2015 Symposium on Advances in Genomics, Epidemiology and Statistics

Abstracts

May 29, 2015

Funding for this conference was made possible in part by grant no. 1 R13 HG007809-01A1 from the National Human Genome Research Institute, National Institute of Environmental Health Sciences and National Institute on Deafness and Other Communication Disorders. The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

The Children's Hospital *of* Philadelphia®
RESEARCH INSTITUTE

An expanded sequence context model broadly explains variability in polymorphism levels across the human genome

*Varun Aggarwala and Benjamin F. Voight

Genomics and Computational Biology Graduate Group, University of Pennsylvania School of Medicine

The rate of single nucleotide polymorphism varies ~1000 fold across the human genome and fundamentally impacts evolution and incidence of genetic disease. Previous studies have only considered the immediate flanking nucleotides around a polymorphic site –the site's trinucleotide sequence context–to study polymorphisms levels genome-wide. Moreover, the impact of larger sequence context has not been fully clarified, even though context substantially influences rates of polymorphism. Using a novel statistical framework and data from the 1000 Genomes Project, we demonstrate that a heptanucleotide context explains up to 93% of variability in substitution probabilities, revealing novel mutation promoting motifs including ApT dinucleotides, CAAT, and TACG sequences. Our approach reveals previously undocumented variability in C-to-T substitutions at CpG sites, not immediately explained by differential methylation intensity. Using our model, we present a substitution tolerance score for genes and a novel tolerance score for amino acids to guide variant prioritization in clinical resequencing studies.

* Eligible for pre-doc poster award

Evaluating the Prognostic Gene Signatures in Bladder Cancer using "Curated Bladder Data" R package

Ragheed Al-Dulaimi, MD, MPH, Msc; Shakiba Muhammadi, MD, MPH

University of Utah, Salt Lake City, Utah

Bladder Cancer (BC) is one of the most common malignancies worldwide with an estimated 74,690 new cases and 15,580 deaths in the United States only. Multiple gene signatures have been identified previously in BC. Despite high number of publications, none of these signatures have been fully accepted into routine clinical practice. For clinicians to adopt a genetic signature as a meaningful tool for prognostic evaluation and treatment decision makings for BC, the bar is high.

We saw the need to conduct meta-analysis of several gene expression datasets to evaluate these gene signatures as predictive markers for disease progression and mortality. We used both data driven and knowledge guided predictive modeling. A systematic search of Pubmed for previously published literature of gene predictors of prognosis and survival from BC was conducted. The gene signatures identified from the literature were tested in the cox proportional hazard model. We used gene expression datasets from a pre-existing R package of Curated bladder data to estimate the hazard ratios (HR) of death from BC for these genes. The calculated HRs were compared to the ones from previous publication. In addition, genes were ranked based on the correlation between each gene and death from BC, and the ones with the highest ranks were used to build the best multivariate predictive model of mortality from BC. R statistical language v2.9 was used for analysis including curated Bladder data and metaphor packages.

Can rare variants account for signals from common variants?

**Ferdouse Begum, Ingo Ruczinski, Mary L. Marazita, Jeffrey C. Murray, Terri H. Beaty, Margaret A. Taub

Johns Hopkins University, Baltimore, MD

While genome-wide association studies (GWAS) have successfully identified polymorphic markers associated with complex diseases, few are directly causal. High-throughput sequencing holds the promise of identifying directly causal mutations. Several statistical methods for sequencing data under case-control study designs are available, but fewer methods for family-based studies exist. Also, rare variants (RV) capable of producing deleterious gene products cannot be identified in conventional association study designs due to lack of power. Our method looks for excess of putatively functional RVs on transmitted risk haplotypes tagged by a common marker showing significant association in the allelic transmission disequilibrium test (TDT), compared to corresponding untransmitted risk and non-risk haplotypes. We illustrate our approach using 1,409 case-parent trios ascertained through a child with a non-syndromic oral cleft from 3 ethnic groups (Chinese, Filipino and European) sequenced on 6.3 Mb in 13 candidate genes/regions. We used counts of putatively deleterious RVs transmitted with the common risk allele and compared them to counts of RVs on both untransmitted risk and non-risk haplotypes and observed some significant differences. We used permutation-based tests to assess evidence that non-random sets of RVs are associated with transmission to the affected child. We will expand this approach to consider the linkage disequilibrium structure and further characterize functionality of RVs, which may offer a viable method for resolving causal variants that could explain associations found in GWAS.

