, Germany

[4]Institute for Farm Animal Research and Technology, University Rostock, Germany

[5]Center for Clinical Trials Lübeck, University Lübeck, Germany

In farm animal science, mapping of expression quantitative trait loci (eQTL) becomes increasingly important for studying molecular mechanisms of complex traits, such as milk production or carcass traits in cattle. We investigated 145 female animals from an F2 resource population derived from a cross between Charolais (beef cattle) and German Holstein (dairy cattle) founder breeds. SNP genotyping of 37204 SNP was accomplished using Illumina BovineSNP50 Beadchip, and gene expression profiles of 10069 adrenal cortex transcripts were obtained from Affymetrix GeneChip®Bovine v1 Array. The expression values were decorrelated by means of a sire-dam model at which we adjusted for relatedness, age and season year of slaughtering. Residuals were used to perform the eQTL analysis. An adaptive location test was applied to adjust for varying degrees of skewness and tail length of the gene expression distributions. A total of 1048 eQTLs were identified which were associated with the expression of 641 adrenal cortex transcripts. Ingenuity pathway analysis of transcripts differentially expressed among genotypes highlighted molecular and cellular functions related to carbohydrate and lipid metabolism to be affected by eQTLs within the F2 cross population.

Categories: *Association: Genome-wide, Gene Expression Arrays, Gene Expression Patterns, Pathways, Quantitative Trait Analysis*

P 72

Evidence for polygenic effects in two genome-wide association studies of breast cancer using genetically enriched cases

Olivia Leavy[1], Luigi Palla[1], Julian Peto[1], Douglas Easton[2], Frank Dudbridge[1]
[1]London School of Hygiene and Tropical Medicine
[2]University of Cambridge

Over recent years genome-wide association studies have proven to be successful in finding associations between genetic variants and phenotypes. However, much of the heritability remains to be explained for complex diseases. Polygenic scoring allows testing for substantive polygenic effects among the markers that are not individually significant in GWAS. This has been successfully applied to many complex diseases, but to date has not been demonstrated in breast cancer. We studied two datasets: the UK2 study and the British Breast Cancer Study (BBCS), both containing women who have at least two close relatives that have developed breast cancer. In the BBCS dataset most of the cases have bilateral breast cancer. The disease prevalence applicable to these studies therefore will be lower than the general prevalence for breast cancer. Methods given by Dudbridge (2013) can be used to estimate the genetic variance explained by the entire GWAS using information on training and replication datasets. The training and replication datasets were created by internally splitting each of the BBCS and UK2 datasets. Using different values of the prevalence for familial breast cancer, these being lower than the prevalence of breast cancer, we estimated the genetic variance explained to be between 11.5% and 47.3% for the BBCS and at least 35.5% for the UK2 study. Given the low heritability of breast cancer, these values are larger than typically seen in complex diseases and seem to reflect the stronger genetic effects present in familial cases. This is the first significant association of genome-wide polygenic scores for breast cancer and confirms the value of using genetically enriched cases in GWAS.

Categories: *Association: Genome-wide, Cancer*

P 73

Do Boundaries Matter for Tiled Regression?

Alexa JM Sorant[1], Heejong Sung[1], Tae-Hwi Schwantes-An[1], Alexander F Wilson[1]
[1]Computational and Statistical Genomics Branch, NHGRI, NIH

Current methods of analyzing today's vast quantities of genetic data include regression-based variable selection methods producing linear models incorporating the chosen predictors. One such method, Tiled Regression, begins by considering separately relatively small segments of the genome called tiles, using stepwise regression to choose a set of independent significant SNPs, if any, within each tile and then combining them for further selection at higher levels. A natural way to define tiles is to create boundaries around recombination hotspots, so that genetic variants likely to be highly correlated due to linkage disequilibrium are initially considered together. However, such grouping may not be critical to the ultimate selection of genetic components of a trait model. To study the effects of alternative boundary definitions, we used a simulated mini-GWAS genome including 306,097 SNPs in 4000 unrelated individuals, with two kinds of phenotypes generated for each. For examination of type I error we generated 2000 non-genetic traits based on a normal distribution. For examination of power, we generated 2000 traits from a simple additive model of genetic effects contributed by 7 independent SNPs with locus-specific heritabilities ranging from .0005 to .0108. We analyzed each trait with TRAP (v. 1.3) using several different tile boundary schemes, including the usual hotspot-based definition, combining sets of ten consecutive tiles into larger tiles, and a definition based on a fixed length in base pairs corresponding to the average size of the original tiles. With analyses of 400 replicates completed, we observed virtually no difference in either type I error or power resulting from the different tile boundary definitions.

Categories: *Association: Genome-wide, Multilocus Analysis, Prediction Modelling, Quantitative Trait Analysis*

P 74

METAINTER: meta-analysis tool for multiple regression models

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

The need to summarize the results of related Genome-wide association studies (GWAS) has encouraged rapid development of new meta-analytic methods and tools. Application of the fixed or random effects models to single-marker association tests is a standard practice. More complex methods involving multiple parameters have been used seldom in view of the absence of a respective meta-analysis pipeline. Meta-analysis based on combining p-values can be applied to any association test. However, in order to be powerful, meta-analysis methods for high-dimensional models should incorporate additional information such as study-specific properties of parameter estimates, their effect directions, standard errors and covariance structure. In this context, a method for the synthesis of linear regression slopes (MSRS) has been recently proposed in the educational sciences. We elaborate this method for multiple logistic regression models and introduce a software tool METAINTER, which implements MSRS for an arbitrary number of model parameters as well as three further meta-analysis methods. METAINTER provides meta-analysis p-values and common parameter estimates of multiple regression models, and can be used to test the homogeneity of studies results. The software can directly be applied to analyze the results of single-SNP tests, global haplotype tests, tests for and under gene-gene or gene-environment interaction. Via simulations for two-SNP models we have shown that MSRS has correct type I error and its power comes very close to that of the joint analysis of the entire sample. We support the results by a real data analysis of six GWAS of type 2 Diabetes.

Categories: *Association: Genome-wide, Case-Control Studies, Data Integration, Gene - Gene Interaction*

P 75

Successful replication of GWAS hits for multiple sclerosis in 10,000 Germans using the exome array

Theresa Holste[1], Dorothea Buck[2], Antonios Bayas[3], Thomas Bettecken[4], Andrew Chan[5], Sabine Fleischer[6], Andre Franke[7], Ralf Gold[5], Christiane Grätz[8], Christoph Heesen[6], Karl-Heinz Jöckel[9], Bernd C Kieseier[10], Tania Kümpfel[11], Wolfgang Lieb[12], Markus M Nöthen[13], Friedemann Paul[14], Vilmos Posevitz[15], Martin Stangel[16], Konstantin Strauch[17,18], Björn Tackenberg[19], Florian T Bergh[20], Hayrettin Tumani[21], Melanie Waldenberger[22,23], Frank Weber[24], Brigitte Wildemann[25], Uwe Zettl[26], Frauke Zipp[8], Bertram Müller-Myhsok[24], Heinz Wiendl[15], Bernhard Hemmer[2], Andreas Ziegler[1,27] on behalf of the German Competence Network for Multiple Sclerosis (KKNMS)

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany
[2]Klinikum rechts der Isar, Department of Neurology, Technische Universität München, Munich, Germany
[3]Department of Neurology, Klinikum Augsburg, Augsburg, Germany
[4]Max Planck Institute of Psychiatry, Munich, Germany
[5]Neuroimmunologisches Labor, St. Josef-Hospital, Universitätsklinikum der Ruhr-Universität Bochum, Bochum, Germany
[6]Klinik und Poliklinik für Neurologie, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany
[7]Institut für Klinische Molekularbiologie, Christian-Albrechts-Universität zu Kiel, Germany
[8]Klinik und Poliklinik für Neurologie, Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Mainz, Germany
[9]Institut für Medizinische Informatik, Biometrie und Epidemiologie, Universitätsklinikum Essen, Essen, Germany
[10]Neurologische Klinik, Heinrich-Heine Universität, Düsseldorf, Germany
[11]Institut für Klinische Neuroimmunologie, Ludwig-Maximilians-Universität München, München, Germany
[12]Institut für Epidemiologie and Biobank popgen, Christian-Albrechts-Universität zu Kiel, Germany
[13]Institut für Humangenetik, Universitätsklinikum Bonn, Bonn, Germany
[14]NeuroCure Clinical Research Center, Charité - Universitätsmedizin Berlin, Germany
[15]Klinik für Allgemeine Neurologie, Universiätsklinikum Münster, Münster, Germany
[16]Klinik für Neurologie, Medizinische Hochschule Hannover, Hannover, Germany
[17]Institute of Genetic Epidemiology, Helmholtz Zentrum München – German, Research Center for Environmental Health, Neuherberg, Germany
[18]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany
[19]Klinik für Neurologie, Philipps-Universität Marburg, Marburg, Germany
[20]Klinik und Poliklinik für Neurologie, Universitätsklinikum Leipzig, Leipzig, Germany
[21]Klinik und Poliklinik für Neurologie der Universität Ulm, Ulm, Germany
[22]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
[23]Institute of Epidemiology II, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[24]Max-Planck-Institut für Psychiatrie, München, Germany
[25]Neurologische Klinik, Universität Heidelberg, Heidelberg, Germany
[26]Klinik für Neurologie und Poliklinik, Universitätsklinikum Rostock, Universität Rostock, Rostock, Germany
[27]Zentrum für Klinische Studien, Universität zu Lübeck, Lübeck, Germany

Background: Several genome-wide association studies (GWAS) were conducted in the past few years to identify genetic variants associated with multiple sclerosis (MS). The objective of this study was the

replication of observed findings using the exome array. Methods: 4,476 German MS cases and 5,714 German controls were genotyped using Illumina's HumanExome v1-Chip. Genotype calling was performed with Illumina's Genome StudioTM Genotyping Module, followed by zCall. Results: Replication was successful for 9 regions beside the HLA region that are listed in the Catalog of Published Genome-Wide Association Studies as associated with MS. Criteria for replication were SNPs with $p<10\text{-}5$ that were either identical to reported SNPs or in linkage disequilibrium with $r2 > 0.8$ to reported SNPs or were located in the reported gene. Many SNPs in various HLA genes reached genome-wide significance ($p<5x10\text{-}8$). Collapsing methods for rare variants gave similar results. Overall, replication of reported findings was possible using the exome array. One association identified in this study was not reported before in any previous GWAS. Specifically, we found genome-wide significance to the gene MMEL1 which was found to be associated with MS in a candidate gene study by Ban (2010 Genes Immun 11:660-4) and SNPs in the vicinity (145 kb) to MMEL1 were identified by Sawcer (2011 Nature 476:214-9). Conclusion: In this study, findings of previous GWAS could be replicated in a large German consortium using the exome array. This is especially important because the German population shows only low levels of population substructure and is therefore well suited for the investigation of complex diseases.

Categories: *Association: Genome-wide*

P 76

Shared Genetic Effects Underlying Age at Menarche, Age at Natural Menopause and Blood Pressure

Erin K Wagner[1], Jin Xia[1], Yi-Hsiang Hsu[2], Chunyan He[1]
[1]Indiana University Richard M. Fairbanks School of Public Health
[2]Harvard Medical School

Age at menarche (AM) and age at natural menopause (ANM) are both associated with the risk of cardiovascular disease and its risk factors including blood pressure (BP). BP is known to increase rapidly during puberty, and early menarche is associated with elevated BP in adolescent and adulthood. BP is also known to increase more steeply around age at menopause. Earlier menopause is associated with higher blood pressure, although it is still unclear whether menopause accelerates BP increase or increased BP leads to earlier menopause. The observed synchronization between reproductive aging and BP development raises questions about the possibility of common regulating mechanisms shared by these processes. Using data from genome-wide association studies, we performed a bivariate meta-analysis of these traits to identify genes with pleiotropic effects for AM, ANM and BP. We identified 6 novel loci at or near ARNTL (11p15.2), FTO (16q12.2), DCAKD (17q21.31), ZNF652 (17q21.32), 14q32.2 and 20q13.32 (intergenic regions) were associated with AM and BP ($P<5X10-8$). For the bivariate analysis for ANM and BP, we found multiple variants within 200kb region at the 6p21.33 locus were significantly associated with ANM and BP. This region harbors genes including PRRC2A, BAG6, DDAH2, VW7, and HSPA1B. Our results suggest shared genetic effects for AM, ANM and BP. The findings may help improve the understanding of the genetic architecture and molecular mechanisms underlying these traits.

Categories: *Association: Genome-wide, Multivariate Phenotypes*

P 77

Identification of combined Common- and Rare- Genetic variances associated with renal function in Han Chinese

Guanjie Chen[1], Zhenjian Zhang[2], Adebowale Adeyemo[1], Yanxun Zhou[2], Ayo Doumatey[1], Guozheng Liu[2], Amy Bentley[1], Daniel Shriner[1], Congqing Jiang[2], Charles N Rotimi[1]

[1]Center for Research on Genomics and Global Health, National Human Genome Research Institute, NIH, Bethesda, Maryland, USA

[2]Suizhou Central hospital, Suizhou, Hubei, China

The public health burden of Chronic Kidney Disease (CKD) is increasing in developing countries including China with an overall prevalence of 10.8% defined as eGFR less than 60 mL/min per 1·73 m² or presence of albuminuria; thus, about 120 million Chinese have CKD. Both genetic and non-genetic factors including economic status, area of residence, age, hypertension, diabetes and history of CVD contribute to the development of CKD. Here, we investigate the contribution of rare and common exonic variants to susceptibility to CKD by analyzing exome array data in 991 Han Chinese genotyped with the Affymetrix Axiom Exome Genotyping Arrays. A total of 64,397 SNPs that passed QC filters with minor allele count ≥ 5 within 17,266 gene sets were carried forward for analysis; 6,649 gene sets had common variants only (8802 SNPs), 8802 gene sets had both common and rare variants, and 1815 gene sets had only rare variants. Common variants analysis was implemented in PLINK assumed additive genetic model. The common and rare gene sets analysis was implemented in the Simultaneous Analyses of Common and Rare Variants in complex traits (SCARVAsnp) statistical package. Analyses were adjusted for age, sex, BMI, and hypertension status. We identified significant associations (pvalue<2.57×10-6) in DIDO1, MOG, and GAB2, and suggestive significant association (pvalue<2.57x10-5) in DNAH5, LAMC3, and TRAP1 with observed the lowest p value of 3.6×10-10. We replicated seven of the sixteen and six of the ten genes reported to be associated with renal disease respectively in European and Chinese ancestry studies. These findings promise to provide novel insight into the genetic basis of CKD in Chinese and perhaps other human populations

Categories: *Association: Genome-wide, Multilocus Analysis*

P 78

Pathway and gene-gene interaction analysis reveals new candidate genes for melanoma

Myriam Brossard[1], Shenying Fang[2], Amaury Vaysse[1], Qingyi Wei[3], Hamida Mohamdi[1], Marie-Françoise Avril[4], Mark Lathrop[5], Jeffrey E Lee[2], Christopher I Amos[6], Florence Demenais[1]
[1]INSERM, UMR-946, Paris, France; Université Paris Diderot, Paris, France
[2]MD Anderson Cancer Center, Houston, Texas, USA
[3]Department of Medicine, Duke University School of Medicine, Durham, USA
[4]H*ôpital* Cochin, Université Paris Descartes, Paris, France
[5]Genome Quebec Innovation Centre, McGill University, Montreal, Canada
[6]Geisel College of Medicine, Dartmouth College, New Hampshire, USA

GWAS have identified 17 loci associated with melanoma, but these loci account for a small part of melanoma risk. These GWAS used single-SNP analysis which may be underpowered to detect SNPs with small effect and/or interacting with other SNPs. To identify new candidate genes for melanoma risk, we combined pathway analysis and tests of gene-gene interactions within melanoma-associated pathways. Pathway analysis was based on the gene-set enrichment analysis (GSEA) approach, using the Gene Ontology (GO) database. GSEA was applied to single-SNP statistics obtained from melanoma GWAS of the MELARISK study (3,976 subjects) and MDACC study (2,827 subjects). To identify GO categories enriched in association signals, the false discovery rate (FDR) was computed using 100,000 SNP permutations. We tested all SNP-SNP interactions within the identified GOs using INTERSNP. One million Hapmap3-imputed SNPs were assigned to 22,000 genes, which were assigned to 316 Level 4-GO categories. We identified 5 GOs with FDR≤5% in the two studies: response to light stimulus, regulation of mitotic cell cycle, induction of programmed cell death, cytokine activity and oxidative phosphorylation. A total of 110 genes were driving the enrichment signals in these GOs. Nine of these genes were found to occur frequently with melanoma-related terms through PubMed mining, of which 5 are new candidates for melanoma risk (TP63, MAPK1, IL6, IL15, NDUFA2). Gene-gene interaction analysis within each of the 5 identified GOs showed evidence for interaction for 4 SNP pairs (P≤10-4 in MELARISK and replication at 5% in MDACC). Two of these pairs, CMTM7-TNFSF4 (combined P=3x10-7) and TERF1-AFAP1L2 (combined P=2x10-6), are biologically relevant. Funding: INCa_5982, LNCC, FRM

Categories: *Association: Genome-wide, Cancer, Gene - Gene Interaction, Multilocus Analysis, Pathways*

P 79

Leveraging evolutionarily conserved, cell type-specific, regulatory region data to detect novel SNP-TFPI associations

Jessica Dennis[1], Alejandra Medina-Rivera[2], Vinh Truong[1], Lina Antounians[2], Pierre Morange[3], David Trégouët[4], Michael Wilson[2], France Gagnon[1]

[1]Dalla Lana School of Public Health, University of Toronto, Canada

[2]Genetics & Genome Biology Program, SickKids Research Institute, Toronto, Canada

[3]Faculty of Medicine, University of the Mediterranean, Marseille, France

[4]Université Pierre et Marie Curie, Paris, France

Low plasma levels of tissue factor pathway inhibitor (TFPI), a key regulator of the extrinsic coagulation cascade, increase the risk of venous and arterial thrombosis. TFPI plasma levels are highly heritable, but the genetics underlying this heritability are poorly understood. Genetic variants in evolutionarily conserved, cell type-specific gene regulatory regions are important to complex traits. Incorporating this information in genome-wide association studies (GWAS) may increase power. We experimentally ascertained regulatory regions in human and rat aortic endothelial cells (EC; a primary source of TFPI) using ChIP-seq for epigenetic histone modifications and transcription factors. We then conducted a GWAS of SNPs associated with TFPI in 253 individuals from 5 French-Canadian families ascertained on venous thrombosis (VT), prioritizing SNPs in these regulatory regions via stratified false discovery rate (sFDR) control. We tested SNPs with sFDR <0.25 for replication in 1170 French VT patients and, in both study samples, tested the significance of our prioritization scheme by comparing the median t-statistic of prioritized SNPs and SNPs selected from comparable random regions. None of the 39 SNPs associated with TFPI in the discovery sample replicated at an FDR <0.05. Although our prioritization scheme did not help identify TFPI-associated SNPs, defining novel approaches sFDR approaches is of great interest. Since TFPI is up-regulated in inflamed vascular EC, we will next prioritize SNPs in experimentally determined inflammation-specific vascular EC genes and their regulatory regions.

Categories: *Association: Genome-wide, Bioinformatics, Data Integration, Epigenetic Data, Epigenetics, Sequencing Data*

P 80

A software package for genome-wide association studies with Random Survival Forests

Marvin N Wright[1], Andreas Ziegler[1,2]
[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany
[2]Zentrum für Klinische Studien, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany

In recent years, Random Forests have been successfully used to analyze genome-wide association studies (GWAS) with dichotomous and quantitative endpoints. For censored survival endpoints software is available, namely Random Survival Forests and Conditional Inference Forests. However, due to computational burdens and memory issues, these tools are not capable of handling high-dimensional data on GWAS scale. Consequently, we are not aware of any study applying one of them to genome-wide data. We therefore introduce the new software package Random Jungle 3, which embeds the functionality of Random Survival Forests into the computationally efficient framework of Random Jungle. Compared to the original implementation, the runtime is reduced considerably, making the analysis of GWAS data possible. We validate the new software in extensive simulation studies. Finally, we apply it to a real dataset to assess the importance of involved single nucleotide polymorphisms (SNPs).