** Eligible for post-doc poster award

Birth Month Affects Lifetime Disease Risk: A Phenome-Wide Method

*Mary Regina Boland[1,5], Zachary Shahn[4], David Madigan[4-5], George Hripcsak[1,5], Nicholas P Tatonetti[1-3, 5]

[1]Department of Biomedical Informatics, [2]Department of Systems Biology, [3]Department of Medicine, [4]Department of Statistics, [5]Observational Health Data Sciences and Informatics (OHDSI), Columbia University, New York, NY, USA

An individual's birth month has a significant impact on their lifetime disease risk. Previous studies reveal relationships between birth month and several diseases including atherothrombosis, asthma, attention deficit hyperactivity disorder, and myopia, leaving most diseases completely unexplored. We developed a hypothesis-free method that systematically investigates disease-birth month patterns across all conditions. Our dataset includes 1,749,400 individuals with records at New York-Presbyterian/Columbia University Medical Center. We modeled associations between birth month and 1,688 diseases using logistic regression. Significance was assessed using a chi-squared test with multiplicity correction. We found 55 diseases with a significant birth month dependency. Of these 39 were reported in the literature, and a remaining 16 diseases were completely unreported. We found distinct incidence patterns across disease categories. Individuals born in birth months with higher cardiovascular disease incidence (February-June) were also associated with decreased life-expectancy in the literature corroborating our findings. Neurological diseases, pregnancy conditions and asthma associations revealed by our method were validated by European studies in the literature. Overall, we found that individuals born in May and July had the lowest overall disease risk. Lifetime disease risk is affected by birth month. Seasonally-dependent developmental mechanisms may help explain these associations.

* Eligible for pre-doc poster award

Fast and sensitive metagenomic sequence assignment with Centrifuge

**Florian Breitwieser, Daehwan Kim, Li Song, Steven Salzberg

Johns Hopkins University

Centrifuge is a fast, sensitive and leightweight program for identifying organisms from metagenomic sequencing data down to the species level. Metagenomics sequencing is becoming more and more important in both biological research and clinical investigations. However, tools for identifying bacterial and viral reads are typically computational expensive, very space-consuming or not very sensitive. Centrifuge circumvents computational bottlenecks by building on an indexing scheme based on the Burrows-Wheeler transform and the FM-index. This enables small initial seed sizes and thus senstive alignment, as well as rapid extension. The Centrifuge database with all RefSeq bacterial and viral genomes and the full human genome consumes 5.1GB memory. The runtime for samples with millions of reads is typically below an hour on standard hardware. In addition to the read-level assignment results, Centrifuge provides a report with accurate species abundance estimates, employing a count sketching algorithm. On test data, Centrifuge performs comparable or better than existing tools in identfying and quantifying species.

** Eligible for post-doc poster award

Using an integrated gene-based sequence kernel association test (iSKAT) to identify subtype specific single nucleotide variants in glioma

Yian Ann Chen, Jamie K. Teer, Zachary J. Thompson, Rebekah L. Baskin, Yonghong O. Zhang, Kate J. Fisher, Zhihua Chen, Alvaro N. Monteiro, Kathleen M. Egan

Moffitt Cancer Center

One of the challenges for genome-wide association analyses is that the effect directions and allele frequencies (e.g., rare vs. common) of true causal variants are unknown. Built on a family of powerful approaches, sequence kernel association test (SKAT), we have devised an omnibus approach, integrated-SKAT (iSKAT), to perform association tests using next-generation sequence data. This includes a suite of 12 methods: Burden test, SKAT, SKAT-O, SKAT-C (Combined sum test of rare- and common-variant effects), SKAT-A (Adaptive sum test), SKAT-AR, three methods weighted by functional scores, and three rare-variant only methods. Minimum FDR was used to adjust for the multiple comparison across methods.

We applied iSKAT to investigate sub-type specific susceptibility loci between the low-grade glioma (LGG) and glioblastoma (GBM) as a proof of principle study. We downloaded the germline exome sequence data (N = 612) data from TCGA, and aligned the sequence reads using the Burrows-Wheeler Aligner (BWA). Insertion/deletion realignment, quality score recalibration, and variant identification were performed with the Genome Analysis ToolKit (GATK). We used 80% SNV call rate for quality control. After PCA was performed, 544 Caucasian samples were included. A total of 224K SNVs in 18,053 genes was included in the analyses. Ten genes were significantly associated with glioma subtypes (minFDR of 10%). Among which, there are 9 significant SNVs with predicted possible damaging functions in 6 genes. Five of these genes are differentially expressed with statistical significance between LGG and GBM, and consistently supported by both microarray and RNAseq platforms. These findings were promising.

Genetic association mapping in admixed populations

Guimin Gao and Wenan Chen

Virginia Commonwealth University

Admixed populations, such as African Americans, are populations formed by recent admixture of two or more ancestral populations. For gene mapping in admixed populations, admixture mapping tests based on admixture linkage disequilibrium (LD) can only identify a large chromosomal segment (usually several Mbs) harboring a causal variant. Association tests that correct for local ancestry can use background LD existing in the ancestral populations and map a causal variant into a small region, within less than a few hundred Kbs. However, these association tests may have relatively low statistical power in detecting causal variants with admixture mapping signals. To improve power, several joint association tests that combine information from admixture mapping tests and association tests that correct for local ancestry have been proposed. In this study, we show that in genome wide association studies (GWAS), when testing the null hypothesis that a SNP is not in background LD with the causal variants, several existing methods, including the association tests adjusting for global ancestry and the joint association tests, cannot control well the type I error rate or family-wise error rate (FWER) in the strong sense. Furthermore, we propose weighted multiple testing procedures to incorporate information from admixture mapping tests into association tests that correct for local ancestry in the context of GWAS. Our simulation studies indicate that the proposed association testing procedures not only can control FWER, but also improve statistical power compared to the association tests that correct for local ancestry.

Why do fragile X carrier frequencies differ between Asian and non-Asian populations?

Diane P. Genereux (1) and Charles D. Laird (2, 3)

1. Department of Biology, Westfield State University; 2. Department of Biology, University of Washington; Center on Human Development and Disability, University of Washington

Asian and non-Asian populations have been reported to differ substantially in the distribution of fragile X alleles into the normal (< 55 CGG repeats), premutation (55–199 CGG repeats), and full-mutation (> 199 CGG repeats) size classes. Our statistical analyses of data from published general-population studies confirm that Asian populations have markedly lower frequencies of premutation alleles, reminiscent of earlier findings for expanded alleles at the Huntington's Disease locus. To examine historical and contemporary factors that may have shaped and now sustain allele-frequency differences at the fragile X locus, we develop a population-genetic/epigenetic model, and apply it to these published data. We find that founder-haplotype effects likely contribute to observed frequency differences via substantially lower mutation rates in Asian populations. By contrast, any premutation frequency differences present in founder populations would have disappeared in the several millennia since initial establishment of these groups. Differences in the reproductive fitness of female premutation carriers arising from fragile X primary ovarian insufficiency (FXPOI) and from differences in mean maternal age may also contribute to global variation in carrier frequencies. We conclude with a discussion of how information from double-stranded DNA methylation patterns can deepen our understanding of fragile X syndrome and other genetic/epigenetic diseases.

Machine Learning Derived Disease Risk Prediction of Anorexia Nervosa

**Yiran Guo[1], Zhi Wei[2], Brendan Keating[1,3], The Genetic Consortium for Anorexia Nervosa[4], The Wellcome Trust Case Control Consortium 3[4], Hakon Hakonarson[1,3]

1. The Center for Applied Genomics (CAG), Abramson Research Center, The Children's Hospital of Philadelphia (CHOP), Philadelphia, PA 19104, USA; 2. Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA; 3. Department of Pediatrics, School of Medicine University of Pennsylvania, Philadelphia, PA 19104, USA

Anorexia nervosa (AN) is a complex psychiatric disease with a moderate to strong genetic contribution. In addition to conventional genome wide association (GWA) studies, researchers have been using machine learning methods in conjunction with genomic data to predict risk of diseases in which genetics play an important role. In this study, we collected whole genome genotyping data on 3,940 AN cases and 9,266 controls from the Genetic Consortium for Anorexia Nervosa (GCAN), the Wellcome Trust Case Control Consortium 3 (WTCCC3) and the Children's Hospital of Philadelphia (CHOP), and applied machine learning methods for predicting AN disease risk. The prediction performance is measured by area under the receiver operating characteristic curve (AUC), indicating how well the model distinguishes cases from unaffected control subjects. Logistic regression model with the lasso penalty technique generated an AUC of 0.693, while Support Vector Machines and Gradient Boosted Trees reached AUC's of 0.691 and 0.623, respectively. Using different sample sizes, our results suggest that larger datasets are required to optimize the machine learning models and achieve higher AUC values. To our knowledge, this is the first attempt to assess AN risk based on genome wide genotype level data. Future integration of genomic, environmental and family-based information is likely to improve the AN risk evaluation process, eventually benefitting AN patients and families in the clinical setting.

** Eligible for post-doc poster award

Candidate Gene Resequencing, SNP Chip or GBS for Diverse Maize Germplasm

Christine Hainey

DuPont Pioneer, Wilmington, DE

In 2013 the US achieved average corn yields of 7.73 Metric tons per hectare (MT/Ha), compared to sub-Saharan Africa where average yields range from 1.5 MT/Ha in Kenya to 3.82 MT/Ha in South Africa. In order to raise yields in sub-Saharan Africa, a collaboration was established through Improved Maize for African Soils (IMAS). It is a partnership between CIMMYT, Kenya Agricultural & Livestock Research Organization (KALRO), Agricultural Research Council of South Africa (ARC), and DuPont Pioneer Hi-Bred with a grant from the Bill & Melinda Gates Foundation and USAID. This collaboration employed a multifaceted approach to locate regions in the genome associated with traits improving yield. These experiments consisted of genotyping over 400 diverse inbred lines from CIMMYT, KALRO and ARC with three different genotyping platforms. These 400 lines were than phenotyped in Africa for numerous agronomic traits and an extensive hydroponic seedling assay was conducted at the DuPont Pioneer facility in Johnston, IA.