Categories: *Association: Genome-wide, Bioinformatics, Data Mining, Machine Learning Tools*

P 81

Identification of novel common and rare genetic variants associated with renal function in Han Chinese

Guanjie CHEN[1], Zhenjian Zhang[2], Adebowale Adeyemo[1], Yanxun Zhou[2], Ayo Doumatey[1], Jie ZHOU[1], Amy Bentley[1], Daniel Shriner[1], Charles Rotimi[1]

[1]Center for Research on Genomics and Global Health, National Human Genome Research Institute, NIH, Bethesda, Maryland, USA

[2]Suizhou Central hospital, Suizhou, Hubei, China

The public health burden of CKD is increasing in developing countries including China with an overall prevalence of 10.8% defined as eGFR less than 60 mL/min per 1·73 m² or presence of albuminuria; thus, about 120 million Chinese have CKD. Both genetic and non-genetic factors including economic status, area of residence, age, hypertension, diabetes and history of CVD contribute to the development of CKD. Here, we investigate the contribution of rare and common exonic variants to susceptibility to CKD by analyzing exome array data in 991 Han Chinese genotyped with the Affymetrix Axiom Exome Genotyping Arrays. A total of 64,397 SNPs that passed QC filters with minor allele count ≥ 5 within 17,266 gene sets were carried forward for analysis; 6,649 gene sets had common variants only, 8802 gene sets had both common and rare variants, and 1815 gene sets had only rare variants. Common variants analysis was implemented in PLINK assumed additive genetic model. The common and rare gene sets analysis was implemented in the Simultaneous Analyses of Common and Rare Variants in complex traits (SCARVAsnp) statistical package. Analyses were adjusted for age, sex, BMI, and hypertension status. We identified significant associations (pvalue<2.57×10-6) in DIDO1, MOG, and GAB2, and suggestive significant association (pvalue<2.57x10-5) in DNAH5, LAMC3, and TRAP1 with observed lowest p value of 3.6 ×10-10. We replicated seven of the sixteen and six of the ten genes reported to be associated with renal disease respectively in European and Chinese ancestry studies. These findings promise to provide novel insight into the genetic basis of CKD in Chinese and perhaps other human populations.

Categories: *Association: Genome-wide, Multilocus Analysis*

P 82

A Genome-Wide Association Study to Explore Gene-environment Interaction with Parental Smoking and the Risk of Childhood Acute Lymphocytic Leukemia

Jessica L Barrington-Trimis[1]
[1]University of Southern California, Keck School of Medicine, Department of Preventive Medicine

Genetic susceptibility to parental smoking around pregnancy and risk of childhood acute lymphocytic leukemia (ALL) has not been fully explored. In this analysis, we used novel methods to scan the genome for gene-parental smoking interactions. Participants were Hispanic cases and controls participating in the California Childhood Leukemia Study. Cases (N=380) were <15 years of age at diagnosis, and controls (N=454) were matched to cases on date of birth, gender, and maternal race. Genome-wide genotyping was conducted using DNA from archival dried blood spot samples using the Illumina Human OmniExpress v.1 platform. Data were evaluated for the presence of multiplicative gene-parental smoking interaction using statistically efficient two-step scanning methods. We sought to replicate our most significant SNPs in two case-only studies of childhood ALL in France (ESCALE, n=441), and Australia (AUS-ALL, n=285). We identified two SNPs for replication for maternal smoking prior to and during pregnancy. One SNP was statistically significant in the AUS-ALL replication, with the strongest results for maternal smoking during pregnancy, restricting to B-cell progenitor ALL (summary interaction OR [CCLS/AUS-ALL] = 4.40; 95% CI: 2.53, 7.64). Genotyping data for this SNP was not available in the ESCALE study. A second SNP was suggestive of a potential interaction in the AUS-ALL replication (P=0.078, B-cell ALL), but not in the ESCALE study where the interaction OR was in the opposite direction. Results indicate potential novel susceptibility loci for maternal smoking during pregnancy and risk of B-cell ALL. Additional studies should be conducted to confirm these results in larger study populations of similar ethnic background.

Categories: *Association: Genome-wide, Cancer, Case-Control Studies, Gene - Environment Interaction*

P 83

Network-based analysis of GWAS data : Does the gene-wise association significance modeling matters?

Julie HAMON[1], Yannick ALLANORE[2], Maria MARTINEZ[1]
[1]INSERM UMR1043, Hôpital Purpan, Toulouse
[2]INSERM 1016, Hôpital Cochin, Paris

Integrating prior biological knowledge into Genome-Wide Association data may unravel sets of genes having collectively or in interaction a role on the disease. Several network-based approaches have been proposed depending on the type of known information that is used to combine the genes such as protein-protein interaction (PPI) network or gene functions pathways. These studies rely on the association of each gene with the disease, i.e., on an individual Gene-Wise P-value (GWP) which can be derived under different alternatives. Here, we aim to compare such different strategies in our Systemic Sclerosis GWAS data. We built a two-stage network study by randomly splitting our GWAS data into a scan and a replication dataset. In the scan dataset we performed a PPI network-based approach using a dense module search strategy with different GWP values: for instance using the smallest single-SNP P-value either unadjusted (Min) or Bonferroni-adjusted (Bonf) or using the Fisher's method to combine all single-SNP P values.The results were compared according to the length (number of genes) of the enriched sub-modules and the characteristics of their genes. The top (5 and 10%) most enriched modules were tested for enrichment analysis in the replication dataset. We finally mapped the genes from the replicated sub-networks to KEGG pathways. Overall, we found low consistency across the results from the different strategies: different sets of genes are selected but also different KEGG pathways are identified.

Categories: *Association: Genome-wide, Gene - Gene Interaction, Pathways*

P 84

Heritability estimates and genetic association for 60+ complex traits in a young healthy sibling cohort

Jun Z Li[1], Qianyi Ma[1], Ayse B Ozel[1], Karl C Desch[2], David Ginsburg[3]
[1]Department of Human Genetics, University of Michigan, Ann Arbor
[2]Department of Pediatrics and Communicable Disease, University of Michigan, Ann Arbor
[3]Howard Hughes Medical Institute, Department of Internal Medicine, University of Michigan, Ann Arbor

As genotyping becomes more efficient, sample recruitment and phenotyping remain a major limiting factor. In a GWAS of bleeding and blood clotting traits we sought to increase the utility of the cohort by collecting > 60 self-reported complex traits through web-based questionnaires. The cohort of 1,191 healthy young subjects consists of 509 sibships, 80% Europeans, and age of 14-35 yrs. The traits include 16 quantitative traits (e.g., weight, height, age of menarche, hematological measures RBC, HCT, MCV, MCH, MCHC, RDW, WBC, HGB, PLT, MPV), 21 ordinal traits (e.g., Smoking, BleedingTendency, SkinTags, Acne, TanningTendency, SkinColor, Freckles, DentalCaries, VisionCorrection, EatingSweets, EatingSaltyfood, Athleticability, Aphthousulcers), and 27 nominal traits (e.g., Immunization, ToothExtraction, EyeColor, HairColor, Hairline, EarLobeCreased, EarLobeAttachment, Dimples, Dyslexia, Migraines, Stuttering, Allergies, Flatfeet, Handedness, PhoticSneeze, BrainFreeze, InterlockingFingers, etc.). We used the known relatedness to estimate heritability using Merlin-regress and found that >1/2 of the traits have H^2 > 40%. Since the samples have been genotyped over ~800K SNPs in the original GWAS we used SNP data to calculate the actual genetic relatedness, and estimated the variance explained by all the genotyped SNPs using GCTA. With all subjects, pedigree-based estimates were similar to SNP-based estimates; but the latter were often reduced when we select one subject from each sibship to analyze the unrelated subsets. For many traits we identified common variants of significant association. This study demonstrates the feasibility of simultaneous analysis of dozens of traits via web-based profiling.

Categories: *Association: Genome-wide, Heritability, Multivariate Phenotypes, Quantitative Trait Analysis*

P 85

Large-scale exome chip genotyping reveals novel coding variation associated with endometriosis

Andrew P Morris[1], Reedik Mägi[2], Nilufer Rahmioglu[3], Anubha Mahajan[3], Neil Robertson[3], Marie Peters[4], Merli Saare[4], Andres Salumets[4], Krina T Zondervan[3]

[1]Department of Biostatistics, University of Liverpool, Liverpool, UK

[2]Estonian Genome Centre, University of Tartu, Tartu, Estonia

[3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

[4]Department of Obstetrics and Gynaecology, University of Tartu, Tartu, Estonia

Genome-wide association studies have identified nine loci harbouring common variants implicated in endometriosis, which together explain only ~3% of the heritability of the condition. To investigate the contribution of coding variation to endometriosis pathogenesis, we undertook genotyping with the Illumina Exome Chip of two studies of European ancestry: (i) 910 cases from the Oxford Endometriosis Gene study and 13,334 population controls from the UK Exome Chip Consortium; and (ii) 326 cases and 711 population controls from the Estonian Biobank. Within each study, we evaluated the association of endometriosis with: (i) individual coding variants; and (ii) burden/over-dispersion of loss of function (all frequencies) and rare non-synonymous (minor allele frequency [MAF] less than 1%) variants within genes using SKAT-O. Association summary statistics were combined across studies by meta-analysis. We conducted pathway analysis on the basis of single variant meta-analysis summary statistics using MAGENTA. No individual coding variants achieved exome-wide significant evidence of association ($p<5x10-7$, Bonferroni correction for 100,000 variants). The strongest signals include missense variants in TAF1L (D141N, p=1.5x10-5, MAF=0.077%) and BMP3 (Y67N, p=3.2x10-5, MAF=2.7%). We observed exome-wide significant evidence of association ($p<2.5x10-6$, Bonferroni correction for 20,000 genes) with burden/over-dispersion of loss of function variants in C16orf89 (p=1.1x10-6) and rare non-synonymous changes in NECAB3 (p=1.7x10-7), ZNF485 (p=1.1x10-6), and RSAD2 (p=2.1x10-6). MAGENTA analyses highlighted potential involvement of cell adhesion/structure, immune function and cancer-related pathways in endometriosis.

Categories: *Association: Genome-wide, Case-Control Studies*

P 86

Dissecting the Obesity Disease Landscape: Identifying Gene-Gene Interactions that are Highly Associated with Body Mass Index (BMI)

Rishika De[1], Shefali Setia Verma[2], Sarah Pendergrass[2], Fotios Drenos[3], Michael Holmes[4], Folkert Asselbergs[5], Brendan Keating[4], Marylyn Ritchie[2], Diane Gilbert-Diamond[1]
[1]Dartmouth College
[2]Pennsylvania State University
[3]University College London
[4]University of Pennsylvania
[5]University Medical Center Utrecht

Though obesity is estimated to have a heritability of 40-70%, less than 2% of its variation is explained by the BMI-associated loci that have been identified so far. Hence, interactions between genes, i.e. epistasis, may explain a larger portion of the heritability of BMI. We analyzed genetic information from 18,686 individuals across 5 cohorts – ARIC, CARDIA, FHS, CHS, MESA – to identify interactions between SNPs (Single Nucleotide Polymorphisms). Participants were genotyped using a targeted approach via the gene-centric IBC array (ITMAT-Broad-CARe). SNPs were filtered using two parallel approaches – one based on the strength of their main effects of association, and the other a knowledge-based approach called Biofilter that identifies biologically plausible SNP-SNP models. Filtered SNPs were analyzed using QMDR (Quantitative Multifactor Dimensionality Reduction) to detect SNP-SNP interactions that are highly associated with BMI. QMDR is a nonparametric, genetic model-free method that detects non-linear interactions in the context of a quantitative trait. We identified 6 novel interactions with a Bonferroni corrected p-value of association < 0.05. These interactions also replicated previously identified BMI-associated independent signals - rs12617233 in FLJ30838, rs997295 in MAP2K5, and rs1799998 in CYP11B2. Our results highlighted interactions between genes involved in mitochondrial dysfunction (POLG2), aldosterone synthase functioning (CYP11B2), cell proliferation (MAP2K5), insulin resistance (IGF1R, CAV3), vascular development (MAP2K5, EZR), cell adhesion (EZR) and apoptosis (EZR). This study highlights a novel approach to discovering gene-gene interactions within the obesity disease landscape.

Categories: *Association: Genome-wide, Bioinformatics, Gene - Gene Interaction, Genomic Variation, Quantitative Trait Analysis*

P 87

Investigation of Parent-of-Origin effects in Autism Spectrum Disorders

Siobhan Connolly[1], Elizabeth A Heron[1]
[1]Trinity College Dublin

The detection of parent-of-origin effects aims to identify whether or not the functionality of alleles, and in turn associated phenotypic traits, depends on the parental origin of the alleles. Genome-Wide Association Studies (GWAS) have had limited success in explaining the heritability of many complex disorders and traits but successful identification of parent-of-origin effects using trio (mother, father, offspring) GWAS may help shed light on this missing heritability. Autism Spectrum Disorders (ASDs) are considered to be heritable neurodevelopmental disorders and a number of trio GWAS datasets exist for examining this heritability. Here, we have investigated parent-of-origin effects in large trio GWAS datasets that have previously been analysed for parent-of-origin effects using statistical approaches that did not have the capacity to detect epigenetic effects such as maternal-offspring genetic effects and all assumptions of the approaches may not have been satisfied. Here the approach of Estimation of Maternal, Imprinting and Interaction Effects Using Multinomial Modelling (EMIM) is used to identify SNPs associated with ASD through a parent-of-origin mechanism which has the potential to aid in understanding more fully the genetic underpinnings of ASD.

Categories: *Association: Genome-wide, Psychiatric Diseases, Transmission and Imprinting*

P 88

Integrative clustering of multiple genomic data using Non-negative Matrix Factorization

Prabhakar Chalise[1], Brooke L Fridley[1]
[1]University of Kansas Medical Center, Kansas City, KS USA

We propose a novel approach for integrative clustering of multiple genomic data sets to classify the disease subtypes. The method uses Non-negative Matrix Factorization (NMF) technique by extending the existing method for single data in order to utilize the strengths across multiple data types. This analysis approach was applied to the cancer genome atlas (TCGA) studies on ovarian cancer involving 499 subjects that have both gene expression (90797 probes) and methylation (27338 probes) assays on tumor samples available. To get clinically meaningful clusters, top 500 most associated probes from each data set were selected by fitting cox proportional hazards model with time to recurrence (TTR) of the disease as end point for each probe adjusting for age and cancer stage. The integrative method resulted in three optimum clusters of samples. The phenotypic differences of TTR among these clusters were assessed by Kaplan Meier plot followed by log rank test (p =1.0×10-11). Further, each expression and CpG probe was assessed across the three clusters using analysis of variance followed by multiple testing adjustments (Benzamini and Hochberg). Among other significant probes, the genes TXNDC9 (p=7.23×10-35) and CRBN (p=8.44×10-32) and the CpG probes near genes PLOD2 (p=1.91×10-4) and KRTAP11-1 (p=3.24×10-4) were found to be most significantly different across the clusters. Further studies are needed to determine the functional relevance of these genes in the ovarian cancer etiology.

Categories: *Association: Genome-wide, Cancer, Data Integration*

P 89

Tools for robust analysis in genome-wide association studies using STATA

Niki Dimou[1], Pantelis Bagos[1]
[1]University of Thessaly

Within the context of genetic association studies (GAS) and genome-wide association studies (GWAS) there is a variety of statistical techniques in order to conduct the analysis but a common problem is the lack of knowledge concerning the model of inheritance. Several approaches have been proposed for deriving robust procedures that will detect the true underlying model of inheritance and, at the same time perform the analysis maximizing the power and preserving the nominal type I error rate. The primary goal of this work is to implement as many as possible robust methods within the statistical package STATA and subsequently to make the software available to the scientific community. Robust methods based on the MAX statistic, the MERT statistic, the MIN2, as well as the GMS and the GME procedures were implemented in STATA and immediate commands were constructed. The main difficulty in implementing the above-mentioned methods is the fact that they are computationally intensive since (with the exception of MERT) the asymptotic properties of the estimators cannot be derived analytically and other methods are needed. Concerning MAX, GMS and GME, we used a several fast Monte Carlo simulation methods in order to calculate accurate p-values, whereas for MIN2, we relied on numerical integration. This is the first complete effort to implement procedures for robust analysis and selection of the appropriate genetic model in GAS or GWAS using STATA. Since there are only a few available software implementations of the robust methods for meta-analysis of GAS or GWAS our future goal is to extend our software in the context of meta-analysis using STATA. The software is available at http://www.compgen.org/tools/robust-meta-analysis.

Categories: *Association: Genome-wide*

P 90

Development of a three-way mixed modelling approach integrating genetic and clinical variables in analysis of early treatment outcomes in epilepsy.

Ben Francis[1], Andrea Jorgensen[1], Andrew Morris[1], Anthony Marson[1], Michael Johnson[2], Graeme Sills[1], EpiPGX consortium
[1]University of Liverpool
[2]Imperial College London

Remission from seizures (12 months of seizure freedom) is indicative of therapeutic response to an antiepileptic drug (AED) when treating epilepsy patients. Clinical factors including gender and epilepsy type have been attributed to the remission outcome and potential pharmacogenetic factors are now being investigated.

A total of 964 patients from the Standard and New Antiepileptic Drug (SANAD) study, a randomised trial that compared treatments with various AEDs in patients with newly diagnosed epilepsy, were genotyped to investigate genetic biomarkers for time to remission, as well as other longitudinal phenotypes, including time to AED withdrawal and time to first seizure. Analysis was initially undertaken using a traditional one component survival model for time to remission, however, no genome-wide significant SNPs were found.

This method may lack power as the population is considered homogeneous. The presence of three sub-populations for time to remission is apparent; those who experience remission immediately, those who experience remission eventually and those who do not experience remission at any point during follow-up. To consider these sub-populations, a three component model is required for survival analysis. Mixture modelling with a cure fraction was selected as the optimal methodology to derive a three component model.

The further adapted methodology proposed in this abstract will be applied to a larger population of patients with newly-diagnosed epilepsy that is now available via the EpiPGX consortium (www.epipgx.eu), and which includes the SANAD cohort as well as other cohorts of patients being collected worldwide to investigate genetic biomarkers of epilepsy.

Categories: *Association: Genome-wide, Multivariate Phenotypes, Population Stratification, Prediction Modelling, Psychiatric Diseases*

P 91

Meta-analysis of low frequency and rare coding variants and pulmonary function.

Victoria E Jackson[1], Louise V Wain[1], Ian Sayers[2], Ian P Hall[3], Martin D Tobin[1], SpiroMeta Consortium
[1]Department of Health Sciences, University of Leicester, Leicester, United Kingdom
[2]Division of Respiratory Medicine, University of Nottingham, Nottingham, United Kingdom
[3]Division of Therapeutics and Molecular Medicine, University of Nottingham, Nottingham, United Kingdom

Pulmonary function measures are an important predictor of mortality and morbidity and are used in the diagnosis of a number of diseases, including chronic obstructive pulmonary disease (COPD). A number of large-scale genome-wide association studies (GWAS) have successfully identified single nucleotide polymorphisms (SNPs) influencing pulmonary function in 26 regions; however these so far identified regions only account for a small proportion of the estimated heritability. One hypothesis is that the so-called "missing heritability" might be found in rare variants with large effects. A genotyping array has recently been developed as a cost-effective way to investigate the effects of rare variants in large sample sizes. The variants included in the array design were selected as they were observed numerous times in the sequenced exomes or genomes of a set 12,000 individuals from 16 sample collections and arepredominantly low frequency and rare exonic SNPs. We carried out a meta-analysis of exome array data and three pulmonary function measures (FEV1, forced vital capacity (FVC) and the ratio of FEV1 to FVC (FEV1/FVC)) in over 30,000 individuals of European ancestry, from 12 studies, who had been genotyped using the Illumina HumanExome beadchip. We have utilised single variant association analysis methods, traditionally employed in GWAS, along with gene-based methods, which for the joint effect of several variants in a gene; the latter method is considered a more powerful approach to identify rare variants associated with a trait. We present emerging findings from these analyses.

Categories: *Association: Genome-wide, Quantitative Trait Analysis*

P 92

Using Polygene Scores and GCTA to Identify a Subset of SNPs that Contribute to Genetic Risk

Elizabeth A Heron[1], Alison K Merikangas[1], Ricardo Segurado[2]
[1]Department of Psychiatry & Neuropsychiatric Genetics Research Group, Trinity College Dublin, Dublin 2, Ireland
[2]Centre for Support and Training in Analysis and Research, University College Dublin, Dublin 4, Ireland

Polygene scores1 are a means of summarising the combined effect of a group of markers, in this case single nucleotide polymorphisms (SNPs), that as individual markers perhaps do not reach statistical significance in a genome-wide association study (GWAS), but in aggregate are associated with case status. The polygenic scoring method offers a means by which a reduced set of markers can be identified that offer good prediction for a particular trait and can perhaps narrow the focus of the associated genetic risk factors. Genome-wide Complex Trait Analysis (GCTA)2 is a method by which the proportion of phenotypic variance that is explained by SNPs can be estimated. Thus, a given set of SNPs can be compared with another set of SNPs to determine which set explains more of the genetic component of the variability in the phenotype. The aim of this paper is to combine these two methodologies to identify a subset of SNPs that both contribute significantly to the genetic component of the phenotypic variance but that also offer good prediction for a phenotypic trait of interest. A number of GWAS datasets together with simulated data will be used to explore this approach which offers the potential to aid in the difficult task of identifying risk variants for complex disorders. 1. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009; 460: 748– 52. 2. Yang J, Lee SH, Goddard ME and Visscher PM. GCTA: a tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011 Jan 88(1): 76-82.