This comprehensive approach to association mapping indicates a single genotyping approach for a diverse breeding program is not sufficient to pick up exotic alleles associated with these phenotypic traits in an African maize breeding program.

r2VIM: Variable selection method for identifying interaction effects

**Emily Holzinger, Silke Szymcak, James Malley, Abhijit Dasgupta, Qing Li, Joan Bailey-Wilson

Computational and Statistical Genomics Branch, NHGRI/NIH, Baltimore, MD

Standard analysis methods for genome wide association studies (GWAS) are not robust to complex disease models, which likely contribute to the heritability of complex human traits. Machine learning methods, such as Random Forests (RF), are an alternative analysis approach. One caveat to RF is that there is no standardized method of selecting variables so that false positives are reduced while retaining adequate power. To this end, we have developed a variable selection method called r2VIM. This method incorporates recurrency and variance estimation for optimal threshold selection. We assess how this method performs in simulated data with close to completely epistatic effects (i.e. no marginal effects).

Our findings indicate that the optimal selection threshold can often identify interactions while reducing the number of false positives in the selected variables. However, the optimal threshold is highly dependent on the simulated genetic model, which is unknown in biological data. To address this, we also test a permutation procedure to generate null VIM distributions based on the actual genotype data to guide threshold selection. We permute the phenotype and re-run r2VIM to get a new estimate of the null variance. This is then used to choose a selection threshold for the non-permuted analysis. We tested the permutation method on a subset of the simulated data used in the initial analysis. The results suggest that the permutation procedure can guide optimal threshold selection in data with strong interaction effects in a manner that retains locus detection power and a low false positive selection rate.

** Eligible for post-doc poster award

MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations

**Xing Hua[1], Paula L. Hyland[2], Jing Huang[3], Bin Zhu[1], Neil E. Caporaso[2], Maria Teresa Landi[2], Nilanjan Chatterjee[1] and Jianxin Shi[1]

[1]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, [2]Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics and [3]Laboratory of Cancer Biology and Genetics, Center for Cancer Research, National Cancer Institute, National Institute of Health, Bethesda, Maryland, 20892, USA

We present a powerful and flexible computational framework for systematically identifying mutually exclusive gene sets (MEGS) with mutations in tumor sequencing studies. Our method is based on a likelihood ratio test and a model selection procedure. Extensive simulations demonstrated that our method outperformed existing methods for both power and the capability of identifying the exact MEGS, particularly for highly imbalanced MEGS. Our method can be used for *de novo* discovery, pathway-guided searches or for expanding established small MEGS. We applied our method to the TCGA whole exome sequencing data and identified multiple previously unreported non-pairwise MEGS in multiple cancer types.

** Eligible for post-doc poster award

CODEX: a normalization and copy number variation detection method for whole-exome sequencing.

*Yuchao Jiang, Derek A Oldridge, Sharon J Diskin, Nancy R Zhang.

University of Pennsylvania, Philadelphia, PA; The Children's Hospital of Philadelphia, Philadelphia, PA.

Background: High throughput sequencing of DNA coding regions has become a common way of assaying genomic variation in the study of cancer. Copy number aberration (CNA) is an important type of genomic change, but detecting and characterizing CNA from whole-exome sequencing (WES) is challenging due to the high level of biases and artifacts.

Methods: We propose CODEX, a normalization and CNA calling procedure for WES data. The Poisson latent factor model in CODEX includes terms that specifically remove biases due to GC content, exon capture and amplification efficiency, and latent systemic artifacts. CODEX also includes a Poisson likelihood-based recursive segmentation procedure that explicitly models the count-based WES data. CODEX can be used to detect both germline and somatic CNAs in cancer samples with or without matched normal.

Results: Compared to existing approaches, CODEX is shown to be more effective in removing the biases in WES, and attains better sensitivity and specificity in detecting copy number aberrations by in silico spike-in studies. We further evaluate performance on 222 neuroblastoma samples with matched normal from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) Project. We carry out systematic genome-wide analysis and detailed characterization of both Germline and somatic copy number events. With a focus on a well-studied rare somatic CNV within the *ATRX* gene, we show that the cross-sample normalization procedure of CODEX is more effective in removing noise than the standard pipeline of normalizing the tumor against the matched normal, and that the segmentation procedure performs well in detecting CNVs with recurrent complex nested structures. For detecting germline mutations, CODEX is compared to existing methods on a population analysis of HapMap samples from the 1000 Genomes Project, and shown to be perform well on three microarraybased validation data sets.