Categories: *Association: Genome-wide, Association: Unrelated Cases-Controls, Case-Control Studies, Causation*

P 93

Challenging Issues in GWAS of Human Aging and Longevity

Anatoliy I Yashin[1], Deqing Wu[1], Konstantin Arbeev[1], Alexander Kulminski[1], Liubov S Arbeeva[1], Svetlana V Ukraintseva[1]
[1]Duke University

Anatoliy I. Yashin, Deqing Wu, Konstantin G. Arbeev, Liubov S. Arbeeva, Alexander M. Kulminski, Svetlana V. Ukraintseva During last decade substantial progress in genetic analyses of complex traits has been observed. Encouraged by this progress the genome wide association studies (GWAS) of human aging and longevity have been performed. The results of these studies were much less impressive, however. Strong associations of genetic variants linked to APOE, FOXO3A and to several other genes with human life span observed in a number of studies were accompanied by many associations that have not reached the level of genome wide statistical significance. Most research findings suffered from the lack of replication in studies of independent populations. In this paper we investigate reasons that might be responsible for slow progress in genetic analyses of data on aging and longevity traits. We showed that one such reason deals with the fact that bio-demographic aspects of aging and longevity traits have been ignored. The genetic structure of study population gets modified as a result of mortality selection process which takes place in any genetically heterogeneous populations when some genes influence mortality risk. Such modification affects the results of association studies. We discuss benefits of using bio-demographic concepts and models in GWAS of human aging and longevity. Using simulated data, and then the Framingham Heart Study data we show how estimates of genetic associations with life span can be improved. Other reasons including multifactorial nature of aging and longevity traits, high genetic heterogeneity of these traits, pleiotropic effects of genetic variants on mortality risks at different age intervals are discussed.

Categories: *Association: Genome-wide*

P 94

Heritability estimates on Hodgkin lymphoma: a genomic versus population based approach

Hauke Thomsen[1], Miguel Inacio da Silva Filho[1], Asta Försti[1], Michael Fuchs[2], Elke Pogge von Strandmann[2], Per Hofmann[3], Stefan Herms[3], Jan Sundquist[4], Andreas Engert[2], Kari Hemminki[1]
[1]German Cancer Research Center (DKFZ), Division of Molecular Genetic Epidemiology, Heidelberg, 69120, Germany
[2]Department of Internal Medicine I, University Hospital of Cologne, Cologne, 50924, Germany
[3]Institute of Human Genetics and Department of Genomics, University of Bonn, 53127, Germany
[4]Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, 94305, USA.

Genome-wide association studies (GWAS) have identified several single-nucleotide polymorphisms influencing the risk of Hodgkin lymphoma (HL) and demonstrated the association of common genetic variation for this type of cancer. Such evidence for inherited genetic risk is also provided by the family history and very high concordance between monozygotic twins. However, little is known about the genetic and environmental contributions. A common measure for describing the phenotypic variation due to genetics is the heritability. Using GWAS data on 906 HL cases by considering all typed SNPs simultaneously, we have calculated that the common variance explained by SNPs accounts for more than 35% of the total variation on the liability scale in HL (95% confidence interval 6–62%). These findings are supported by similar heritability estimates of about 0.40 (95% confidence interval 0.17-0.58) based on Swedish population data. Our estimates support the underlying polygenic basis for susceptibility to HL, and show that heritability based on the population data is somehow larger than for the genomic data due to the possibility of some missing heritability in the GWAS data. Besides that there is still major evidence for multiple loci causing HL on chromosomes other than chromosome 6, which need to be detected. Due to limited findings in prior GWAS it seems to be worth to check for more loci causing susceptibility to HL

Categories: *Association: Genome-wide, Cancer, Case-Control Studies, Heritability, Population Genetics*

P 95

Are we able to guide treatment choice to reduce antidepressant-induced sexual dysfunction in males using genome-wide data from randomised controlled trials?

Andrew A Crawford[1], Sarah Lewis[1], Karen Hodgson[2], Peter McGuffin[2], David Nutt[3], Tim J Peters[4], Philip Cowen[5], Michael C O'Donovan[6], Nicola Wiles[1], Glyn Lewis[7]
[1]School of Social and Community Medicine, University of Bristol, Bristol
[2]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London
[3]Department of Neuropsychopharmacology, Imperial College, London
[4]School of Clinical Sciences, University of Bristol, Bristol
[5]Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford
[6]MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Cardiff
[7]Division of Psychiatry, University College London, London

Antidepressants are effective at reducing depressive symptoms but are also frequently associated with increased sexual dysfunction. Treatment emergent sexual dysfunction is drug specific (selective serotonin reuptake inhibitor (SSRI) or noradrenaline reuptake inhibitor (NARI)) and may be genetically determined. Identifying genetic markers able to guide treatment choice would be clinically important. The GENPOD study randomly allocated 601 depressed individuals to citalopram (SSRI) or reboxetine (NARI). Analysis was restricted to white, European men with data on sexual dysfunction (n=105). Genome-wide data were analysed using logistic regression in an additive genetic model, with an interaction term between genotype and drug. Replication analysis used data from the GENDEP study (n=202). Quantile-quantile plots suggest that population stratification was generally well controlled (lambda = 1). No association reached a genome-wide level of significance. Genetic variants near the PTPRD (P=0.0001) and PCDH9 (P=5.29x10-5) genes provided the strongest evidence of an association however, there was no evidence in our replication cohort (P>0.1). The lack of biological plausibility in our identified genes combined with a lack of evidence in our replication study lead us to conclude that larger trials are required before pharmacogenetics may be able to guide clinical practice in this area. The utilisation of data from a randomised controlled trial and the inclusion of an interaction term in our regression model allowed us to identify genetic variants whose association with sexual dysfunction differed by antidepressant, which are clinically important, but also increased our chances of obtaining spurious associations.

Categories: *Association: Genome-wide*

P 96

A GENERAL METHOD FOR TESTING GENETIC ASSOCIATION WITH ONE OR MORE TRAITS

Zeny Feng[1], William WL Wong[2]
[1]Department of Mathematics and Statistics, University of Guelph
[2]Leslie Dan Faculty of Pharmacy, University of Toronto

Genetic association study is an essential step for finding genetic factors that are associated with a complex trait. Many methods have been proposed for analysing data collected from different study designs. In this talk, we will present a very general method that based on the quasi-likelihood scoring approach for analysing data collected from a broad range of study designs. The proposed method can also be used to simultaneously test on multiple traits. Simulation studies and real data analysis will be included to show the performance of the proposed method.

Categories: *Association: Genome-wide, Multivariate Phenotypes*

P 98

A Generalized Similarity U test for Multiple-trait Sequencing Association Analyses

Changshuai Wei[1], Qing Lu[1]
[1]Department of Epidemiology and Biostatistics, Michigan State University

Sequencing-based studies are emerging as a major tool for genetic association research on complex diseases. These studies pose a great challenge to traditional statistical methods (e.g., single variant analysis) due to the high-dimensionality of the data and the low frequency of the genetic variants. A joint test has been shown to be more suitable for sequencing studies; jointly testing multiple variants increases the power and reduces the dimensionality. Meanwhile, there is a growing need for statistical methods that are distribution-free and that can handle multiple phenotypes. In this paper, we propose a generalized similarity U test, referred to as GSU. GSU first summarizes the genetic information and multiple traits into a genetic similarity and a trait similarity, and then combines the two similarities in the framework of a weighted U statistic. We derived the asymptotic distribution of GSU under a null hypothesis, so as to efficiently calculate the significance level. We also studied the asymptotic behavior of GSU under alternatives, and provided sample size and power calculations for the study design. To evaluate the performance of GSU, we conducted extensive simulation studies and compared them with the existing methods. Through simulation, we found that GSU had an advantage over existing methods in terms of power comparisons and its robustness to trait distribution. Moreover, GSU is computationally more efficient than existing methods. Finally, we applied GSU to sequencing data from the Dallas Heart Study, identifying 4 genes jointly associated with 5 metabolic-related traits.

Categories: *Association: Unrelated Cases-Controls, Case-Control Studies, Multilocus Analysis, Multivariate Phenotypes, Sequencing Data*

P 99

Modeling X-chromosome data in Random Forest Genetic Analysis

Joanna M Biernacka[1], Gregory Jenkins[1], Stacey J Winham[1]
[1]Mayo Clinic

The X chromosome is routinely excluded from genome-wide association studies. Random Forests (RF) have been proposed for genetic analysis involving many variants. We illustrate that for traits associated with sex, RF analysis yields biased results for X chromosome SNPs, and propose three extensions of RF to model X SNPs, based on (1) the principle of X chromosome inactivation (XCI), (2) stratification by sex, and (3) incorporation of sex as a variable in RF. We compare the performance of these approaches to traditional RF using simulations and analysis of data from the Study of Addiction: Genes and Environment (SAGE). Comparison of the SAGE data results for autosomal vs. X SNPs shows that traditional RF ranks X SNPs too high, whereas the three new approaches rank the X SNPs similar to autosomal SNPs. To evaluate the alternative approaches to incorporating X-SNP data in RF, we investigate variable importance (VI) measures for autosomal and X SNPs in simulated data with and without X SNP effects. We perform simulations under varying degrees of sex-trait association, case/control ratios, and patterns of linkage disequilibrium. All methods correctly estimate VI if sex is not associated with the trait, but when sex is associated with the trait, traditional RF leads to inflated VI for the X chromosome. Incorporating sex in RF does not properly correct this bias, whereas the methods based on XCI and stratified RF do not inflate the VI of X SNPs. Thus, we conclude that if sex is not associated with the trait in the sample, regular RF may be used to analyze X SNP data. Otherwise, either stratification of the forest or extension based on XCI should be used to avoid overestimation of X SNP importance. Future investigation will compare the power of these two methods.

Categories: *Association: Unrelated Cases-Controls, Case-Control Studies, Data Mining, Machine Learning Tools*

P 100

Empirical Bayes Scan Statistics for Detecting Clusters of Disease Risk Variants in Genetic Studies, with Applications to CNVs in Autism

Iuliana Ionita-Laza[1], Kenneth McCallum[1]
[1]Columbia University

Recent developments of high-throughput sequencing technologies offer an unprecedented detailed view of the genetic variation in various human populations, and promise to lead to significant progress in understanding the genetic basis of complex diseases. Despite this tremendous advance in data generation, it remains very challenging to analyze and interpret these data due to their sparse and high-dimensional nature. Here we propose several empirical Bayes scan statistics to identify genomic regions significantly enriched with rare disease risk variants. We show that the empirical Bayes methodology can be more powerful than existing methods especially so in the presence of many non-disease risk variants, and in situations when there is a mixture of risk and protective variants. Furthermore, the empirical Bayes approach has greater flexibility to accommodate covariates such as functional prediction scores and additional biomarkers. We apply the proposed methods to a whole exome-sequencing study on autism spectrum disorders and identify several new genes that reside in copy number variable regions associated with autism. In particular, genes SYNGAP1 and RNF135 are both strong candidate genes for autism and have been identified by the proposed methods.

Categories: *Association: Unrelated Cases-Controls, Sequencing Data*

P 101

Fine mapping of chromosome 5p15.33 region for lung cancer susceptibility based on a targeted deep sequencing and custom Axiom array

Linda Kachuri[1], Christopher I Amos[2], Loic LeMarchand[3], Shelley Tworoger[4], Geoffrey Liu[5], James D McKay[6], Paul Brennan[6], John K Field[7], John R McLaughlin[8], Yafang Li[2], Robert E Denroche[9], Philip C Zuzarte[9], John McPherson[9], Rayjean J Hung[1]
[1]Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada
[2]Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA
[3]University of Hawaii, Honolulu, HI, USA Shelley
[4]Harvard School of Public Health, Boston, MA, USA
[5]Ontario Cancer Institute, Princess Margaret Cancer Center, Toronto, ON, Canada
[6]International Agency for Research on Cancer, Lyon, France
[7]Institute of Translational Medicine, University of Liverpool, Liverpool, UK
[8]Public Health Ontario, Toronto, ON, Canada
[9]Genome Technologies, Ontario Institute for Cancer Research, Toronto, ON, Canada

Background: Genome-wide association studies have consistently linked single nucleotide polymorphisms (SNPs) in ch5p15.33 with increased lung cancer risk. This region contains two known cancer susceptibility genes: telomerase reverse transcriptase (TERT) and cleft lip and palate transmembrane 1-like (CLPTM1L), however, the causal mechanisms underlying these risk variants have not been fully elucidated. Methods: We carried out a fine mapping of 5p15.33 first by deep sequencing 288 lung cancer case-control pairs, and subsequently genotyping 4608 SNPs (1125 de novo variants: 953 SNPs, 172 indels not previously described) using a custom Affymetrix Axiom array in 3063 cases and 2940 controls of European ancestry from 5 studies: MSH-PMH, EPIC, MEC, LLPC, HPFS & NHS. Odds ratios (OR) adjusted for age, sex and cigarette pack-years were estimated using logistic regression. Sequence kernel association tests (SKAT) were used to localize the effects of rare variants. Results: 17 SNPs met the multiple testing corrected threshold ($p<4.1\times10\text{-}4$). Of these, two newly identified variants were strongly associated with lung cancer risk after conditioning on the effects of known risk variants: ch5:1253720 (OR: 0.25, $p=6.6\times10\text{-}6$) located in the TERT exon, and ch5:1384599 (OR: 0.03, $p=3.1\times10\text{-}4$) downstream of CLPTM1L. 13 of the 17 significant SNPs were located in CLPTM1L. The SKAT analysis points to risk variants within the TERT exon ($p=8.8\times10\text{-}4$), downstream of CPTM1L ($p=1.5\times10\text{-}4$) and microRNA4457 ($p=4.7\times10\text{-}4$). Conclusions: In this study we identified several novel variants that were independently and significantly associated with lung cancer risk. Our findings refined the association between the TERT/CLPTM1L region and lung cancer risk.

Categories: *Association: Unrelated Cases-Controls, Cancer, Fine Mapping, Sequencing Data*

P 102

Genetic variants in inflammation-related genes and interaction with NSAID use on colorectal cancer risk and prognosis

Yesilda Balavarca[1], Nina Habermann[1], Dominique Scherer[1], Katharina Buck[1], Petra Seibold[2], Katja Butterbach[3], Barbara Burwinkel[4], Katrin Pfuetze[4], Michael Hoffmeister[3], Elisabeth Kap[2]
[1]Division of Preventive Oncology, National Center for Tumor Diseases (NCT/DKFZ), Heidelberg, Germany
[2]Division of Cancer Epidemiology, Unit of Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[3]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany
[4]Molecular Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

Inflammation has been shown to contribute to colorectal carcinogenesis. Non-steroidal anti-inflammatory drugs (NSAIDs) are associated with reduced inflammation. Thus, we studied association of inflammation-related genes and their interaction with NSAID use regarding colorectal cancer (CRC) risk and survival. We analyzed 15 genes (169 SNPs) of 1756 CRC patients and 1781 controls enrolled in a case-control study with follow-up of patients (DACHS). CRC risk was assessed by multivariable unconditional logistic regression and overall survival by multivariable Cox regression models. P values for non-candidate SNPs were adjusted for multiple testing (p adj.) CRP (rs1205, p=0.04; rs1800947, p=0.02) and PTGS1 (rs10513402, p adj.=0.01) variants were associated with increased CRC risk. Subjects with the variant allele in CRP (rs1800947, p=0.004) and in PTGIS (rs477627, p=0.04), respectively, showed lower CRC risk with the use of NSAIDs. After 5-years follow up, variants in PTGS1 were associated with poorer overall survival (rs1330344, p adj.=0.02 and rs3119773, p adj.=0.04). NSAID use was associated with improved overall survival of patients with the variant allele in CCL2 (rs3760396, p. adj=0.01) and with decreased overall survival of patients with the variant allele in IL23R (rs12041056, p=0.008). In patients with disease stage I-III, those with variants in IL18 (rs1293344, p adj.=0.01) and in IL23R (rs10889665, p adj.=0.02) showed improved overall survival. We showed that genetic variations in inflammation-related genes and their interactions with NSAIDs are associated with CRC risk and survival. This information may aid in tailoring prevention strategies to subjects who will benefit most from NSAID use.

Categories: *Association: Unrelated Cases-Controls, Cancer, Gene - Environment Interaction, Multifactorial Diseases*

P 103

Association analysis of exome chip data of Polycystic Ovary Syndrome in Estonian Biobank

Reedik Mägi[1], Andrew P Morris[2], T Karaderi[3], Triin Triin Laisk-Podar[4], Triin Tammiste[4], Andres Metspalu[1], Andres Salumets[4], Cecilia M Lindgren[5]
[1]Estonian Genome Center, University of Tartu, Tartu, Estonia
[2]Department of Biostatistics, University of Liverpool, Liverpool, UK
[3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[4]Department of Obstetrics and Gynaecology, University of Tartu, Tartu, Estonia
[5]Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA

Polycystic ovary syndrome (PCOS) is a common multifactorial disease affecting up to 10% of women of reproductive age. To investigate the contribution of potentially causal coding variants to PCOS, we have genotyped 167 cases and 711 population controls (363 females) from the Estonian Biobank with the Illumina exome array. We conducted single variant and burden tests of association using SKAT-O within genes for (i) loss of function (LOF) and (ii) rare non-synonomous (NS) variants with minor allele frequency (MAF) <1%). The association analyses were adjusted for first two principal components to account for the population stratification. In the autosomal analysis, both male and female samples were used in the control group but in the X chromosome analysis, only female samples were used. Altogether 55,345 polymorphic variants were successfully tested in single variant analysis. It revealed one missense variant which was showing exome-wide evidence of association (p<5x10-7, Bonferroni correction for 100,000 variants): exm233350 in the nebulin coding NEB gene (p=4.9 x 10-9, MAF=0.05%). Mutations in NEB have previously been associated with myopathy and muscle structure. None of the associations were statistically significant in the gene-based tests after multiple testing correction for 20,000 genes. The strongest associations came from aggregating non-synonymous rare variants within POLK (p=4.3 x 10-5) and PELI3 (7.3 x 10-5) gene, which are DNA replication and immune response related genes. Our study suggests that rare variants can contribute to the genetic component of PCOS, but cannot explain previously reported association signals in established GWAS loci.

Categories: *Association: Unrelated Cases-Controls, Case-Control Studies*

P 104

A MODEL FOR CO-SEGREGATION OF CRYPTORCHIDISM AND TESTIS CANCER IN FAMILIES

Duncan C Thomas[1], Victoria K Cortessis[1]
[1]University of Southern California

Testicular germ cell tumors (TGCT) and cryptorchidism (CO) are highly familial, but loci identified by genome-wide association studies explain only 15-22% of TGCT heritability. To understand segregation of the traits and dependency of TGCT on CO, we developed a novel statistical model incorporating major genes, polygenes, and nongenetic frailties accounting for dependence between testes, for each trait and the transition from CO to TCGT. From 17,844 TCGT cases in the California Cancer Registry, we obtained informed consent and personal and family history from 5,702 (17,844 family members), and extended pedigrees for 697 of those cases reporting bilateral TGCT, CO, or family history of either trait (23,143 members). Adjusting for this complex ascertainment, we found strong evidence for polygenic effects for CO and TCGT and a major gene modifying the effect of CO on TCGT risk. Genotypes for 9 TGCT risk variants in 1,639 members of 527 families were used in an extended model incorporating multiple genes assumed to be in LD with the SNPs and to segregate without recombination. This revealed significant associations of CO with TERT and CENPE, baseline TGCT risk with KITLG and UCK2, and suggestive evidence that TERT modifies the effect of CO on TGCT. These support a genetic basis for familial aggregation of both traits and dependency between traits. The model can address other precursors (e.g. polyps for colorectal cancer, mammographic density for breast cancer).

Categories: *Ascertainment, Association: Candidate Genes, Association: Family-based, Bayesian Analysis, Cancer, Familial Aggregation and Segregation Analysis, Linkage and Association, Markov Chain Monte Carlo Methods*

P 105

Joint analysis of secondary phenotypes: an application in family studies

Renaud R Tissier[1], Roula S Tsonaka[1], Jeanine J Houwing-Duistermaat[1]
[1]LUMC, The Netherlands

A case-control design is typically used in order to test associations between the case-control status (primary phenotype) and genetic variants. In addition to this primary phenotype secondary traits are available and associations are studied between the genetic variants and these secondary traits. However, when analyzing these phenotypes the case-control design has to be taken into account especially when the marker tested is associated with the primary phenotypes or when there is correlation between the primary and secondary phenotype. Methods are available for secondary phenotype analysis in case-control studies. These methods are not directly applicable to more complex designs, such as multiple cases family studies. Here a proper secondary analysis is complicated by the biased sampling design, the within families correlations and the mixed type of outcomes: binary primary phenotype and continuous secondary phenotypes. We propose a bias correction approach for secondary phenotype analysis in family studies which allows investigation of genetic effects across multiple secondary traits. We adopt the retrospective likelihood method to correct for ascertainment of the families and use a correlated probit model to model jointly the mixed type primary and secondary phenotypes. The estimates of the parameters can be pooled with results from studies comprising randomly selected subjects by standard meta analysis tools. We studied the performance via simulations and estimated the effects of several SNPs on a triglyceride available in the Leiden Longevity Study, a case family-control study. We conclude that the use of an ad-hoc will lead to bias especially in case the SNP is associated with the primary phenotype

Categories: *Ascertainment, Association: Candidate Genes, Association: Family-based, Association: Unrelated Cases-Controls*

P 106

Prediction of imprinted genes based on the genome-wide methylation analysis

Natalia Tšernikova[1], Neeme Tõnisson[2], Kaie Lokk[2], Andres Salumets[3], Andres Metspalu[1], Reedik Mägi[4]

[1]Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia and Estonian Genome Center, University of Tartu, Tartu, Estonia

[2]Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

[3]Competence Centre on Reproductive Medicine and Biology, Tartu, Estonia; Department of Obstetrics and Gynecology, University of Tartu, Tartu, Estonia and Institute of Biomedicine, University of Tartu, Tartu, Estonia

[4]Estonian Genome Center, University of Tartu, Tartu, Estonia;

Genomic imprinting is an epigenetic gene-marking phenomenon that is established in germline. Our hypothesis is that imprinted genes can be predicted by the methylation level. We expect semi-methylation in imprinted genes. In order to prove this hypothesis we analysed the DNA methylation in well-known imprinted genes across the tissue panel from the same individuals. 17 tissues from 4 individuals were collected during autopsy. DNA methylation analysis of the total 72 tissue samples was performed with the Illumina Infinium HumanMethylation450 BeadChip. We used Levene's test for comparison of known imprinted genes with the rest of the genes captured by 450K methylation array. As a result, all imprinted genes (n=92) demonstrated less variability in the methylation level ($p < 0.01$) across all tissues. We also visualized CpG patterns of known imprinted genes across all tissues. Each CpG was annotated to its exact location in the genome in exon, gene body or UTR region. Visualized CpG patterns also confirmed tissue-specific nature of imprinted genes. For example, gallbladder shows medium methylation of KCNQ1DN gene, while in isciatic nerve the CpG sites are not methylated. Using this mapping method, we narrowed down the list of potential candidate genes to 3 000. We found that some genes meet the criteria for candidate imprinted genes in all somatic tissues, while other genes meet those criteria only in some of the tissues. As the next step we are using the RNAseq data to further narrow down the list of candidate genes. Our method can be regarded as a tool to identify the tissue specificity of the already established imprinted genes as well as to discover new imprinted genes across the whole human genome.

Categories: *Bioinformatics, Epigenetic Data, Epigenetics*

P 107

Addiction and Mental Health Genes form Genomic Hotspots with Drugable Targets.

Latifa F Jackson[1], Aydin Tozeren[1]
[1]Drexel University

Dopamine, alcohol and opiate addiction are well characterized co-morbidities with depression, bipolar and schizophrenia disorders. While each of these disorders are known to have a strong genetic basis, there is little systematic information that addresses the genetic intersections of these co-morbid disorders. We can harness curated gene sets derived from single gene and genome wide association studies to identify the genomic regions disproportionately participating in addiction and mental health disorders. Opiate, dopamine and alcohol addiction disorder gene sets and mental health gene sets (schizophrenia, depression and bipolar disorder) obtained from National Center for Biotechnology Information Gene, were combined, then projected onto the genome, and the regions of interest were annotated with gene ontology categories and cellular pathways. functional annotations among the resulting addiction and mental health genomic hotspot regions form a bioinformatics portrait of the genetic intersections likely to contribute to observed co-morbidities. We identify eight genomic hotspots, with an overabundance of addiction and mental health genes ($p<0.005$). Hotspot genes and their co-located candidate counterparts are involved in significant core neurological functions ($p=0.05$): neurological transmission, responses to organic substances and cell-cell signaling. We further annotated all hotspot genes for their associated drug binding sites to identify whether these binding sites were candidates for therapeutic intervention. We found 16 drug binding sites, which sorted into four thematic classes: an illicit drug binding site, four mental health drug sites, three immune response sites, and five cancer binding sites with strong addiction or mental health counter-indications. Our analyses demonstrate the utility of considering a hotspot approach in identifying genomic regions contributing to the intersection of addiction and mental health and provide gene candidates for potential drug targets. Keywords: Bioinformatics, Addiction, Schizophrenia, Bipolar Disorder, Depression, Genome, Alcohol, Causal Inference, Candidate Genes

Categories: *Bioinformatics, Gene - Gene Interaction, Genomic Variation*

P 108

Recurrent shared rare variants in 9 genes detected by whole exome sequencing of multiplex oral clefts families

Joan E Bailey-Wilson[1], Emily R Holzinger[1], Qing Li[1], Margaret M Parker[2], Jacqueline B Hetmanski[2], Mary L Marazita[3], L Leigh Field[4], Ajit Ray[5], Elisabeth Mangold[6], Markus M Nöthen[6]
[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD, USA
[2]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
[3]Center for Craniofacial and Dental Genetics, Department of Oral Biology, University of Pittsburgh, Pittsburgh, PA USA
[4]Emeritus, University of British Columbia, Vancouver, BC Canada
[5]Emeritus, University of Toronto, Toronto, Ontario, Canada
[6]Institute of Human Genetics, University of Bonn, Bonn, Germany

Non-syndromic cleft lip with/without cleft palate (CL/P) is a complex trait. Genome-wide association studies (GWAS) have identified several genetic risk factors for CL/P and recently we identified a novel, potentially damaging variant in CDH1 in one Indian multiplex CL/P family. Here, we used whole exome sequence (WES) data on 2 or 3 related (2o or more distant) affected individuals per family to identify genes containing shared rare variants (RV). Fifty-five families of Indian (12), Filipino (11), German (19), Syrian (10), European-American (1) and Asian (2) descent containing 114 individuals, 4 duplicate controls and 2 unrelated CEPH HAPMAP controls were sequenced on the Illumina Hi-Seq 2500 and processed through GATK. Ingenuity 'Variant Analysis' was used to identify RVs shared under a recessive model by all sequenced affected individuals in a family. Genes where such sharing was observed in the same gene in at least 2 separate multiplex families (different RVs per family) were considered potentially related to CL/P. After filtering based on variant quality and frequency (MAF<0.05), we identified 9 genes exhibiting recessive RV sharing in all affected individuals in at least two families: ARHGEF12, CCT4, HSD3B7, MAN1B1, RREB1, SNRPC, STARD9, ZDHHC11, and ZNF835. These RVs are not present in either of the sequenced HapMap controls. Follow-up will include Sanger sequencing and genotyping of these RVs in other affected and unaffected individuals in these same families to determine if the RVs segregate with disease.

Categories: *Bioinformatics, Data Mining, Genomic Variation, Sequencing Data*

P 109

Evaluation of variant calling from thousands of low pass whole genome sequencing (WGS) data using GATK haplotype caller

Hua Ling[1], Kurt Hetrick[1], Peng Zhang[1], Elizabeth Pugh[1], Jane Romm[1], Kimberly Doheny[1]
[1]Center for Inherited Diseases Research (CIDR), Johns Hopkins University

Low pass WGS (~2-8x) on large numbers of samples has become an attractive strategy in genetic studies of complex traits. Given the same amount of sequence yield, it may provide more power in detecting disease associated variants than deep sequencing (30x) a smaller number of samples. It can also be used to build a reference panel for imputing additional samples to further boost power. Recent advances in GATK enable us to do joint variant calling and analysis on multiple WES samples using Haplotype Caller (HC) that was computationally prohibitive. HC is desirable over Unified Genotyper (UG) not only because of its higher accuracy in variant calling (especially for INDELs), but it offers greater flexibility by allowing for adding in more samples at a later stage without re-processing the cohort. To evaluate the feasibility and performance of calling thousands of low pass WGS samples under current hardware and software supports, we used 2,535 low pass WG BAM files from the 1KGP for chr11. We generated genomic VCF files for each individual sample with HC, and created joint calls using GenotypeGVCFs followed by variant filtering with VQSR. The mean coverage per sample ranged from 2.8 to 38 (median of 6.67), with 70% samples have mean coverage below 8x. More than 96% and 85% bases have depth greater than 2X and 4X, respectively. For NA12878 (mean coverage of 4.91), low pass WGS made ~75% of SNVs on chr11 that are called by GenomeInABottle (v2.18). By comparing this to array data and 30x WGS generated onsite for a subset of samples and reviewed using IGV, we can characterize sensitivity and concordance at different levels of MAF and sequencing depth for both SNVs and INDELs.

Categories: *Bioinformatics, Data Quality, Genomic Variation, Sequencing Data*

P 110

Integration of fMRI and SNPs indicated potential biomarkers for Schizophrenia diagnosis

Hongbao Cao[1], Yu-PingWang[2], Vince Calhoun[3], Yin Yao Shugart[1]
[1]National Institute of Health
[2]Tulane University
[3]University of New Mexico

Integrative analysis of multiple data types can take advantage of their complementary information and therefore may provide greater power to identify potential biomarkers. However, due to the diversity of the data modality, data integration is challenging. Here we address the data integration problem by developing a generalized sparse model (GSM) using weighting factors to integrate multi-modality data for biomarker selection. To prove the feasibility, we applied the GSM model to a joint analysis of two types of schizophrenia data sets: 759075 SNPs and 153594 functional magnetic resonance imaging (fMRI) voxels in 208 subjects (92 cases/116 controls). To solve this small-sample-large-variable problem, we developed a novel sparse representation based variable selection (SRVS) algorithm, aiming to identify biomarkers associated with schizophrenia. To validate the effectiveness of the selected variables, we performed multivariate classification followed by a ten-fold cross validation. Results showed that our proposed SRVS method can be used to identify novel biomarkers and offer stronger capability in distinguishing schizophrenia patients from healthy controls. Moreover, better classification ratios were achieved using biomarkers from both types of data, suggesting the importance of integrative analysis. Especially, with norm based penalty, our SRVS method generated highest classification accuracy in discriminating schizophrenia patients from healthy controls. This suggests that norm may be the best choice as penalization term for the proposed SRVS method. Further biological experimental work is needed to validate the biomarkers identified in the paper.

Categories: *Bioinformatics, Case-Control Studies, Data Integration*

P 111

EWAS to GxE: A robust strategy for detecting gene-environment interaction models for age-related cataract

Molly A Hall[1], John R Wallace[1], Sarah A Pendergrass[1], Richard Berg[2], Terrie Kitchner[2], Peggy Peissig[2], Murray Brilliant[2], Catherine A McCarty[3], Marylyn D Ritchie[1]
[1]Center for Systems Genomics, The Pennsylvania State University, University Park, PA
[2]Marshfield Clinic, Marshfield WI
[3]Essentia Rural Health, Duluth, MN

Gene-environment interactions (GxE) are essential to elucidating the nature of complex traits, but computational demands and multiple testing make uncovering these interactions difficult. We address this using an environment wide association study (EWAS) to identify putative environmental factors in a high-throughput manner followed by a test for GxE with genome-wide SNPs for association with cataract. We performed a dietary EWAS by evaluating 57 dietary exposures from a Dietary History Questionnaire using logistic regression, adjusted for age, sex, and type 2 diabetes (T2D) in 2,629 samples (932 controls, 1,697 cases) of European descent from the Marshfield Clinic Personalized Medicine Research Project, part of the Electronic Medical Records & Genomics (eMERGE) Network. Seven dietary measures were predictive of cataract (p-value < 0.05); a monounsaturated omega-9 fatty acid, erucic acid (FA22:1) (p=5.5×10-4) passed our Bonferroni corrected p-value threshold. We then tested FA22:1 for GxE using 498,829 SNPs in a subset of samples for whom genetic data was available (831 controls, 1,511 cases) using logistic regression adjusted for age, sex, and T2D status. Twenty SNP-FA22:1 models were significant (p < 1.0×10-4). The most significant GxE model was FA22:1 and rs726712, an intronic SNP in LPP (p=2.9×10-5). The erucic acid-cataract association is novel; although two polyunsaturated fatty acids have been found in cataractous human lenses. LPP encodes a protein involved in cell-cell adhesion, a process with multiple published associations with cataract. These findings indicate the role of GxE in susceptibility to cataract and demonstrate the utility of EWAS for investigating the GxE interplay of complex diseases.

Categories: *Bioinformatics, Case-Control Studies, Gene - Environment Interaction*

P 112

RNA-seq analysis of lung adenocarcinoma reveals differential gene expression in nonsmoker and smoker patients

Yafang Li[1], Xiangjun Xiao[1], Christopher I Amos[1]
[1]Dartmouth college

Lung adenocarcinoma is a complex disease that caused by both genetic and environmental effect. The RNA-seq technology provides us a powerful tool for transcriptome analysis of lung cancer. In this study, we used R Bioconductor edgeR to analyze RNA-seq from paired normal and tumor tissue in 34 nonsmoker and 40 smoker patients with lung adenocarcinoma (GEO: GSE40419). We divided the samples into pilot and replication study for each group, and there is high consistence between the results from replication and pilot studies. The gene differential expression analysis identified 179 genes that showed differential expression only in tumors from nonsmoker patients; 780 genes that are differentially expressed in both smoker and nonsmoker tumor tissue versus normal tissue; and 1869 genes that exclusively varied in tumor tissue from smoker patients versus normal tissue. 77% and 59% of the identified genes are down regulated in nonsmoker and smoker groups, respectively. Among the common genes, the genes tend to have a larger logFC change in smoker patients than nonsmoker patients. The smoker and nonsmoker patient specific genes with large logFC are also identified in our analysis. Our study provides a systematic analysis of whole genome gene differential expression. It provides target genes for subsequent biological studies to decipher the aberrations that are present in lung adenocarcinoma.

Categories: *Bioinformatics, Cancer, Case-Control Studies, Gene Expression Patterns, Sequencing Data*

P 113

Using random forests to identify genetic links between Alzheimer's disease and type 2 diabetes

Burcu F Darst[1], Chen Yao[2], Rebecca L Koscik[3], Barbara B Bendlin[4], Bruce P Hermann[3], Asenath La Rue[3], Sterling C Johnson[5], Mark A Sager[3], Corinne D Engelman[1]
[1]Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA
[2]Department of Dairy Science, University of Wisconsin, Madison, WI, USA
[3]Alzheimer's Diseases Research Center, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA; Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA
[4]Alzheimer's Diseases Research Center, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA; Geriatric Research Education and Clinical Center, Wm. S. Middleton Memorial VA Hospital, Madison, WI, USA
[5]Alzheimer's Diseases Research Center, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA; Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA; Geriatric Research E

Increasing evidence suggests that type 2 diabetes (T2D) is a risk factor for Alzheimer's disease (AD), but the genetic mechanism linking these conditions is unknown. Using random forests (RF), we investigated whether interactions between single nucleotide polymorphisms (SNPs) in a pathway linked to both AD and T2D, and risk factors for T2D influence cognition in a cohort of middle-aged adults enriched for a parental history of AD. We analyzed a sample of 836 participants from the Wisconsin Registry for Alzheimer's Prevention with data on predictors of T2D, 30 SNPs in the SORL1 and SORCS1 genes, and 4 cognitive factors. These variables were input into RF, a machine-learning algorithm that calculates importance scores based on the variance explained by each variable in a model while allowing for interactions. Because RF does not specifically identify interacting variables, we used a novel approach that identifies interactions by determining how often a pair of variables descends together in a RF. Rs7907690 in SORCS1,and rs2282649 and rs1010159 in SORL1,appeared in the top 25 descendant pairs for all 4 cognitive factors, frequently paired with waist-hip ratio, HOMA-IR (a measure of insulin resistance), age, and physical activity. Many of the interactions identified with descendant pairs consisted of discordantly ranked pairs, with one variable having a high importance score and the other having a low importance score. These results suggest that interactions between SNPs associated with AD and T2D and risk factors for T2D may contribute to the relationship between AD and T2D and that the descendant pair method captures interactions that the standard RF method does not.

Categories: *Bioinformatics, Diabetes, Gene - Environment Interaction, Machine Learning Tools, Psychiatric Diseases*

P 114

Study oF Human MGP promoter variants in CAD patients: From Experiment to prediction

Bita Sadat Hosseini[1], Abazar Roustazadeh[2], Mohammad Najafi[3]
[1]Biochemistry Department, Iran University of Medical Sciences, Tehran, Iran
[2]Jahrom University of Medical Sciences, Jahrom, Iran
[3]Biochemistry Department, Cellular and Molecular Research Center, Iran University of Medical Sciences, Tehran, Iran

Background: Matrix Gla protein (MGP) is known as a calcium scavenger within sub-endothelial space of vessels and is suggested to reduce the risk of coronary artery diseases. In this study, we compared the MGP promoter high minor allele frequency (MAF) variants and the changes on the predicted transcription factor elements in patients with coronary artery disease. Methods: The MGP promoter genotypes and haplotypes were detected by ARMS-RFLP PCR techniques. The Jaspar profiles (similarity >80) were used for scoring the polymorphic variants within the transcription factor elements. Results: The MGP polymorphic haplotypes and genotypes had not significant differences between control and patient groups (P=0.4 and P=0.1 respectively). Furthermore, the results showed that the genotype and haplotype distributions of the MGP promoter high-MAF polymorphisms, as confirmed in the prediction studies are not significantly associated with the coronary artery disease. Discussion: The prediction and population results showed that the allele changes within the elements have not significantly related to the transcription factor scores and stenosis of coronary arteries.

Categories: *Bioinformatics, Cardiovascular Disease and Hypertension, Haplotype Analysis*

P 115

A novel functional data analysis approach to detecting gene by longitudinal environmental exposure interaction

Peng Wei[1]
[1]University of Texas School of Public Health

Most complex diseases are likely the consequence of the joint actions of genetic and environmental factors. Identification of gene-environment (GxE) interactions not only contributes to a better understanding of the disease mechanisms, but also improves disease risk prediction and targeted intervention. In contrast to the large number of genetic susceptibility loci discovered by genome-wide association studies, there have been very few successes in identifying GxE interactions which may be partly due to limited statistical power and inaccurately measured exposures. While existing statistical methods only consider interactions between genes and static environmental exposures, many environmental factors, such as air pollution and diet, change over time, and cannot be accurately captured at one measurement time point. There is a dearth of statistical methods for detecting gene by time-varying environmental exposure interactions. Here we propose a powerful functional logistic regression (FLR) approach to model the time-varying effect of longitudinal environmental exposure and its interaction with genetic factors on disease risk. Capitalizing on the powerful functional data analysis framework, our proposed FLR model is capable of accommodating longitudinal exposures measured at irregular time points and contaminated by measurement errors. We use simulations to show that the proposed method can control the Type I error and is more powerful than alternative ad hoc methods. We demonstrate the utility of this new method using data from a case-control study of pancreatic cancer to identify the windows of vulnerability of lifetime body mass index on the risk of pancreatic cancer as well as genes which may modify this association.