Conclusions: The cross-sample normalization procedure of CODEX, when applied to the matrix of tumor and normal samples, is more effective in reducing noise than normalizing each tumor to its matched normal. The somatic deletions in the *ATRX* region have a nested structure, which CODEX was able to recover. Through multiple types of validation, CODEX is shown to be applicable to a wide range of study designs for copy number estimation using WES data.

* Eligible for pre-doc poster award

Detecting Differentially Expressed Proteins

**Kai Kammers

Department of Biostatistics, Johns Hopkins University, Baltimore, MD

High throughput proteomics is one key component in contemporary multi-omics approaches. The problem of identifying differentially expressed proteins in massspectrometry based experiments is ubiquitous, and most commonly these comparisons are carried out using t-tests for each peptide or protein. Sample sizes are often small however, sometimes as small as 4 or 8 samples total, which results in great uncertainty for the estimates of the standard errors in the test statistics. The consequence is that proteins exhibiting a large fold change are often declared non-significant because of a large sample variance, while at the same time small observed fold changes might be declared statistically significant, because of a small sample variance. Additionally, adjustments for multiple testing reduce the list of significant peptides or proteins.

We review and demonstrate how much better results can be achieved by using “moderated” t-statistics, arguably the analytical standard for gene expression experiments. This empirical Bayes approach shrinks the estimated sample variances for each peptide or protein towards a pooled estimate, resulting in far more stable inference particularly when the number of samples is small. Using real data from labeled proteomics experiments (iTRAQ and TMT technology) and simulated data we show how to analyze data from multiple experiments simultaneously, and discuss the effects of missing data on the inference. We also present easy to use open source software for normalization of mass spectrometry data and inference based on moderated test statistics.

** Eligible for post-doc poster award

Allelic diversity of Human Disease Genes in the Amish

**Rachel Kember

University of Pennsylvania, Philadelphia, PA

Identifying genetic variants that play a role in disease is the main goal of many geneticists. A genetic isolate with greater genetic and phenotypic homogeneity, such as the Old Order Amish (OOA) founder population, is ideal for studying the role of such variants. The OOA contain a subset of the variants found in the general population, and therefore rare variants can be found in higher frequencies. By utilizing a combination of whole genome sequence for 80 subjects (30 parent-child trios) with dense SNP genotype data for 394 family members to provide an accurately imputed and phased whole genome sequence, we created a catalog of Amish SNPs and CNVs in disease loci. We have identified variants in both the heterozygous and homozygous state that are known to be disease causing, in addition to mutations in disease genes that are private to this population. By utilizing the rich pedigree data for the large extended multigenerational pedigree we can capture the distribution of these disease mutations and follow their transmission across multiple generations.

** Eligible for post-doc poster award

Follow-Up and Replication Study of Caries in the Permanent Dentition

**D. D. Lewis [1] J.R. Shaffer [1] E. Feingold [1,2] M. Cooper [3,4] M.M Vanyukov [1,5,6] B.S. Maher [7] R.L. Slayton [8] M.C. Willing [9] S.E. Reis [10,11] D.W. McNeil [12] R.J. Crout [13] R.J. Weyant [14] S.M. Levy [15,16] A.R. Vieira [3,4] M.L. Marazita [1,3,4,6,11]

[1] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA [2] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA [3] Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA [4] Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA [5] Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, Pittsburgh, PA [6] Department of Psychiatry, School of Medicine, University of Pittsburgh, Pittsburgh, PA [7] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD [8] Department of Pediatric Dentistry, School of Dentistry, University of Washington, Seattle, WA [9] Division of Genetics and Genomics, Medicine, Department of Pediatrics, School Medicine, Washington University at St. Louis, St. Louis MO [10] Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA [11] Clinical and Translational Science Institute, School of Medicine, University of Pittsburgh, PA [12] Dental Practice and Rural Health, West Virginia University, Morgantown, WV [13] Department of Periodontics, School of Dentistry, West Virginia University, Morgantown, WV [14] Department of Dental Public Health and Information Management, School of Dental Medicine, University of Pittsburgh, PA [15] Department of Preventive and Community Dentistry, University of Iowa College of Dentistry, Iowa City, IA [16] Department of Epidemiology University of Iowa College of Public Health, Iowa City, IA