Categories: *Cancer, Gene - Environment Interaction*

P 116

Leveraging Family Structure for the Analysis of Rare Variants in Known Cancer Genes from WES of African American Hereditary Prostate Cancer

Cheryl D Cropp[1], Shannon K McDonnell[2], Sumit Middha[2], Danielle Karyadi[3], Stephen N Thibodeau[4], Janet Stanford[5], Kathleen A Cooney[6], Joan E Bailey-Wilson[1], John D Carpten[7], for the International Consortium of Prostate Cancer Genetics

[1]Computational and Statistical Genomics Research Branch, National Human Genome Research Institute/National Institutes of Health, Baltimore, MD

[2]Departmant of Health Science Research, Mayo Clinic, Rochester, MN

[3]Cancer Genetics Branch, National Human Genome Research Institute/National Institutes of Health, Bethesda, MD

[4]Departmant of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN

[5]Public Health Sciences Division, Epidemiology Program, Fred Hutchinson Cancer Research Center, Seattle, WA

[6]University of Michigan Comprehensive Cancer Center, Ann Arbor, MI

[7]Integrated Cancer Genomics Division, Translational Genomics Research Institute (TGen), Phoenix, AZ

"Leveraging Family Structure for the Analysis of Rare Variants in Known Cancer Genes from WES of African American Hereditary Prostate Cancer" Prostate cancer (PRCA) is the second leading cause of cancer death in North American men and it disproportionately affects African American (AA) men, who have higher incidence and mortality rates compared to men without known African ancestry. Disentangling the environmental and genetic factors in AA with hereditary PRCA remains elusive. The African American Hereditary Prostate Cancer Study (AAHPC) was developed to further explore the role of genetics in the causation of hereditary PRCA in AA. AAHPC is in partnership with the International Consortium for Prostate Cancer Genetics (ICPCG) to conduct collaborative studies in PRCA genetics in multiplex families. As part of an ICPCG sequencing study of 539 affected individuals from 366 PRCA pedigrees, we performed whole exome sequencing on 16 AAHPC affected men from 12 pedigrees. Post-variant calling quality control was implemented using Golden Helix SVS 8 software with filters set for removal of variants with Read Depth < 10, Quality Score < 20, Quality Score:Read Depth Ratio < 0.5, Call Rate < 0.75. Variants were additionally filtered by minor allele frequency (MAF) based on the NHLBI ESP650051-V2 exomes variant frequencies for AA population using a MAF threshold of 1%. After QC, 174,047 variants remained for further analysis. In these analyses, we focused on 13 known cancer causing genes. Two AAHPC families had > 1 affected members sequenced (3 per family). Under a dominant model, Family 1 shared 14 variants in these genes among all affecteds while Family 2 shared 17 variants among all affected men. Additional studies are underway to determine if predicted damaging variants in these genes are shared in other ICPCG AA families to help unravel the genetic heterogeneity of hereditary PRCA in AA.

Categories: *Cancer, Data Mining, Genomic Variation, Sequencing Data*

P 117

Association of breast cancer risk loci with survival of breast cancer patients

Myrto Barrdahl[1], Federico Canzian[2], Sara Lindström[3], Irene Shui[3], Rudolf Kaaks[1], Daniele Campa[1]
[1]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]Genomic Epidemiology Group. German Cancer Research Center (DKFZ), Heidelberg, Germany
[3]Department of Epidemiology, Harvard School of Public Health , Boston MA, USA

The survival of breast cancer patients is largely influenced by tumor characteristics, such as TNM stage, tumor grade and hormone receptor status. However, there is growing evidence that inherited genetic variation might influence the disease prognosis and response to treatment. Several lines of evidence suggest that polymorphisms influencing breast cancer risk might also be associated with breast cancer survival. With the aim of further exploring this possibility, we selected 35 polymorphisms associated with breast cancer risk and investigated their role in the disease over-all survival. We studied 10,255 breast cancer patients from the National Cancer Institute Breast and Prostate Cancer Cohort Consortium (BPC3) of which 1,379 had fatal breast cancer. We also conducted a meta-analysis of almost 35,000 patients and 5,000 deaths, combining results from the current study and from the Breast Cancer Association Consortium (BCAC). In BPC3 we observed a significant association between the C allele of LSP1-rs3817198 and reduced death hazard (HRper-allele=0.70; 95% CI: 0.58-0.85; Ptrend=2.84×10-4). This association was supported by the observation that the C allele of this SNP increases the expression of the tumor suppressor cyclin-dependent kinase inhibitor 1C (CDKN1C). The meta-analysis showed a significant association between TNRC9-rs3803662 and an increased death hazard (HRMETA=1.21; 95% CI: 1.09-1.35; P=2.47×10-4 comparing homozygotes for the minor allele vs. homozygotes for the major allele). In conclusion, we show that there is little overlap between SNPs associated with breast cancer risk and SNPs associated with breast cancer prognosis, with the possible exceptions of LSP1-rs3817198 and TNRC9-rs3803662.

Categories: *Cancer*

P 118

Evidence of gene-environment interactions in relation to breast cancer risk, results from the Breast Cancer Association Consortium

Myrto Barrdahl[1], Anja Rudolph[1], Nick Orr[2], Paul Pharoah[3], Per Hall[4], Montserrat Garcia-Closas[5], Marjanka Schmidt[6], Roger Milne[7], Doug Easton[8], Jenny Chang-Claude[1]
[1]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, London, UK
[3]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[4]Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[5]Sections of Epidemiology and Genetics, Institute of Cancer Research and Breakthrough Breast Cancer Research Centre, London, UK
[6]Division of Molecular Pathology and Division of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[7]Genetic and Molecular Epidemiology Group, Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[8]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland, USA

Several new susceptibility alleles for breast cancer (BC) risk have been identified by the Breast Cancer Association Consortium (BCAC) through imputation of genetic data to 1000Genomes and fine-mapping of known susceptibility loci. We investigated whether the identified single nucleotide polymorphism (SNP) associations are modified by established BC risk factors. We assessed multiplicative interaction between 12 BC risk factors and 74 SNPs, of which 54 were imputed (from 17 known regions) and 29 genotyped (in 22 regions). We used data from up to 25,539 invasive BC cases and 29,664 controls from 21 studies in BCAC. The risk factors were: age at menarche, parity, number of full-term pregnancies (FTP), age at first FTP, breastfeeding, BMI, height, oral contraceptive use, current postmenopausal hormone use (estrogen and estrogen-progesterone), current smoking and cumulative lifetime alcohol intake. Interactions between SNPs and BC risk factors were evaluated using likelihood-ratio tests to compare logistic regression models with and without interaction terms. All models were adjusted for study, age and ancestry informative principal components. We found a suggestive interaction between a SNP in the 9q31 region and current smoking (Pinteract=5.3×10-5), which was significant after Bonferroni correction of the significance threshold (P<5.6×10-5). In particular, the G-allele was inversely associated with BC risk among smokers (ORper-allele: 0.68, 95% CI: 0.57-0.81, P=1.7×10-5) but not among non-smokers (ORper-allele: 0.95, 95% CI: 0.89-1.03, P=0.2). In conclusion, the findings of our study provide indications that the association between common genetic variants and BC risk may vary across the levels of the BC risk factors.

Categories: *Cancer, Gene - Environment Interaction*

P 119

Integration of pathway and gene-gene interaction analyses reveal biologically relevant genes for Breslow thickness, a major predictor of melanoma prognosis

Amaury Vaysse[1], Shenying Fang[2], Myriam Brossard[1], Wei V Chen[2], Hamida Mohamdi[1], Eve Maubec[1], Marie-Françoise Avril[3], Christopher I Amos[4], Jeffrey E Lee[5], Florence Demenais[1]
[1]INSERM U946, Paris, France; Université Paris Diderot, Paris, France;
[2]MD Anderson Cancer Center (MDACC), Houston, Texas, USA;
[3]Hôpital Cochin, Université Paris Descartes, Paris, France
[4]Geisel College of Medicine, Dartmouth College, New Hampshire, USA
[5]MD Anderson Cancer Center (MDACC), Houston, Texas, USA

Breslow thickness (BT), a measure of invasion of melanoma in the skin, is a major predictor of melanoma survival. To date, the genetic factors underlying BT are largely unknown. We conducted a GWAS of BT in the French MELARISK study (966 cases) and the US MDACC study (1546 cases). We first performed single-SNP analysis that was followed by multi-marker analysis to characterize pathways and gene-gene interactions associated with BT. Pathway analysis was based on the gene set enrichment analysis (GSEA) method, using the Gene Ontology (GO) database. All gene pairs within each of the melanoma-associated GOs were tested for interaction using a linear regression model. Single SNP analysis of Hapmap3-imputed SNPs in MELARISK showed evidence for five loci that reached P<10-5 but none of these associations was replicated in MDACC, suggesting the existence of many variants with small effect. In the GSEA analysis, one million imputed SNPs were assigned to 22,000 genes, which were assigned to 316 Level 4-GO categories. Three GO categories were found to be enriched in genes associated with BT (FDR≤0.05 in both studies): hormone activity, cytokine activity and myeloid cell differentiation. A total of 61 genes were driving these pathways. Interestingly, expression of four of these genes (CXCL12, TNFSF10, VEGFA, CDC42) was reported to be associated with melanoma progression in tumors. Cross-gene SNP-SNP interaction analysis within each of the three identified GOs showed evidence for interaction for three SNP pairs (P≤10-4 in MELARISK and replication at P≤ 0.05 in MDACC). One of these gene pairs (SCINxCDC42, combined P=2x10-6) has biological relevance since SCIN and CDC42 proteins are involved in the actin dynamics with opposite roles. Funding: INCa_5982

Categories: *Cancer, Gene - Gene Interaction, Multilocus Analysis, Pathways, Quantitative Trait Analysis*

P 120

JAG1 polymorphism is associated with incident neoplasm in a southern Chinese population

Chor-Wing Sing[1], Vivian Wai-Yan Lui[1], Pak-Chung Sham[1], Kathryn Choon-Beng Tan[1], Annie Wai-Chee Kung[1], Ian Chi-Kei Wong[1], Bernard Man-Yung Cheung[1], Johnny Chun-Yin Chan[1], Ching-Lung Cheung[1]
[1]The University of Hong Kong

Aim: Jagged 1 (JAG1) is a ligand of notch receptors that regulates cell division, differentiation, and survival. Over-expression of JAG1 has been linked to increased risk of cancer. We previously showed that rs2273061 of JAG1 was associated with bone mineral density (BMD) using genome-wide association, and the SNP was associated with JAG1 expression in bone and blood cells. We hypothesized that this SNP has association with neoplasm. Methods: The SNP rs2273061 of JAG1 was genotyped in two independent cohorts without history of neoplasm at baseline. The cohorts were followed (median 10.8 years) for development of neoplasm using electronic medical database of the Hong Kong Hospital Authority. Ascertainment of neoplasm was based on ICD9 code 140-239. Cox proportional hazards regression models adjusted for age, sex, BMI, and lumbar spine BMD Z-score were used for association analysis. AUC was used to test predictive accuracy of the models. Result: In the first cohort (n=731; 80 incidents; 7620 person-year), minor allele (G) of rs2273061 was significantly associated with neoplasm (HR=0.68; 95%CI:0.47-0.98). The result was validated (HR=0.81; 95% CI:0.66-0.99) in replication cohort (n=1,885; 241 incidents; 19966 person-year). Meta-analysis showed a more significant association (HR=0.78; 95% CI:0.65-0.93; p=0.005). AUC for basic clinical model (age+sex+BMI) in predicting neoplasm was 0.618 (95%CI: 0.588-0.648). The addition of rs2273061 genotype to the basic model increased AUC to 0.635 (95%CI: 0.605-0.665), and the increment was statistically significant. Conclusion: JAG1 polymorphism has association with incident neoplasm However, further study is required to evaluate any functional effects of rs2273061 on tumor formation.

Categories: *Cancer*

P 121

Epigenome-wide methylation array analysis reveals few methylation pattern differences between hyperplastic polyps and sessile serrated adenomas/polyps

Jing Li[1], Angeline S Andrew[1], Amitabh Srivastava[2], Jason H Moore[1]
[1]Institute for Quantitative Biomedical Sciences, Dartmouth College
[2]Brigham and Women's Hospital

The colorectal 'serrated polyps' arise via a neoplastic pathway and were historically not considered with malignant potential. Major subtypes of serrated colorectal polyps, hyperplastic polyps (HPs) and sessile serrated adenomas/polyps (SSA/Ps), are classified based on morphological distinctions. Recent studies have identified SSA/P as a high-risk subtype of serrated colorectal polyps that can develop into colorectal cancer. Our goal was to determine whether HPs and SSA/Ps have distinct underlying DNA methylation signatures. To evaluate the subtype-specific DNA methylation status, DNAs from 35 HPs and 42 SSA/Ps were extracted and Illumina Infinium HumanMethylation450 BeadChip arrays were used to profile the methylation status for >485,000 CpG loci. Principal component analysis revealed that the top principal components, which account for the largest amount of variability of methylation status, are not significantly associated with subtype (p=0.414). Also, linear mixed-effects models showed that the methylation pattern is not significantly different between subtypes, after controlling for age, gender, polyp size, anatomic side and batch. We also compared SSA/Ps and HPs using the probes that map to the CIMP panel loci and found no statistically significant differences in methylation status by morphology. Comparing the normal vs. serrated colorectal polyps revealed 18 probes with significantly different methylation levels below the Bonferroni threshold (p=<1.06e-7). Our results suggests that dysregulated methylation is prevalent, involving a number of non-CIMP CpGs and likely occurs early in serrated neoplasia. The data do not support the hypothesis that SSA/Ps and HPs arise via different epigenetic pathways.

Categories: *Cancer, Epigenetic Data, Epigenetics*

P 122

The effect of bile acid sequestrants on the risk of cardiovascular events: A meta-analysis and Mendelian Randomization analysis

Guillaume Pare[1], Stephanie Ross[1], Matthew D'Mello[1], Sonia S Anand[1], John Eikelboom[1], Alexander FR Stewart[2], Nilesh J Samani[3], Robert Roberts[2]
[1]McMaster University
[2]University of Ottawa
[3]University of Leicester

Statins are used to lower low density lipoprotein cholesterol (LDL-C) but they may be poorly tolerated or ineffective. Bile acid sequestrants (BAS) act to reduce the intestinal absorption of cholesterol but previous trials were underpowered to demonstrate an effect on clinical outcomes. We conducted a systematic review and meta-analysis of randomized controlled trials (RCTs) to assess the effect of two approved BAS, cholestyramine and colesevelam, on plasma lipid levels. We then applied the principles of Mendelian Randomization to estimate the effect of BAS on reducing the risk of coronary artery disease (CAD) by quantifying the effect of rs4299376 (ABCG5/ABCG8), which affects the intestinal cholesterol absorption pathway targeted by BAS, on both LDL-C and CAD. Nineteen RCTs with a total of 7,021 study participants met the inclusion criteria. Cholestyramine 24g/d was associated with a 23.5 mg/dL reduction in LDL-C (95% CI: -26.8,-20.2; N=3,806) and a trend towards reduced risk of CAD (OR: 0.81, 95% CI: 0.70-1.02; P=0.07; N=3,806) while colesevelam 3.75g/d was associated with a 22.7 mg/dL reduction in LDL-C (95% CI: -28.3,-17.2; N=759). Based on the genetic association of rs4299376 with a 2.75 mg/dL decrease in LDL-C and a 5% decrease in risk of CAD outcomes, we estimated that cholestyramine may be associated with an OR for CAD of 0.63 (95% CI: 0.52 - 0.77; P= 6.3x10-6; N=123,223) and colesevelam with an OR of 0.64 (95% CI: 0.52-0.79, P: 4.3x10-5). These estimates were not statistically different from previously reported trends from BAS clinical trials (P>0.05). The cholesterol lowering effect of BAS can thus be expected to translate into a clinically relevant reduction in the risk of CAD.

Categories: *Cardiovascular Disease and Hypertension, Mendelian Randomisation*

P 123

Mendelian Randomisation study of the causal influence of kidney function on coronary heart disease

Pimphen Charoen[1], UCLEB Consortium, Juan-Pablo Casas[1], Dorothea Nitsch[1], Frank Dudbridge[1]
[1]Dept Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK

Kidney function is known to correlate with coronary heart disease (CHD). However it is not yet clear whether kidney function reflects a causal pathway because this observed association could due to other confounding factors, such as BMI and blood pressure. Therefore we applied Mendelian Randomisation (MR) which allows disentangling of cause and effect in the presence of potential confounding, to determine whether kidney function has causal role to CHD. To our knowledge, this is the first MR study to investigate the causal influence of kidney function on CHD. The level of kidney function was measured by an estimated glomerular filtration rate (eGFR). To enhance the statistical power by increasing the sample size up to 200K, the summary statistics of associations between genetic variants and CHD from our UCL-LSHTM-Edinburgh-Bristol (UCLEB) Consortium were combined with the CARDIoGRAMplusC4D Consortium which is available publicly. Eighteen SNPs previously reported to be associated with eGFR were then established as instruments where their causal effects can be combined using the method proposed by Burgess et al. 2013 as well as a more general model which allows flexible scaling on an estimated causal effect. We observed no significant evidence of causal influence of eGFR on CHD. This may be due to the limited explanatory power of our genetic instrument, despite our large sample size, but also implies that the association observed between kidney function and CHD could due to confounding factors or reverse causation.

Categories: *Cardiovascular Disease and Hypertension, Causation, Mendelian Randomisation*

P 124

Shared genetic risk of myocardial infarction and blood lipids using empirically derived extended pedigrees: results from the Busselton Health Study

Gemma Cadby[1], Phillip E Melton[1], Jennie Hui[2], John Beilby[3], Arthur W Musk[4], Alan L James[5], Joseph Hung[6], John Blangero[7], Eric K Moses[1]
[1]Centre for Genetic Origins of Health and Disease, University of Western Australia
[2]Busselton Population Medical Research Institute Inc
[3]PathWest Laboratory Medicine WA
[4]Department of Respiratory Medicine, Sir Charles Gairdner Hospital
[5]Department of Pulmonary Physiology and Sleep Medicine, Sir Charles Gairdner Hospital
[6]School of Medicine and Pharmacology, University of Western Australia
[7]Texas Biomedical Research Institute

Quantitative endophenotypes related to complex diseases provide increased power for gene localisation and identification compared with dichotomous disease status. In this study, we employed empirically derived identical by descent (IBD) measures to estimate the heritabilities and genetic correlations between blood lipid endophenotypes (HDL-C, LDL-C and triglycerides) and myocardial infarction (MI) in 4671 individuals who attended the 1994/95 Busselton Health Study (BHS). IBD estimates were derived from genome-wide association data using LDAK software. MI events were identified from hospital morbidity and death registry data obtained from the Western Australian Health Department Data Linkage Unit. Heritability and genetic correlation between traits were calculated after adjusting for significant covariates (e.g. age, sex, lipid medication, smoking status). Approximately 75% of the 4671 individuals were related to at least one other BHS participant (up to and including third degree relatives). Between 1970 and 2011, 331 individuals had at least one MI event. Heritability of HDL-C, LDL-C and triglycerides were 0.54, 0.48, and 0.34, respectively (all $P<0.001$). HDL-C and triglycerides both showed a significant shared genetic correlation with MI of -0.43 ($P=0.01$) and 0.46 ($P=0.03$), respectively. HDL-C, LDL-C and triglycerides were highly heritable in the BHS and similar to earlier reported estimates, demonstrating the viability of using empirically derived IBDs. HDL-C and triglycerides both showed genetic correlation with MI, suggesting these are valuable endophenotypes for CVD-risk gene discovery.

Categories: *Cardiovascular Disease and Hypertension, Heritability*

P 125

Analysis of Case-Base-Control designs

Najla S Elhezzani[1], Wicher P Bergsma[2], Mike Weal[3]
[1]King's college London and King Saud university
[2]The London school of economics
[3]King's college London

Case-control studies compare individuals with a trait of interest (cases) with others who don't have it (controls). However, In many genetics association studies the control group is taken as a sample from the population where individuals have unknown trait status (bases). This approach appeared to be successful when the trait is rare. However, if the prevalence is high then using the bases as a set of controls will lead to unreliable results as power is compromised in this case. Accordingly, we proposed the case-base-control design which allows the three sample types to be used in a single analysis. To test wether genotype frequencies differ between cases and controls taking into account the bases, we derived the score test. The test reduces to Cochran-Armitage test when the CBC reduces to the CC. The score statistics shows a good adherence to the asymptotic distribution. We investigated the maximum likelihood estimates of the underlying parameters analytically and numerically using expectation-maximization algorithm. We derived the Wald's and likelihood ratio tests. We found that using moderate sample sizes, LRT was slightly more powerful than others, However for large samples the power of all tests not only becomes similar but also independent of the prevalence. Finally, we compared the CBC design with the usual case-control design. We found that only if the prevalence is well-specified and the proportion of cases in the bases is different from that in the experiment (cases and controls), then the CBC would provide more power compared to the CC. Looking at the case of having a large set of bases, we found that if prevalence is well specified then, the optimal design will be gained by using only cases if prevalence is low and only controls if it is high.