Recent genome-wide association studies (GWAS) of permanent dentition caries have identified novel loci (AJAP1, TGFBR1, NR4A3, LYZL2, IFT88, ISL1, CNIH, BCOR, BCORL1, and INHBA) for further study. The aim of this study is to replicate these putative genetics associations in six independent studies of non-Hispanic whites and blacks. In this study, we interrogated 158 single nucleotide polymorphisms (SNPs) in 13 race- and age stratified samples from six independent studies (*n* =3600). All statistical analyses were performed separately for each sample, and results were combined across samples by meta-analysis. CNIH was significantly associated with caries via meta-analysis across eight adult samples, with four SNPS showing significant associations in white adults after gene-wise adjustment for multiple testing ($p < 0.001$). These results corroborate the previous GWAS study, although the functional role of CNIH in caries etiology remains unknown. BCOR also showed significant association in four SNPs, with the strongest evidence of association was observed in white adults (p=9.11E05). Mutations in this gene results in an X-linked dominant Mendelian disorder oculofaciocardiodental (OFCD) syndrome which is responsible for several dental abnormalities. Furthermore, in adults, genetic association was observed for IFT88 in individual white samples ($p < 0.005$). IFT88 is thought to be involved in craniofacial, salivary gland and tooth development. Overall, this study strengthens that hypothesis that IFT88 influences caries risk.

** Eligible for post-doc poster award

Modified Random Forest Algorithm to Identify GxG Interaction in Matched Case-control and Case-Parent Trios Studies

**Qing Li[1], Emily Holzinger[1], Joan E. Bailey-Wilson[1]

[1]Statistical Genetics Section, Computational and Statistical Genomics Branch/National Human Genome Research Institute/NIH

Random forests (RF) is a machine-learning method useful to detect complex interactions among genetic markers related to a disease trait. Extensive theoretic work and applications of RF have been conducted for case-control samples. To extend this method to family-based genetic studies, we implement it in a flexible software framework that can easily facilitate a modified sampling scheme and feature selection criteria. In our implementation, we build an ensemble procedure that use an R package, rpart, which has a function to conduct classification tree analysis and we proposed different feature selection criteria (i.e. splitting criteria in a classification tree) to accommodate a matched case-control and case-parents trio design. For matched case-control, or case-parent trio data, we sample the set of samples (in a matched set, or matched case, pseudo-controls set) to be fit by each classification tree. Different classification criteria are also proposed to accommodate the matched study design. Then, we summarized results from all trees, and calculated the importance score to determine whether certain genetic variants are associated with susceptibility to the disease trait. In addition to existing criteria for classification, we proposed to use the likelihood ratio test from conditional logistic regression. To evaluate our method, we simulated matched case-control, and case-parent trio data, and applied our method to select the most important predictors. The results are compared with Logic Regression and trio Logic Regression.

** Eligible for post-doc poster award

Conservation of genomic characteristics in primary and contralateral breast tumors

**Maeve Mullooly[1,2], Ruth M Pfeiffer[1], Sarah J Nyante[3], Louise Brinton[1], Robert N Hoover[1], Andrew Glass[4], Amy Berrington de Gonzalez[5], Mark E Sherman[6], Gretchen L Gierach[1].

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [2]Cancer Prevention Fellowship Program, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, USA. [3]Department of Radiology, University of North Carolina School of Medicine, Chapel Hill, NC, USA. [4]Group Health Research Institute, Seattle, WA, USA. [5]Kaiser Permanente Northwest Center for Health Research, Portland, OR, USA. [6]Breast and Gynecologic Cancer Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, USA.

Among women diagnosed with breast cancer, contralateral breast cancers (CBC) are the most common second cancer diagnosed. Currently the natural history of CBC is unclear and almost always represents new cancers. It has been hypothesized that if CBC occurs randomly, the tumors will have unrelated molecular characteristics. Conversely, if activation of specific pathways is necessary for tumor formation, tumors will share molecular changes related to these pathways. To address this question and to study the molecular epidemiology of CBC, we developed an interdisciplinary collaboration with the Kaiser Permanente Northwest Health Plan. We identified women in this Plan, diagnosed with an invasive breast cancer between 1990 and 2008 and subsequently diagnosed with a CBC that consented to genetic research. Paired tumor blocks from the primary and CBC as well as detailed diagnosis and treatment information via medical record linkage, including clinical features, history of mammography, chemotherapy, radiotherapy and hormonal treatments have been collected where possible (n=83). Using tumor specimens, we aim to carry out whole genome sequencing to identify global somatic mutational signatures associated with the development of contralateral breast tumors. Further we aim to carry out a comprehensive comparison of the molecular profiles of primary and contralateral tumors to identify whether unique features are conserved in both, and to characterize the effect of treatments on the molecular profile of the contralateral tumor. We will present clinical characteristics of participants involved in this study and will present the proposed approach that we will use to carry out this study.