Categories: *Case-Control Studies, Maximum Likelihood Methods, Population Genetics, Prediction Modelling, Sample Size and Power*

P 126

Polymorphisms in HTR3A, CYP1A2, DRD4 and COMT and response to clozapine in treatment-resistant schizophrenia: a gene-gene interaction analysis

R Veera Manikandan[1], Anto P Rajkumar[2], Lakshmikirupa Sundaresan[1], Chithra C[1], Anju Kuruvilla[1], Alok Srivastava[1], Poonkuzhali Balasubramanian[1], Kuruthukulangara S Jacob[1], Molly Jacob[1]
[1]Christian Medical College, Vellore, India
[2]Aarhus University, Aarhus

Variable responses to clozapine in patients with schizophrenia are complex and poorly understood phenomena. The findings of pharmacogenetic studies on the use of this drug are poorly replicated. Effects of individual polymorphisms have rarely proved explanatory. One possible explanation may be multi-factorial involvement of genetic and environmental influences.The aim of this study is to evaluate the role of possible second and third order genetic interactions (epistasis) between polymorphisms in CYP1A2 (*1F, *1D, *1E, *1C), HTR3A (rs1062613 and rs2276302), DRD4 (120-bp duplication) and COMT (Val158Met) genes over clinical response, serum levels and adverse effects of clozapine in patients with treatment-resistant schizophrenia (TRS). The model-based multidimensionality reduction (MB-MDR) method has recently been shown to be superior to traditional parametric regression methods in detecting higher order gene-gene interactions. We used this approach in a sample of 93 patients with TRS to explore the epistatic effects of the polymorphisms of interest on clinical phenotypes of clozapine. The MB-MDR analysis showed a significant interaction between Val158Met, CYP1A2*1D and rs1062613 polymorphisms and clinical response to clozapine ($p=0.002$). In addition, multiple significant second and third order interactions were observed with regard to the adverse effects of clozapine ($p<0.05$). All the reported interactions were found to be significant after 1000 permutations. The observed multiple significant interactions emphasizes the importance of epistatic analysis in pharmacogenetic studies of clozapine. Such an approach may be useful in predicting a patient's response to clozapine therapy.

Categories: *Case-Control Studies, Gene - Gene Interaction, Genomic Variation, Multifactorial Diseases, Multilocus Analysis, Multivariate Phenotypes, Population Genetics, Psychiatric Diseases, Quantitative Trait Analysis*

P 127

Joint modeling of longitudinal and time-to-event phenotypes in genetic association studies: strengths and limitations

Osvaldo Espin-Garcia[1, 2], Zhijian Chen[2], Andrew D Paterson[3], Shelley B Bull[1,2]
[1]Dalla Lana School of Public Health, University of Toronto
[2]Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital
[3]The Hospital for Sick Children; Dalla Lana School of Public Health, University of Toronto

Genome-wide association study designs that evaluate multiple endpoints in observational settings are becoming more common. While oftentimes examination of single outcomes is sufficient for the purposes of the study, there are cases where joint analysis is informative in the simultaneous evaluation of genetic association with multiple endpoints. In particular, the study of time-to-event and longitudinal data arises naturally in cohort studies, but the use of joint analysis has remained rather unexplored in genetic association. The motivation for joint analysis comes to light under different scenarios. The objective may be to distinguish whether a SNP has a direct association with a time-to-event phenotype, and/or an indirect association through an intermediate quantitative trait (QT). This can be thought of as a form of causal inference: if the SNP association with time-to-event is negligible when the QT is well modelled in the survival analysis, then it cannot have a direct causal effect on time to event. Alternatively, genetic association with a QT may be of primary interest, but a clinical event causes informative censoring of the trait. In this work, we focus on the joint model proposed by Wulfsohn and Tsiatis (1997) and widely used in clinical studies of CD4+ counts and time to AIDS. We discuss estimation and causal interpretation of genetic association parameters in the joint model, examine statistical properties such as efficiency and bias of the effect estimates compared to their single-outcome-analysis counterpart, and quantify potential improvement in power to detect genetic association. In addition, we review software implementation and computational feasibility in the context of genome-wide analysis.

Categories: *Causation, Maximum Likelihood Methods, Multivariate Phenotypes, Quantitative Trait Analysis*

P 128

Perinatal depression and omega-3 fatty acids: A Mendelian randomisation study

Hannah Sallis[1,2], Colin Steer[3], Lavinia Paternoster[1], George Davey Smith[1], Jonathan Evans[2]
[1]MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, UK
[2]Centre for Academic Mental Health, School of Social and Community Medicine University of Bristol, UK
[3]Centre for Child and Adolescent Health, School of Social and Community Medicine, University of Bristol, UK

Introduction There have been numerous studies investigating the association between omega-3 fatty acids (FAs) and depression, with mixed findings. We propose an approach which is largely free from issues such as confounding or reverse causality to investigate this relationship using observational data from a pregnancy cohort. Methods The Avon Longitudinal Study of Parents and Children (ALSPAC) cohort collected information on FA levels from antenatal blood samples and depressive symptoms at several time points during pregnancy and the postnatal period. Conventional epidemiological analyses were used in addition to a Mendelian randomisation (MR) approach to investigate the association between levels of two omega-3 FAs (docosahexaenoic acid (DHA) and eicosapentaenoic acid (EPA)) and perinatal onset depression, antenatal depression and postnatal depression. We constructed a weighted allele risk score using independent SNPs identified as associated ($p<5x10-6$) in a recent genome-wide association study of omega-3 FAs by the CHARGE consortium. Results and Discussion Weak evidence of a positive association with both EPA (n=2377; OR=1.07; 95% CI: 0.99-1.15) and DHA (n=2378; OR=1.08; 95% CI: 0.98-1.19) with perinatal onset depression was found using a multivariable logistic regression adjusting for social class and maternal age. However, the strength of association was found to attenuate when using an MR analysis to investigate DHA. In conclusion, we found weak evidence of a positive association between omega-3 FAs and perinatal onset depression. However, without confirmation from the MR analysis, we are unable to draw conclusions regarding causality.

Categories: *Causation, Mendelian Randomisation, Psychiatric Diseases*

P 131

A Gene-Environment Interaction Between Copy Number Burden and Ozone Exposure Provides a High Risk of Autism

Dokyoon Kim[1], Heather Volk[2], Sarah Pendergrass[1], Molly A Hall[1], Shefali S Verma[1], Santhosh Girirajan[1], Irva Hertz-Picciotto[3], Marylyn Ritchie[1*], Scott Selleck[1]

[1]Department of Biochemistry & Molecular Biology, the Pennsylvania State University, University Park, PA
[2]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA; Department of Pediatrics, Children's Hospital Los Angeles, University of Southern California, Los Angeles, CA
[3]Department of Public Health Sciences, University of California, Davis, Davis, CA

Autism is a disorder of neural development as a complex genetic trait with a high degree of heritability as well as a documented susceptibility from environmental factors.The relative contributions of genetic factors, environmental factors and the interactions between them to a risk of autism are poorly understood.While most autism related copy number variations (CNV) identified to date, each with a substantial risk, are highly penetrant for this disorder, they constitute rare events contributing modestly to the overall heritability.Genome-wide analysis of CNV has demonstrated a continuous risk of autism associated with the level of copy number burden, measured as total base pairs of duplication or deletion.In addition, environmental exposure to air pollutants has been identified as a risk factor for developing autism, including particulate pollutants and nitrogen dioxide.We have examined the relative contribution of CNV (measured as total base pairs of copy number burden), exposure to air pollution, and the interaction between air pollutant levels and copy number burden in a population based case-control study, Childhood Autism Risks from Genetics and Environment (CHARGE).A significant and sizable interaction was found between duplication burden and ozone exposure (OR 2.78, $P<0.005$), greater than the main effect for either copy number duplication (OR 2.41, 95% CI: 1.36~4.82) or ozone alone (OR 1.19, 95% CI: 0.75~1.89).The overall implication of our findings is that significant gene-environment interaction associated with autism exists and could account for a considerable level of heritability not detected by evaluating DNA variation alone.

Categories: *Copy Number Variation, Gene - Environment Interaction*

P 132

Gene Regulatory Network inference via Conditional Inference Trees and Forests

Kyrylo Bessonov[1], Francesco Gadaleta[1], Kristel Van Steen[1]
[1]University of Liege

Trees are classical data structures allowing effectively classifying and predicting responses. Due to versatility and high performance in classification and prediction, there exist plenty of tree-based methods including popular Conditional Inference Tree (CIT) and Forests (CIF), Random Forests (RF), Randomized Trees (RT), randomized C4.5, etc. In this work we assessed the performance of CIT and CIF methods in correct gene regulatory network (GRN) prediction from expression data by using reference golden standard built from real transcriptional regulatory network of E. coli. The synthetic microarray expression data was obtained from DREAM4 challenge. The performance of each network inference method was assessed via Area Under Receiver Operating Characteristic (AUROC) and Area Under Precision Recall (AUPR) metrics. Our preliminary results show that CIT and CIF successfully predict directed GRNs at acceptable performance rates although not optimal (the best AUROC at 0.68 and AUPR at 0.13 for CIF and the best AUROC at 0.58 and AUPR at 0.18 for CIT). Surprisingly by using the current aggregation scheme of feature importance that prefers features with the highest number of observations, a single CIT was a better performer compared to CIFs in all 5 networks. Nevertheless, the CIFs showed an overall 10% improvement in AUROC. A single CIT has 24% and CIFs have 27% lower overall performance compared to the best performer of DREAM4 Challenge based on cumulative areas of PR and ROC curves. We plan to test other feature importance aggregation techniques in a single tree and in tree ensembles in order to outperform the top DREAM4 algorithms. In addition the effects of expression data standardization to unit variance will be presented. In future, the developed CIF framework will be used to perform data integration analysis of multi-omics datasets.

Categories: *Data Integration, Gene - Gene Interaction, Gene Expression Arrays, Gene Expression Patterns*

P 133

PREDICTING THE GENETIC RISK FOR COMPLEX DISEASES: CHOOSING THE BEST POLYGENIC RISK SCORE FOR TYPE II DIABETES

Kristi Läll[1,2], Krista Fischer[1], Reedik Mägi[1], Tõnu Esko[1]
[1]Estonian Genome Center, University of Tartu
[2] Institute of Mathematical Statistics, University of Tartu

We assess the practical value of the results from large-scale genome-wide association studies (GWAS) in personalised risk prediction for Type 2 Diabetes (T2D). A large number of associated variants (SNPs) across the genome has been identified, each having a relatively weak effect on the T2D risk. This motivates the use of polygenic risk scores, defined as weighted sums of risk allele frequencies. We discuss different options of constructing such scores in practice and study their advantages and disadvantages. The main selection criterion for a marker to be included in the score, is its significance (p-value) in the GWAS meta-analysis, whereas the estimated logistic regression coefficients are used as weights. Most often, only the genome-wide significant markers ($p<5*10-8$) are used in such scores at the moment. Some studies, however, propose including a larger number of independent SNPs, setting the p-value threshold in the range 0.1..0.5 or including all available markers. Different versions of polygenic risk scores for Type II Diabetes (T2D) will be constructed for the cohort of the Estonian Biobank. We will show that increasing the number of markers in the polygenic risk score for T2D improves the predictive ability until a certain p-value threshold. In addition, a significant interaction effect between the optimal polygenic risk score and Body Mass Index (BMI) on the prevalence of T2D is detected. Based on ROC- and reclassification analysis we conclude that most adequate risk prediction should account for age, BMI and polygenic risk score, whereas the predictive ability of the polygenic risk score differs across different BMI categories.

Categories: *Diabetes*

P 134

Epigenome-wide association with soluble cell adhesion molecules among monozygotic twins

Yan V Sun[1], Jack Goldberg[2], Dean Jones[3], Viola L Vaccarino[1]
[1]Emory University Rollins School of Public Health, Atlanta, GA, USA
[2]University of Washington School of Public Health, Seattle, WA, USA
[3]Emory University School of Medicine, Atlanta, GA, USA

Inflammation plays a critical role in the pathogenesis of cardiovascular disease. Epigenetic mechanisms, including DNA methylation (DNAm), have been shown to be critical in the regulation of inflammatory genes, and can be influenced by inflammation. The soluble form of cell adhesion molecules, including vascular adhesion molecule 1 (sVCAM1), intercellular adhesion molecule 1 (sICAM1), and P-selectin (sP-selectin), are established biomarkers for inflammation and endothelial function, and have been linked to cardiovascular events.

To identify epigenetic markers associated with inflammation and endothelial function, we conducted a methylome-wide association study of peripheral blood cells from 140 monozygotic (MZ) middle-aged male twins from the Emory Twin Study. Using two randomly selected subsets consisting of unrelated subjects, we identified and replicated 69 and 23 DNAm sites significantly associated with sVCAM1, and sICAM1 respectively, adjusted for multiple testing, but none for sP-selectin. All 23 sICAM1-associated DNAm sites were also associated with sVCAM1, including sites on genes ANKRD11, KDM2B, CAPS, CUX1, and HLA-DPA1. Two of these DNAm sites, located on UNC5D and TMEM125, were also significant comparing MZ twins who were phenotypically discordant for both sICAM1 (P=1.79×10-7, 2.78×10-6) and sVCAM1 (P=1.70×10-9, 1.71×10-7). These results suggest that sVCAM1 and sICAM1, but not sP-selectin, may share common pathophysiology in inflammation and endothelial function via an epigenetic mechanism. In addition, the epigenetic association with inflammation can be driven by unshared environmental exposures.

Categories: *Epigenetic Data, Epigenetics*

P 135

Genapha/dbASM: web based tools to investigate allele-specific methylation

George Ellis[1], BiLing Chen[1], Kevin Ushey[1], Denise Daley[1]
[1]University of British Columbia

As interest in studying allele-specific methylation (ASM) and it's association with common complex diseases grows, there is a need for a resource that stores and catalogs SNPs and regions that demonstrate allele specific methylation, analogous to NCBI's dbSNP. Additionally, as ASM is a regulatory mechanism that may be associated with hits from genome-wide association studies (GWAS), researchers need a suite of tools to help them evaluate the relationship between GWAS hits and ASM. To facilitate these investigations, we have created a new web resource called dbASM, hosted on the Genapha web server (www.genapha.ca). The aim of dbASM is twofold: 1. Curate from the literature a publicly-accessible database of known sites of ASM. 2. Provide researchers with a web-based platform of tools for exploring ASM and determining regions of interest. We will present the dbASM resource including details on the underlying database construction and datasets, in addition to the web tools and example workflows. The web tools that are currently available are: GWAS Catalog SNP Search, ASM SNP Search, SNP Counter, Methylation Plots Generation, and Sequence Viewer. GWAS Catalog SNP Search allows browsing through NHGRI's Catalog of Published Genome-Wide Association Studies by phenotype and filtering GWAS SNP's based on their relation to suspected sites of ASM. For example, rs11742570 is associated with inflammatory bowel disease (p=2.0 E-82) and demonstrates ASM. ASM SNP Search supports finding SNP's based on: ASM status or interrogability; location compared to genes, a chromosomal region, or other SNP's; and filtering by population minor allele frequencies and sample size. SNP Counter uses asynchronous JavaScript calls to the database to provide real-time counts of types of SNP's in user-selected regions of chromosome. Methylation Plots Generation is a calculates SNP correlation stratifying by genotype with CpG site methlyation patterns, similar to epigenome wide association studies (but without disease status),using CEPH HapMap samples and genotypes and methylation assays on these same samples completed on the Illumina 27K array. Sequence Viewer displays SNP's in the human reference genome (based currently on GRCh37.p10 and dbSNP build 137) with annotations showing ASM SNP's and regions of interrogability via MSRE cut sites for enzymes: HpyCH4IV, AciI, HhaI, and HpaI. These tools are all freely available for use at: http://genapha.icapture.ubc.ca/asm/.

Categories: *Epigenetics*

P 136

A gene-based method for analysis of Illumina 450K methylation data

Celia MT Greenwood[1,2], Kathleen Klein Oros[1] Aurelia Labbe[3], Stephan Busche[4], John Lambourne[4], Christian A Pineau[5,6], Sasha Bernatsky[5,6] , Ines Colmegna[5,6], Antonio Ciampi[3], Tomi Pastinen[7], Marie Hudson[1,5]

[1] Lady Davis Institute for Medical Research, Jewish General Hospital
[2]McGill University, Montreal, QC, Canada
[3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University
[4]McGill University and Genome Quebec Innovation Centre, McGill University
[5]Department of Medicine, McGill University, Montreal, QC, Canada
[6]Research Institute of the McGill University Health Centre
[7]Department of Human Genetics, McGill University

The epigenetic effects of DNA methylation play a critical role in regulating gene expression in human health and disease. The Illumina 450K methylation array allows the quantification of methylation levels at over 480,000 CpG sites throughout the genome. The large number of probes and the inherent correlation structure among nearby probes make it worth considering multiple-probe analyses. Here we propose a region-based method to increase power in detecting patterns of differential methylation, and we compare to univariate analyses in a sample of patients with systemic autoimmune rheumatic diseases SARDS). As part of an ongoing program of research on epigenetic signatures of SARDS, we recruited the following subjects: seropositive rheumatoid arthritis (n=12); systemic sclerosis (n=17); and systemic lupus erythematosus (n=12). Illumina 450K methylation data was obtained on cell-sorted CD4+ T lymphocytes and CD14+ monocytes from all patients at baseline. Similar cell subsets were retested in a group of patients that received Methotrexate treatment. The data was transformed using two alternative methods, a logit transformation and a beta quantile transformation to stabilize variances. We performed probe by probe univariate tests using beta-distribution regressions. For the gene based tests, we fit sparse principal component models using all probes within 5kb of gene boundaries. We then tested for association between the first few PCs and cell type/disease status. Gene-based analysis may have increased power to detect subtle changes in methylation patterns across genomic regions. This set of data provides a unique opportunity to study disease alterations in methylation data un-confounded by cell type differences.

Categories: *Epigenetics, Multivariate Phenotypes*

P 137

Take research to the next level with secondary data analyses: Fine-mapping the specific language impairment gene

William CL Stewart[1], Christopher W Bartlett[1]
[1]The Research Institute at Nationwide Children's Hospital

Mapping the gene mutations responsible for most simple Mendelian disorders was a major step forward in the fields of Human & Medical Genetics. However, as incredible as the mathematical and statistical tools that facilitated this achievement were, a more powerful collection of methods is needed to map the major genes that influence common, complex disease. To this end, we developed what may be the most powerful, integrated suite of statistical genetics software to date. Our suite is designed specifically for the secondary analyses of existing genetic data, although the analysis of newly acquired data is easily performed. The methods within our suite are (1) optimized for parallel computing; (2) rooted in statistical theory with substantial gains for large samples; (3) can integrate linkage, case-control, and family-based association with gene expression data; and, (4) interrogate both copy number and single nucleotide variants. The resulting high-speed, mathematically rigorous, and synergistic capabilities of our suite are likely to define the next-generation of methods development. As a proof of principle, we applied two programs in our suite: EAGLET and POPFAM to the secondary analysis of four large families segregating a specific language impairment gene on chromosome 13. We found that EAGLET reduced the size of the candidate region by 5 megabases, and that POPFAM—which incorporates information from matched controls and references samples, increased our ability to detect associated variants beneath the linkage peak. Overall, this should significantly aid re-sequencing efforts as we close in on the causal alleles.

Categories: *Fine Mapping, Linkage Analysis, Linkage and Association, Markov Chain Monte Carlo Methods, Maximum Likelihood Methods, Multilocus Analysis*

P 138

Detection of Gene-Gene Interaction in Affected Sib Pairs Allowing for Parent-of-Origin Effects

Chih-Chieh Wu[1], Sanjay Shete[2]
[1]National Cheng Kung University
[2]MD Anderson Cancer Center

Genome-wide association studies have discovered several hundred genetic variants associated with common diseases, which in most situations explain a small fraction of the heritability. Gene-gene interactions can play an important role in disease susceptibility and may account for some of the missing susceptibility. Parent-of-origin effects refer to the differential expressions of a gene between two parental chromosomes and have been increasingly observed in mammals. The development of statistical methods is important and needed that are capable of capturing joint actions of individual genetic components underlying the disease susceptibility and allow for parent-of-origin effects. Here, we extended our previous allele-sharing method and presented 3 mathematical two-locus models incorporating parent-of-origin effects: additive, multiplicative, and general models. Our methods are model-free based on allelic identity-by-descent sharing by affected sib pairs. We propose the use of two-locus score method to assess the gene-gene interaction effects using affected sib pairs in the presence of parent-of-origin effects.