** Eligible for post-doc poster award

Does genetics really affect drug efficacy?

Matthew R. Nelson, Toby Johnson, Liling Warren, Arlene R. Hughes, Stephanie L. Chissoe, Chun-Fang Xu, Dawn M. Waterworth

GlaxoSmithKline, King of Prussia, PA

Lack of efficacy is the most common cause of attrition in late phase drug development. Many have believed that genetics could be widely used to drive stratified drug development by identifying and enrolling patients most likely to respond. However, a growing body of evidence demonstrates that only a small proportion of drugs have germline genetic predictors of efficacy with clinically meaningful effects, and so far they have only been found following drug approval. Because we cannot predict which drugs will have their efficacy influenced by clinically useful germline variants, we argue for early, routine, and cumulative screening for genetic efficacy predictors, as an exploratory part of clinical trial analysis. Such a strategy would result in the identification of any clinically relevant predictors that may exist at the earliest possible time, allowing them to be integrated into subsequent clinical development and into an assessment of patient benefit–risk. Although only a small proportion drugs are likely to have clinically useful predictors of genetic efficacy, such discoveries can provide mechanistic insights into drug disposition and patient–specific factors that influence response, and can therefore indirectly support related drug discovery and development efforts. We argue that drug developers should integrate genetics into the clinical development process to ensure that any major genetic factors affecting drug efficacy are recognized and acted on as appropriate.

Nedd4 Family Interacting Proteins limit effector T cell function by promoting JAK degradation

**C.E. O'Leary, C. Riling, G. Deng, L. Spruce, H. Ding, S. H. Seeholzer, and P.M. Oliver

Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, UNITED STATES; Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, UNITED STATES

Ubiquitylation tunes signaling pathways to regulate T cell function and differentiation. Nedd4 family catalytic E3 ubiquitin ligases have known roles in this process. These ligases can be autoinhibited, and are activated by binding Nedd4-family interacting protein 1 (Ndfip1) and Ndfip2. Ndfip1 negatively regulates T cell activation and Th2 polarization. The in vivo role of Ndfip2 is unknown. To investigate this, we generated Ndfip2-/- mice. Ndfip2 is not a dominant regulator of T cell activation or Th2 polarization, however, loss of Ndfip2 exacerbates the inflammatory Ndfip1-/- phenotype, suggesting that Ndfip2 dampens inflammatory processes. Ndfip1/Ndfip2 doubly deficient CD4+ T cells in mixed chimeras are more activated, more proliferative, and produce more cytokine than WT CD4+ T cells in the same host, indicating T cell intrinsic hyperactivation. We developed an unbiased proteomics approach to identify aberrantly ubqituitylated proteins in Ndfip deficient T cells consisting of enriching for polyubiquitylated proteins from Ndfip deficient and sufficient CD4 T cells using Tandem Ubiquitin Binding Entities (TUBEs), whole proteome analysis, and identification of post-translationally modified proteins in CD4+ T cells using ubiquitin remnant immunoprecipitation. This approach identified Jak1 as a ubiquitylated protein differentially regulated in the absence of Ndfip1/2. Jak1 is stabilized in stimulated Ndfip1/2 deficient CD4+ T cells, STAT5 shows increased phosphorylation in Ndfip deficient cells, and doubly deficient T cells are more resistant to Jak inhibition. We propose that, in stimulated T cells, Ndfips activate cognate E3 ligases to degrade Jak1, limiting a cytokine-mediated feed-forward loop that, if perturbed, leads to T cell hyperactivation.

Supported by grants from the NIH to Paula Oliver (R01 AI093566) and Claire O'Leary (T32 5T32GM007229, F31 CA180300)

** Eligible for post-doc poster award

Compound heterozygous mutations in *NEK8* cause end-stage renal disease with hepatic and cardiac anomalies

Ramakrishnan Rajagopalan[1], Christopher M. Grochowski[1], Melissa D. Gilbert[1], Alexandra M. Falsey[1], Kathleen M. Loomes[2,3], David A. Piccoli[2,3], Marcella Devoto[3,4,5,6], Nancy B. Spinner[1,7]

[1]Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA. [2]Division of Pediatric Gastroenterology, Hepatology, and Nutrition, The Children's Hospital of Philadelphia, Philadelphia, PA. [3]Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA. [4]Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA. [5]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA. [6]Department of Molecular Medicine, University La Sapienza, Rome, Italy. [7]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