Categories: *Gene - Gene Interaction*

P 139

Study Designs for Predictive Biomarkers

Andreas Ziegler[1]
[1]University of Lübeck, Institute of Medical Biometry and Statistics

Biomarkers are of increasing importance for personalized medicine, including diagnosis, prognosis and targeted therapy of a patient. Examples are provided for current use of biomarkers in applications. It is shown that their use is extremely diverse, and it varies from pharmacodynamics to treatment monitoring. The particular features of biomarkers are discussed. Before biomarkers are used in clinical routine, several phases of research need to be successfully passed, and important aspects of these phases are considered. Some biomarkers are intended to predict the likely response of a patient to a treatment in terms of efficacy and/or safety, and these biomarkers are termed predictive biomarkers or, more generally, companion diagnostic tests. Using examples from the literature, different clinical trial designs are introduced for these biomarkers, and their pros and cons are discussed in detail.

Categories: *Gene - Environment Interaction, Genetic Data for Clinical Trial Design*

P 140

Does the FTO gene interact with the socio-economic status on the obesity development among young European children? Results from the IDEFICS study

Ronja Foraita[1], Frauke Günther[1], Wencke Gwozdz[2], Lucia A Reisch[2], Paola Russo[3], Fabio Lauria[3], Alfonso Siani[3], Toomas Veidebaum[4], Michael Tornaritis[5], Iris Pigeot[1], on behalf of the IDEFICS consortium

[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany
[2]Copenhagen Business School, Department of Intercultural Communication and Management, Frederiksberg, Denmark
[3]National Research Council, Institute of Food Science, Epidemiology and Population Genetics, Avellino, Italy
[4]National Institute for Health Development, Department of Chronic Diseases, Tallinn, Estonia
[5]Research and Education Institute of Child Health, Strovolos, Cyprus

Various twin studies revealed that the influence of genetic factors on psychological diseases or behavior is more expressed in socio-economically advantaged environments. Other studies predominantly show an inverse relation between socio-economic status (SES) and childhood obesity in western developed countries. The aim of this study is to investigate whether the FTO gene interacts with the socio-economic status (SES) on childhood obesity in a subsample of the IDEFICS cohort (N=4406). A structural equation model (SEM) is applied with the latent constructs obesity, dietary habits, physical activity and fitness habits, and parental SES to estimate the main effects of the latter three variables and a FTO polymorphism on obesity. Further, a multiple group SEM is used to explore whether an interaction effect between the single nucleotide polymorphism rs9939609 within the FTO gene and SES exists. Overall model fit was inconsistent (RMSEA=0.05; CFI=0.79). Significant main effects are shown for SES (standardized βs=-0.057), the FTO homozygous risk genotype AA (βs=0.177) and physical activity and fitness habits (βs=-0.113). The explained variance of obesity is about 9%. The multiple group SEM shows that SES and FTO interact in their effect on childhood obesity ($\Delta\chi 2=7.3$, df=2, p=0.03) insofar as children carrying the protective TT genotype are more susceptible to a favorable social environment.

Categories: *Gene - Environment Interaction*

P 141

Identification of Clusters in Network Graphs by a Correlation-based Markov Cluster Algorithm

Martin L Jäger[1], Ronja Foraita[1]
[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

A common goal in gene expression analysis is to identify groups of genes with correlating expression levels. The Markov Cluster Algorithm (MCL)1 is a method to identify such clusters in undirected network graphs. It converts the graph's adjacency matrix into a probability matrix which is then expanded and inflated until it converges. Clusters can be deduced from the resulting equilibrium state matrix. However, the MCL considers associations between genes only in a dichotomous manner. Hence, our objective is to examine whether the MCL based on the partial correlation identifies more reasonable clusters. A simulation study consisting of three differently sized gene expression networks and six types of clusters is carried out. These types of clusters differ in size, number of clusters existing, underlying distribution and structure. Each cluster type is modelled in a gene expression network consisting of 100 observations and 100, 500 and 1000 genes, respectively. We conduct 1000 replications for each combination of cluster type and network size. The performance of the partial correlation-based MCL is compared to the adjacency-based MCL as well as to k-means clustering and PART (Partitioning Algorithm based on Recursive Thresholding)2 which are applied using the gap statistic3. The adjusted Rand index4 is used to assess the extent to which clusters match the true clusters and to compare the algorithms among each other. References: [1] Van Dongen. PhD thesis, 2000, University of Utrecht. [2] Nilsen et al. Stat Appl Genet Molec Biol, 2013, 12(5): 637-652 [3] Tibshirani et al. J R Stat Soc B, 2001, 63(2): 411-423 [4] Hubert & Arabie. J Classif, 1985, 2(1): 193-218

Categories: *Gene Expression Patterns*

P 142

Develop novel mixture model to estimate the time to antidepressant Onset of SSRIs and the timing effects of key covariates

Yin Yao[1], Meng Yuan Xu[1], Wei Guo[1]
[1]National Institutes of Mental Health

Longitudinal data sets on drug onset—which have only recently become available for research—require multiple-point measurements. We sought to develop a statistical model capable of analyzing longitudinal data with tipping points—specifically, the point in time when a therapeutic drug begins to take effect. We have termed this novel method the 'mixture model'. To take underlying driving factors into account, we also tested the association(s) between time of onset and potential underlying factors. The new mixture model proposed not only models a drug onset but also tests its associations with influential variables such as gender, age, and disease subtype. In order to estimate time of onset, data were divided into three stages: 1) drug naïve state; 2) drug onset; and 3) identifiable drug effects. In addition to estimating when onset occurs, our proposed statistical model takes into account any associations with potentially influential factors. We conducted four simulation studies to test the feasibility of our new method, and also applied it to real-world data from the STAR*D study.The mixture model identified the effect of these different variables on time to onset of drug effects. While the limited sample size makes it difficult to generalize any conclusions from this study, several clinically relevant observations emerged. Our results indicated that for non-anxious and younger patients, the effects of citalopram were apparent earlier—by the sixth week; in contrast, for those individuals classified as having anxiety at baseline, drug effects did not appear until the eighth week of treatment.

Categories: *Genetic Data for Clinical Trial Design, Prediction Modelling*

P 143

Defining recombination hot spot blocks: Just how hot is hot?

Tae-Hwi Schwantes-An[1], Heejong Sung[1], Alexa JM Sorant[1], Jeremy A Sabourin[1], Cristina M Justice[1], Alexander F Wilson[1]
[1]National Human Genome Research Institute / National Institutes of Health

In the past decade, the number of available genetic markers used in genetic studies of human disease has grown exponentially. From dozens of microsatellites for linkage studies to millions of markers in GWAS chips and in whole genome/exome next-generation sequencing in current family/association studies, the increasing density of markers has been instrumental for the fine-mapping of the human genome. However, the increase in marker density has made it increasingly difficult to adjust for multiple tests because of correlations between markers caused by linkage and gametic disequilibrium (LD, GD). Defining and identifying the independent regions of the genome can provide an alternative assessment of the number of "independent" tests for next generation sequencing. One method that can be used to identify regions of "independent" regions in the genome is by identifying blocks of the genome that are flanked by recombination hot spots. Recombination hot spots are defined as regions of the genome that show an increased rate of recombination than expected at random 1cM/Mb (1 centimorgan per megabase). These blocks can be used to identify blocks of the genome that are mostly independent from one another. To identify these independent blocks (regions divided by recombination hot spots), the genome is classified into hot spots (regions above a predefined recombination threshold) and cold spots (regions below a threshold) using recombination rates (cM/Mb); counts and average size of the hot/cold spot blocks can be determined. Increasing the threshold (5%, 10%, 15%, and 20%) increases the average size of the cold spots and decreases the number of hot spots, however the average size of hot spots does not appear to change.

Categories: *Genomic Variation*

P 144

Complex genealogies, simple geometric structures

Marc Jeanpierre[1]
[1]Université. Descartes

Ancient variations always have a long and complicated history. As haplotype decay is essentially stochastic, simple geometric structures that can be described unambiguously in mathematical terms can provide the algebraic framework for analysing the forces shaping the genealogy of a single allele, or a cluster of variants. Considering the simplest example, a three-branch bifurcating tree, there are two possible ways of adding a branch to an existing pair of branches. These two independent and complementary paths of construction are represented by two alternative equations. The possible construction paths therefore reflect the hierarchical organization of the tree. Star-like genealogies are by far the easiest to analyze, as this model bypasses all the difficulties of translating a set of mosaic haplotypes into a specific genealogy. In non-star genealogies, there are always several possible ways to break down a complex genealogy in subtrees. The different construction paths representing alternative sequence of events that may be observed naturally makes use of parameters as branch lengths that can be represented graphically. Subtrees are conditionally independent from upstream nodes and equations representing specific sequences of events may be constructed from bottom to top. The shape of the tree needed to decipher the history of mutation ancestry is a mathematical abstraction. The definition of haplotype blocks as physical entities, with clear borders, as for objects in the physical world, results in an apparent simplification, but is not really helpful because unnecessary reduction of complexity prevents the derivation of meaningful patterns.

Categories: *Haplotype Analysis, Multiple Marker Disequilibrium Analysis*

P 145

Missing heritability partially explained by sequential enrollment of study participants

Damia Noce[1], Martin Gögele[1], Christine Schwienbacher[1], Alessandro De Grandi[1], Yuri D'Elia[1], Peter P Pramstaller[1], Cristian Pattaro[1]
[1]Center for Biomedicine, European Academy of Bolzano/Bozen (EURAC) (affiliated Institute of the University of Lübeck), Bolzano, Italy

In pedigree-based studies the recruitment strategy could play an important role to explain part of the missing heritability. A recruitment carried on over a long time period might pair up with seasonal or day-specific conditions, such as ambient temperature, sample transport conditions and laboratory sample handling, introducing sample stratification similar to the sibship (SS) effect. To quantify the impact of such issues, we analyzed 54 blood parameters from the first 2948 participants of the Cooperative Health Research in South Tyrol (CHRIS) study, enrolled from Aug 2011 until Jul 2013 and connected through an extended pedigree. To maximize participation of complete families we enrolled preferentially close relatives within the same day (up 10 per day). Genetic heritability (h2) was estimated by fitting sex- and age-adjusted variance components models. We additionally included shared environmental effects defined as day of participation (DoP), daily temperature (DT) and SS. We observed a h2 reduction for 49, 28 and 39 traits when accounting for DoP, DT and SS, respectively. When including the DoP, the h2 reduction was >10% for 11 traits and >40% for sodium, chlorine, calcium and mean corpuscular hemoglobin concentration. The SS effect induced >10% h2 reduction for 10 traits and >40% only for cortisol. Despite being associated with some traits, DT did not alter h2 estimates substantially. The day of participation, as a proxy for issues that may happen during the enrollment or measurement phase, can be an important stratification factor, which may induce stronger heritability overestimation than the sibship effect. When appropriate, it should be used to complement the sibship effect to prevent population stratification.

Categories: *Heritability*

P 146

Robust Principal Component Analysis Applied to Population Genetics Processes

Carine Legrand[1], Justo Lorenzo Bermejo[1]
[1]Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

It has been shown that principal genetic components reflect evolutionary processes and the genetic parameters of a population. For example, McVean provided a genealogical interpretation of principal component analysis (PCA) [1]. We have examined the ability of several robust PCA methods to mirror population changes in heterozygosity and admixture. Evolutionary processes were simulated using simuPOP and own scripts [2]. We first examined genetic drift in a single population (CEU haplotypes from HapMap) considering growth, recombination and selection. We also simulated gene flow in South America after the arrival of individuals with European and African ancestries, allowing for a fast population growth in the last century. CEU and YRI samples from the 1000 Genomes Project represented European and African components. PCA results motivated the use of MXL (Mexican) instead of CLM (Colombian) genotypes as surrogates of native South American ancestry. Heterozygosity and admixture were quantified in the evolving populations, and their relationship with the principal genetic components estimated by standard PCA, spherical PCA, and minimum covariance determinant methods was examined. Results from the genetic drift scenario revealed a stronger correlation between heterozygosity and the robust principal genetic components. The simulation of South American admixture also revealed a potential advantage of robust PCA. Results from ongoing sensitivity analyses will be presented at the conference.

[1] McVean G (2009) A Genealogical Interpretation of Principal Components Analysis. PLoS Genet 5(10): e1000686.
[2] Peng B, Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. Bioinformatics, 21(18): 3686-7.

Categories: *Heterogeneity, Homogeneity, Population Genetics, Population Stratification*

P 147

Identifying founders most likely to have introduced disease-causing mutations with the R package GenLib

Claudia Moreau[1*], Jean-François Lefebvre[1*], Héloïse Gauvin[1,2], Michèle Jomphe[3], Christoph Preuss[1], Gregor Andelfinger[1,4], Damian Labuda[1,4], Hélène Vézina[3], Marie-Hélène Roy-Gagnon[1,5]
*These authers contributed equally to this work

[1]CHU Sainte-Justine Research Center, Montreal, Quebec, Canada
[2]Department of Social and Preventive Medicine, Université de Montréal, Montreal, Quebec, Canada
[3]BALSAC Project, Université du Québec à Chicoutimi, Chicoutimi, Quebec, Canada
[4]Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada
[5]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada

Founder populations, such as the French Canadian (FC) population of Quebec, Canada, play an important role in the study of genetic diseases. Their advantages often include access to detailed genealogical records. Once evidence for a disease-causing mutation has been found, genealogical data can be used to identify the founders most likely to have introduced the mutation in the population. Large genealogical data require specialized analytical methods and software. We present the R package GenLib for genealogical analysis. GenLib can compute relevant summary measures describing genealogies and relatedness, including kinship and inbreeding coefficients. It also performs gene-dropping simulations. In this study, we extended the GenLib gene-dropping simulation function to take into account the length of the segment passed IBD through generations and a fitness parameter for homozygotes. This extension allows a more precise estimation through simulations of the probability that the shared segment descended from a specific founder. We illustrate the use of GenLib with genealogical data from 11 patients with the recently identified autosomal recessive syndrome of Chronic Atrial and Intestinal Dysrhythmia (CAID). Average kinship and inbreeding coefficients of these patients were 0.002 and 0.004, respectively. We found that one founding couple had a probability over 80 times larger than that of any other founders to have introduced the mutation in the FC population. This couple immigrated to Quebec City from France around 1621. These results provide information on expected frequencies of the disease in the population and on the diffusion pattern of the mutation on the Quebec territory.

Categories: *Inbreeding, Isolate Populations, Population Genetics*

P 148

Regional IBD Analysis (RIA): linkage analysis in extended pedigrees using genome-wide SNP data

Jakris Eu-ahsunthornwattana[1,2], Heather J Cordell[1]
[1]Institute of Genetic Medicine, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne, NE1 3BZ, UK
[2]Division of Medical Genetics, Department of Internal Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Rama VI Rd, Ratchathevi, Bangkok 10400, Thailand

Exact calculations for traditional linkage analysis are computationally impractical in large, extended pedigrees. Although simulation-based methods can be used, they are not exact and still require significant computational work. For these circumstances, we propose Regional IBD Analysis (RIA), a non-parametric linkage method based on comparison of locally and globally estimated identity by descent (IBD) sharing in affected relative pairs. In this method, genome-wide SNP data are used to calculate the “global” expected IBD sharing probabilities specific to each affected relative pair, against which a "local" set of IBD sharing probabilities, estimated using SNP data within a window of pre-specified width, can be compared. These IBD sharing probabilities can be estimated using a variety of programs/methods: we used PLINK and KING in this study. The global and local IBD sharing probabilities can be used to construct a non-parametric maximum likelihood statistic (MLS)-like test of linkage in each window. We illustrate the use of our method to detect linkage signals in real nuclear-family data and in simulated data based on large extended pedigrees. This method should be useful in studies involving large extended families, with an additional advantage of not having to rely on any prior knowledge about familial relatedness.

Categories: *Linkage Analysis*

P 149

Polygenic risk prediction modeling in pedigrees improves power

Jeffery Staples[1], Chad D Huff[2], Jennifer E Below[3]
[1]The University of Washington
[2]The University of Texas MD Anderson Cancer Center
[3]The University of Texas Health Science Center

As analyses of sequence data in large population based cohorts struggle to achieve sufficient power to detect even large signals from very rare variation, the family-based linkage approach has come back into vogue. In the context of complex disease traits however, ability to detect true signal is impeded by modifying environmental and genetic factors that influence rates of penetrance and phenocopies. Classically, known modifiers of disease risk, e.g. age, have been modeled in liability classes. The era of GWAS has taught us a great deal about common underlying genetic effects on complex traits. These polygenic effects impact risk of disease and can act as modifying factors to rare variation segregating in pedigrees. We show that modeling these effects improves the ability to both detect true linkage signals of large effect rare variants from the genome and correctly identify unlinked markers. In simulations of genotypes for a 1000 different pedigrees (mean size 25, 36% missing samples) we modeled phenotypes by modifying the probability of disease given a large effect dominant risk allele, A, using a simulated aggregate polygenic risk score (pgrs) calculated from 100 different common variants: P(d|aa) =pgrs, P(d|Aa,AA)=pgrs+0.9. We compared LOD scores at the causal variant and an unlinked variant when modeling the pgrs in an individual-specific liability class to scores derived from a single shared liability class. Power to detect the casual variant increased in >60% of our simulations, overall averaging >10% increase and mean LOD gain of >0.25. Incorporating polygenic risk prediction slightly lowered LOD at unlinked markers. Accurate modeling of established polygenic risk factors improves power estimates in linkage studies.

Categories: *Linkage Analysis, Linkage and Association*

P 150

Performance of linkage analysis conducted with whole exome sequencing data

Simon Gosset[1], Edgard Verdura[2], Françoise Bergametti[2], Stephanie Guey[2], Elisabeth Tournier-Lasserve[2], Steven Gazal[3]
[1]INSERM U1137, IAME, Université Paris Diderot, Paris, France
[2]INSERM U1161, Université Paris Diderot, Paris, France
[3]Assistance Publique des Hopitaux de Paris (APHP), Paris, France

Identification of causal variants in Mendelian disorder was usually done by combining linkage analysis (LA) on large families and positional cloning. The progress of the high-throughput sequencing led teams to perform directly whole exome sequencing (WES) for the identification of these variants, particularly for single small families that can be analysed by a simple filter analysis. However, it is essential to minimize the number of candidate variants before starting studies on their functional consequences. To reduce the number of variants that are sequencing errors, not covered in one individual, or without allelic frequency in reference database, and to facilitate the study of recessive diseases with allelic heterogeneity, an additional LA can be performed. Many studies have thus combined their WES filtering with a LA on microsatellites or SNP chips, which uniformly cover the genome. Perform a LA on common polymorphisms present in WES data appears as an attractive strategy to reduce the cost of the analyses. However, it has been rarely done, due to the non-uniform exon coverage of the genome, and to the lack of knowledge of LA power on this kind of data. Our goal was to study the performance of LA conducted with exome genotypes. To achieve this, we performed a simulation study of 2 families (one with a dominant disease, one with a recessive disease) and compared LA results on WES genotypes and data from SNP chips. Our results show that a LA conducted on WES genotypes excludes accurately a high proportion of the genome. In addition, its false positive and false negative evidence of linkage are in the same range that the ones of LA conducted on SNP chips. Finally, an application on real data will illustrate the benefits of this strategy.

Categories: *Linkage Analysis, Sequencing Data*

P 151

Use of exome sequencing data for the analysis of population structures, inbreeding, and familial linkage

Vincent Pedergnana[1,2], Aziz Belkadi[1], Avinash Avinash[3], Quentin Vincent[1], Yuval Itan[4], Bertrand Boisson[4], Jean-Laurent Casanova[1,5], Laurent Abel[1,5]
[1]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, University Paris Descartes, Imagine Institute, Paris, France
[2]Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom
[3]New York Genome Center, New York, NY, USA
[4]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, the Rockefeller University, New York, NY, USA
[5]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, the Rockefeller University

Numerous methods have been proposed to analyze whole exome sequencing (WES) data in order to discover potential causal variants in Mendelian disorders and in more complex traits. These methods could benefit from additional information such as linkage studies in the study of Mendelian diseases. Population stratification could also be an issue in the analysis of WES data when focusing on complex traits. Both linkage and population structure analyses are classically conducted through genome-wide (GW) SNP arrays. Here, we compared the information yielded by WES data to that provided by SNP array data in terms of analyses usually performed by SNP array data such as principal component analyses (PCA), linkage studies, and homozygosity rate estimation. We analyzed 123 subjects originating from six world regions, including North Africa and Middle East which are regions poorly covered by public database and presenting a high consanguinity rate. A number of quality control (QC) filters were tested and applied to the WES data. Compared to results obtained with SNP array data, we found that WES data provided accurate prediction of population substructure and led to highly reliable estimation of homozygosity rates (correlation > 0.94 with the estimations provided by SNP array). Linkage analyses showed that the linkage information provided by WES data was on average 53% lower than the one provided by SNP array at the GW level, but 58% higher in the coding regions. In conclusion, WES data could be used after appropriate QC filters to perform PCA analysis and adjust for population substructure, to estimate homozygosity rates, and to perform linkage analyses at least in coding regions.

Categories: *Linkage Analysis, Population Genetics, Sequencing Data*

P 152

Fast linkage analysis with MOD scores using algebraic calculation

Markus Brugger[1,2] , Konstantin Strauch[1,2]
[1]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany
[2]Institute of Genetic Epidemiology, Helmholtz Zentrum München , German Research Center for Environmental Health, Neuherberg, Germany

Objective The mode of inheritance is often unknown for complex diseases. In the context of parametric linkage analysis, this implies that a MOD-score analysis, in which the LOD score is maximized with respect to the trait-model parameters, can be more powerful. Because the calculation of the disease-locus likelihood for every tested set of trait-model parameters is the most time-consuming step in a MOD-score analysis, we aimed to optimize this part of the calculation to speed-up linkage analysis using the GENEHUNTER-MODSCORE software package. Methods Our new algorithm is based on minimizing the effective number of inheritance vectors by collapsing them into classes. To this end, the disease-locus-likelihood contribution of each inheritance vector is represented and stored in its algebraic form as a symbolic sum of products of penetrances and disease-allele frequencies. Simulations of datasets were used to assess the speed-up of our new algorithm. Results Focusing on MOD-score analysis of single datasets, we were able to obtain speed-ups ranging from 1.94 for affected-sib pairs to 11.52 for affected-sib sextets compared to the original GENEHUNTER-MODSCORE version. When including simulations to calculate empirical p values, the speed-up ranged from 1.69 to 10.36. Speed-up was generally higher for larger pedigrees. Conclusions Computation times for MOD-score analysis including p-value calculation have been prohibitively high so far. With our new algebraic algorithm, the evaluation of many tested sets of trait-model parameters during the maximization in a MOD-score analysis is now feasible within a reasonable amount of time, even when empirical p values are calculated.

Categories: *Linkage Analysis*

P 153

Fetal exposures and perinatal influences on the premature infant microbiome

Diana A Chernikova[1], Devin C Koestler[2], Anne G Hoen[3], Molly L Housman[4], Patricia L Hibberd[5], Jason H Moore[6], Hilary G Morrison[7], Mitchell L Sogin[7], Muhammad Z Ul-Abideen[8], Juliette C Madan[9]
[1]Department of Genetics, Geisel School of Medicine at Dartmouth
[2]Department of Biostatistics, University of Kansas Medical Center
[3]Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth
[4]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth
[5]Department of Pediatrics, Massachusetts General Hospital
[6]Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth
[7]Josephine Bay Paul Center, Marine Biological Laboratory
[8]Geisel School of Medicine at Dartmouth
[9]Department of Pediatrics, Dartmouth-Hitchcock Medical Center

The impact of maternal complications on the premature infant microbiome is still largely unexplored. To investigate the effects of these complications on the gut microbiome, we collected serial stool samples obtained weekly from extremely premature infants enrolled in a prospective longitudinal study from birth through hospital discharge, and then sequenced the V4V6 region of bacterial 16S rRNA genes. Perinatal maternal complications evaluated included prolonged preterm premature rupture of membranes (PPPROM), chorioamnionitis, delivery mode, and peripartum antibiotics. Subjects with prenatal exposure to a non-sterile intrauterine environment (PPPROM and chorioamnionitis) were found to have relatively higher abundance of known pathogenic bacteria across all time points compared to subjects without those exposures, irrespective of exposure to postnatal antibiotics. Compared with those delivered by Cesarean section, vaginally delivered subjects were found to have a significantly lower microbial diversity across all time points, with lower abundance of many bacterial genera, mostly in the family Enterobacteriaceae. Hierarchical clustering analysis showed that samples associated with a non-sterile uterine environment clustered together and had an enrichment of pathogens; furthermore, the cluster's average microbial diversity score that was significantly lower than that of a cluster of samples without the exposure, which instead had an enrichment of important gut commensals. Our results demonstrate that exposure to prenatal pathogens impacts the development of the premature gut microbiome, and highlights opportunities to intervene via breast milk feedings, altered antibiotic regimens, or probiotics.

Categories: *Microbiome Data*

P 154

Combining genotype with allelic association as input for iterative pruning principal component analysis (ipPCA) to resolve population substructures

Kridsadakorn Chaichoompu[1,2], Ramouna Fouladi[1,2], Pongsakorn Wangkumhang[3], Alisa Wilantho[3], Wanwisa Chareanchim[3], Sissades Tongsima[3], Anavaj Sakuntabhai[4], Kristel Van Steen[1,2]
[1]Systems and Modeling Unit, Montefiore Institute, University of Liege, Belgium
[2]Bioinformatics and Modeling, GIGA-R, University of Liege, Belgium
[3]Biostatistics and informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, Thailand
[4]Functional Genetics of Infectious Diseases Unit, Institut Pasteur, France

Single Nucleotide Polymorphisms (SNPs) are commonly used to capture variations between populations and often genome-wide SNP data are pruned based on linkage disequilibrium (LD) patterns. Notably, haplotype composition and the pattern of LD between markers may vary between larger populations but may also play a role within more confined geographic regions. Indeed, knowledge about haplotypes in unrelated individuals can reveal useful information about genetic ancestry. Here, we use iterative pruning principal component analysis (ipPCA) [Intarapanich 2009] to identify and characterize subpopulations in an unsupervised way using a rich set of genetic markers since using reduced sets of genetic markers for these purposes can become challenging, especially when similar geographic regions are involved or when spurious patterns are likely to exist. As input data, either pruned genome-wide SNP data are used or multilocus haplotype information derived from the genome-wide SNP panel. These approaches are applied to real-life data from 4028 Vietnamese individuals [Khor 2012]. Preliminary results indicate that ipPCA applied to pruned SNP data or ipPCA that explicitly uses multilocus information (haplotypes) give complementary information about population substructure for geographically confined populations. Both methods address different aspects of population structure. In conclusion, we propose to combine an LD-based haplotype encoding scheme with the ipPCA machinery to retrieve fine population substructures. Despite the complexities that are associated with haplotype inference, added value can be obtained when the LD structure between SNPs is exploited in the search for relevant population strata.

Categories: *Population Genetics, Population Stratification*

P 155

Spurious cryptic relatedness can be induced by population substructure, population admixture and sequencing batch effects

Di Zhang[1], Shuwei Li[1], Gao T Wang[1], Suzanne M Leal[1]
[1]Center for Statistical Genetics, Baylor College of Medicine

It is important to identify cryptically related individuals in population-based association studies, since inclusion of related individuals can increase type I & II errors. To resolve this problem mixed models have been proposed, but they can be computationally intensive and type I & II errors can be inflated. Another option is to remove related individuals from analysis. Data quality control should include identification of cryptically related individuals. Caution should be used, since population substructure/admixture and sequence data batch effects can cause detection of spurious relatedness. In order to investigate the problem we evaluated the relatedness of 1,092 samples in 1000 Genomes and 2,300 African-American subjects from the NHLBI-Exome Sequencing project via two published methods for kinship inference: (i) the PLINK algorithm which is based on identical-by-descent statistic under the assumption of homogeneous population, and (ii) the KING-robust algorithm which uses an estimate of the genome-wide average heterozygosity across individuals to compute an estimator of kinship coefficient. We identified spurious relatedness due to population substructure/admixture and batch effects with both methods, but the problem was more severe for PLINK. An excess of 3rd degree relatives was observed due population admixture/substructure and batch effects. The kinship coefficients also varied depending on how the analysis was performed and individuals were reclassified, e.g from 1st degree to 2nd degree relatives. In addition to presenting the results of these analyses and showing the severity of the biases in the kinship coefficients, we also demonstrate strategies to avoid the detection of spurious relatedness.

Categories: *Population Genetics, Population Stratification, Sequencing Data*

P 156

Effect of population stratification on validity of a case-only study to detect gene-environment interactions

Pankaj Yadav[1], Sandra Freitag-Wolf[1], Wolfgang Lieb[2], Michael Krawczak[1]
[1]Institute for Medical Informatic and Statistic, Christian-Albrechts University, Kiel, Germany
[2]Institute of Epidemiology, Christian-Albrechts University, Kiel, Germany

Gene-environment (G×E) interaction studies are assumed to partially fill the gap between the estimated heritability of common human diseases and the genetic component hitherto explained by disease-associated variants. The case-only (CO) study has been proposed as a valid approach with increased statistical efficiency over case-control and cohort studies in detecting G×E interactions. However, hidden stratification in the study population can severely compromise a CO study. None of the prior literature explicitly addressed the effect of stratification on a CO study. We therefore systematically assessed through simulations the effect of population stratification (PS) on the validity of a CO approach in G×E interactions studies. Our simulations show that, when study sample is divided by both genetic and exposure factors, a CO study provides an inflated type I error rate. Further, our simulations show that transmission disequilibrium test (TDT) is robust against genetic and/or exposure stratification in detecting G×E interactions.

Categories: *Population Stratification*

P 157

A novel risk prediction algorithm with application to smoking experimentation

Rajesh Talluri[1], Anna Wilkinson[2], Margaret Spitz[3], Sanjay Shete[1]
[1]The University of Texas, M. D. Anderson Cancer Center
[2] University of Texas School of Public Health
[3]Baylor College of Medicine

Risk prediction models are being developed to predict the risk of a variety of cancers, and cardiovascular diseases. However, standard approaches do not account for the variability associated with the cohort being a random sample from the population. We developed a novel risk prediction approach called Resampling-based Model Selection and Aggregation to compute absolute risk. Our approach accounted for variability in the sampled cohort by resampling the data and aggregating the parameter estimates for the resampled datasets. We then used a resampling-based model selection algorithm to select the predictors to include in the final multivariable risk model. This approach guards against over-fitting the model and reduces the variance of the model parameters. The performance of the risk prediction model was evaluated using the area under the receiver operating characteristic curve (AUC). Using the risk prediction model, we computed the absolute risk of smoking experimentation in Mexican American youth. The data included genetic and non-genetic factors that were collected at baseline. The proposed risk prediction model had an AUC of 0.719 (95% confidence interval, 0.637 to 0.801) for predicting absolute risk for smoking experimentation within 1 year.

Categories: *Prediction Modelling*

P 158

Trio-Based Whole Genome Sequence Analysis of a Cousin Pair with Refractory Anorexia Nervosa

P Betty Shih[1], Ashley Van Zeeland[2], Andrew Bergen[3], Tristan Carland[4], Vikas Bansal[1], Pierre Magistretti[5], Wade Berrettini[6], Walter Kaye[1], Nicholas Schork[7]
[1]University of California, San Diego, La Jolla, CA
[2]Cypher Genomics, La Jolla, CA
[3]SRI, Palo Alto, CA
[4]The Scripps Research Institute, La Jolla, CA
[5]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[6]University of Pennsylvania, Philadelphia, PA
[7]J. Craig Venter Institute, La Jolla, CA

Anorexia Nervosa (AN) has an onset during adolescence and is characterized by emaciation, fear of gaining weight despite being underweight, and has the highest mortality rate of all psychiatric illnesses. Despite the serious health and psychosocial consequences of this illness, very few treatments are effective at reversing the core symptoms of AN. AN is highly heritable and show a homogeneous clinical presentation of persistent food refusal and high anxiety traits. However, AN etiology is believed to be heterogeneous as no major susceptibility gene has been consistently replicated in multiple populations. AN symptoms and personality traits tend to be present in unaffected family members of the patients, suggesting that certain shared genetic factors within each family may contribute to unique phenotype risk of the affected. To gain insights into the role "private variants" may play in AN and to maximize genetic information from family members of AN, here we leveraged a family-based study design combined with whole genome sequencing to search for genetic variants that may influence AN risk in an affected cousin pair together with their parents. By capitalizing on the homogeneity of the disease presentation among the two cousins, who both have a diagnosis of refractory AN, we report methods by which we interrogated shared chromosomal segments transmitted to them from their common grandparents that carried likely AN-related functional variants in this family.

Categories: *Psychiatric Diseases, Sequencing Data*

P 159

Power and sample size formulas for detecting genetic association in longitudinal data using generalized estimating equations

Ghislain Rocheleau[1], Loïc Yengo[2], Philippe Froguel[2]
[1]. Université Lille 2, Lille, France
[2]CNRS 8199 - Institute of Biology, Pasteur Institute, Lille, France

Currently, most genetic studies only exploit cross-sectional data to detect novel associations between a SNP and a quantitative trait, even if repeatedly measured outcomes are available for analysis. Instead of focusing on some baseline or single time point measurement, it might be desirable to identify SNPs associated with that trait over time. One possible approach to model correlated measures over time is the generalized estimating equations (GEE), especially if interest lies in detecting the mean differences of the trait as a function of the genotypes. Unlike linear mixed models, GEE models do not require the joint distribution to be fully specified, only the mean and the variance must conform to linear model specifications, along with an appropriate within-cluster correlation matrix. However, in power analysis, this within-cluster correlation matrix is often unknown and is usually modelled as a function of time. Common choices for this matrix include compound symmetry, autoregressive (AR) or moving average (MA) structures. Using asymptotic theory of the Wald test statistic, we derive closed-form formulas for power and sample size estimation under an autoregressive moving average ARMA(1,1) covariance matrix. Interestingly, the ARMA(1,1) covariance matrix is equivalent to an AR(1) covariance matrix plus independent measurement error. We apply our formulas to simulated genotype and phenotype data, and to real data coming from the French cohort D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d'Insulino-Résistance).

Categories: *Quantitative Trait Analysis, Sample Size and Power*

P 160

On the evaluation of predictive biomarkers with dichotomous endpoints: a comparison of the linear and the logistic probability models

Nicole Heßler[1], Andreas Ziegler[1,2]
[1]Institut für Medizinische Biometrie und Statistik, Universitt zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany
[2]Zentrum für Klinische Studien, Universität zu Lübeck, Lübeck, Germany

The standard statistical approach for analyzing dichotomous endpoints is the logistic regression model which has major statistical advantages. However, some researchers prefer the linear probability model over the logistic model in randomized trials for evaluating predictive biomarkers. The main reason seems to be the interpretation of effect estimates as absolute risk reductions which can be directly related to the number needed to treat. In the first part of our presentation, we provide a comprehensive comparison of the two different models for the investigation of treatment and biomarker effects. Using the logistic regression model, Kraft et al. (2007, Hum Hered) showed that the combined 2 degrees of freedom (2df) gene, gene-environment interaction test should be the test of choice for testing genetic effects. In the biomarker treatment setting a gene corresponds to the treatment and environment to biomarker. Using this analogy we extend the study of Kraft et al. in the second part of our presentation. We compare several test statistics including the 2df combination test using the linear probability model. The pros and cons of the combined test are discussed in detail. We demonstrate substantial power loss of the combination test in comparison with either the test for treatment or the test for treatment-biomarker interaction in many scenarios. Although the combination test has reasonable power in all situations considered, its power loss compared to a specialized 1df test can be large. Therefore, the combined test cannot be recommended as the standard approach in studies of treatment-biomarker interaction.

P 161

A two stage random forest probability machine approach for epigenome-wide association studies

Frauke C Degenhardt[1], Andre Franke[1], Silke Szymczak[1]
[1]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany

DNA methylation as the best studied mechanism of epigenetic modification of the genome plays an important role in gene expression, embryonic development and disease control. Nowadays, next generation sequencing technologies can generate methylation data for several millions of CpG sites throughout the genome that might be spatially correlated. Identifying single sites or genomic regions that enable classification of individuals, e.g. as cases or controls is challenging. We propose a two-step random forest probability machine (RFPM) approach to select important regions and sites within these regions. First, a RFPM is trained on sites in each region separately. The estimated probability based on this region (synthetic feature) is then used as input for a genome-wide RFPM and important regions and sites within these regions are identified using appropriate variable importance measures. We evaluate our approach based on methylation data sets from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) and compare it to a more time consuming approach using all sites or summarized methylation ratios per region.

P 162

Statistical approaches for gene-based analysis: A comprehensive comparison using Monte-Carlo Simulations

Carmen Dering[1], Inke R König[1], Andreas Ziegler[1]
[1]Institut für Medizinische Biometrie und Statistik, Unversität zu Lübeck

In recent years several studies detected associations between groups of rare variants and common diseases. These findings resulted in the development of the ”rare variant-common disease” (RVCD) hypothesis, stating that multiple rare variants together may be causal for a common disease. Therefore, many statistical tests, the collapsing methods, were developed which are the topic of this work. We compared fifteen statistical approaches in a gene-based analysis of simulated case-control data of the Genetic Analysis Workshop (GAW) 17 in various collapsing scenarios and 200 replicates. Scenarios differed in minor allele frequency (MAF) threshold and functionality of corresponding collapsed rare variants. Almost all of the investigated approaches showed an increased type-I-error. Furthermore, none of the statistical tests was able to detect true associations over a substantial proportion of replicates in the simulated data. Irrespective of the statistic test used, collapsing methods seem to be generally useless in small case-control studies. Recent work indicates that large sample sizes and a substantial proportion of causing rare variants in the gene-based analysis can yield greater power. However, many of the investigated approaches use permutation which means high computational cost, especially when applying a genome-wide significance level. Overcoming the issue of low power in small case-control studies is a challenging task for the near future.

P 163

Apolipoprotein E gene polymorphism and left ventricular failure in beta-thalassemia: A meta-analysis

Niki Dimou[1], Katerina Pantavou[1], Pantelis Bagos[1]
[1]University of Thessaly

The beta-thalassemia syndromes are a heterogeneous group of genetic disorders characterized by reduced or absent expression of the beta-globin gene. Despite appropriate transfusion and chelation therapy and low ferritin levels, patients still develop organ failure, heart failure being the main cause of death. ApoE acts as a scavenger of free radicals; iron chelation is probably another mechanism of its antioxidant activity.This study was performed to determine whether the decreased antioxidant activity of the apolipoprotein E (APOE) 4 allele could represent a genetic risk factor for the development of left ventricular failure (LVF) in beta-thalassemia homozygotes under a multivariate meta-analysis approach. We included 4 studies with 613 thalassemic patients and 664 controls. According to the echocardiographic findings, patients were divided into three groups: i) asymptotic patients; ii) patients with evidence of LV dilatation; and iii) patients with clinical and echocardiographic findings of LV failure. This classification scheme with the existence of multiple groups as well as multiple alleles, created a multivariate response and subsequently, the need to resort to multivariate methods of meta-analysis. We came up with overall significant results contrasting E4 and E3 vs. E2 allele for each group (Wald test=17.14; p-value=0.009). Multivariate methods suggest a significant role played by the E4 allele when contrasting E4 allele vs. others (OR = 2.49, 95% CI: 1.28, 4.86 and OR = 3.43, 95% CI: 1.84, 6.41 for group II and III respectively, Wald test=16.80; p-value<0.001). Meta-regression analysis failed to provide evidence that the risk conferred by E4 allele is associated with clinical or haematological parameters.

P 164

The Cooperative Health Research in South Tyrol (CHRIS) study

Cristian Pattaro[1], Martin Gögele[1], Deborah Mascalzoni[1], Alessandro De Grandi[1], Christine Schwienbacher[1], Fabiola Del Greco M[1], Roberto Melotti[1], Maurizio F Facheris[2], Peter P Pramstaller[1]
[1]Center for Biomedicine, European Academy of Bolzano (EURAC) (affiliated Institute of the University of Lübeck), Bolzano, Italy
[2]The Michael J. Fox Foundation for Parkinson's Research, New York, New York, USA

The Cooperative Health Research in South Tyrol (CHRIS, www.christudy.it) is a population-based study to investigate the genetic etiology of cardiovascular, metabolic and neurological diseases, started in 2011 in the Venosta valley (Italy). The population is characterized by long-term social stability without major immigration events, families all connected by few very large pedigrees, and homogeneous environmental conditions. Through a community-based communication strategy followed by personal invitation, all 28,000 resident adults are being contacted. We expect more than 10,000 to be voluntarily enrolled. Eighteen self- and interviewer-administered internationally validated questionnaires reconstruct their medical history. Electronic instrumental recordings assess fat intake, cardiac function, and tremor. To enhance power of gene-environment interaction analyses, life-style exposures (nutrient intake, physical activity, life course smoking) are assessed quantitatively. Urine and blood are collected to measure 19 and 54 parameters, respectively, and for biobanking (cryo-preserved urine, DNA, whole and fractioned blood). All participants will be genotyped on a dense SNP array. A subset will undergo whole-genome sequencing to identify rare variants enriched in this population. Involved in the P3G, BBMRI, and BioSHaRE initiatives, the CHRIS study and biobank constitute a valuable resource for scientists willing to investigate genetic factors inhibiting a disease-free and healthy aging.