We studied two brothers who presented in the newborn period with bile duct paucity, renal and cardiac anomalies which were initially suggestive of Alagille syndrome, although no mutation in Jagged1 or Notch2 was identified. Exome sequencing revealed compound heterozygous mutations in the *NEK8*, gene (never in mitosis A-related Kinase 8). The mutations included one frameshift in the last exon of the gene(exon 15), elongating the protein by 86 amino acids, and one missense mutation (T348M), which is localized in the highly conserved RCC1 domain. The RCC1 domain is involved in localization of the protein to the centriole. Mutations in NEK8 have been previously reported in one family with renal-hepatic-pancreatic dysplasia 2 (RHPD2), and in 3 individuals with nepronophthisis (NPHP9). *NEK8* is a ciliary kinase and has been shown to be indispensable for cardiac and renal development based on murine studies. This is the third report of disease causing mutations in the NEK8 gene in humans, although our patient does not have the same phenotype as seen in the previously published cases. The variable expressivity seen in these reported cases highlight the complexity of the genetic mechanisms involved. This report proves the ability of next-generation sequencing technology to provide a more precise diagnosis and highlights the importance of the use of exome sequencing as clinical diagnostic tool for rare disorders.

Linkage Analysis of Whole Exome Sequence Data in Multiplex Autism Families Including Cholesterol Covariates

Claire L. Simpson, Ph.D.

Statistical Genetics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health

The analysis of multiplex autism families may provide valuable insights into the risk of developing autism spectrum disorder (ASD). It is not known whether “sporadic” ASD is truly sporadic or a consequence of reduced penetrance, but our previous analyses of *de novo* variation suggest that multiplex families have a different underlying etiology. It has also been observed that individuals with Smith-Lemli-Opitz syndrome (SLOS) show many autistic features and abnormal cholesterol measures have been demonstrated in individuals with ASD who do not have SLOS. We analyzed whole exome sequence (WES) data in 69 multiplex ASD families from the Autism Genetic Resource Exchange collection using a linkage analysis method that incorporate covariates, to see if some families show stronger evidence of linkage to particular loci in the presence of these covariates. Significant evidence of linkage to ASD when including covariates was found on chromosomes 1 ($p=2.2x10^{-7}$), 2 ($p=7.2x10^{-6}$) and 21 ($p=6.1x10^{-6}$) for hypercholesterolaemia and chromosomes 1 ($p=1.1x10^{-7}$), 2 ($p=1.6x10^{-6}$), and 21 ($p=5.1x10^{-6}$) with hypocholesterolaemia. Additional regions with suggestive evidence of linkage of ASD were also observed for all covariates except ApoB. The signals on chromosomes 1 and 2 are close to known linkage peaks, in particular the chromosome 2 signal is in the 2q21-q33 region which has been repeatedly reported in linkage analyses of ASD. The chromosome 21 locus is close to a previously reported region in ASD families with language regression. These regions are being examined in detail to attempt to identify the linked causal variants.

Computational validation of NGS variant calls using genotype data

Margaret A. Taub, Suyash Shringarpure, Rasika A. Mathias, Ryan Hernandez, Timothy D. O'Connor, Zach A. Szpiech, Raul Torres, Francisco M. De La Vega, Carlos Bustamante, Kathleen C. Barnes

Department of Biostatistics, Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD

Statistical challenges in working with whole genome sequencing data sets arise at all phases of data analysis, from initial measurement of data quality to development of appropriate methods for testing genetic hypotheses and interpreting observed patterns of genetic variation. Variant calling from next-generation sequencing data is susceptible to false positive calls due to sequencing, mapping and other errors. We present a method that uses machine learning techniques, specifically Random Forests, for computationally validating variant calls obtained from a sample of individuals. While existing methods use site quality information from known samples such as HapMap and dbSNP for training classifiers to distinguish between true and false variant calls, our method uses genotype data from the same samples, typically collected for all sequenced individuals, to learn a more accurate classifier. We demonstrate our method on variant calls obtained from 643 high-coverage African-American genomes from the The Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA), sequenced to high depth (~30X). On variant calls obtained using Illumina's single-sample caller CASAVA, our method has a True Positive Rate of 97.5% (at a False Positive Rate of 5%). On variant calls obtained from Real Time Genomics' multisample variant caller, our method obtains a True Positive Rate of 95% (at a False Positive Rate of 5%). We applied our classifier to compare call sets generated with different calling methods, including both single-sample and multi-sample callers, and found that allele frequency is an important determinant of which calling method makes the most accurate calls.

Major-Effect Loci for Lipids also Impact Phenotype Variability in the Old Order Amish

Laura Yerges-Armstrong and Jeffrey O'Connell

Program in Personalized and Genomic Medicine, and Department of Medicine, Division of Endocrinology, Diabetes and Nutrition - University of Maryland School of Medicine, USA. Baltimore, MD 21201

Association studies for complex traits have traditionally focused on identifying loci that shift the mean of a quantitative trait. Several recent papers describe methods for modeling genetic loci that instead impact trait variability. These phenotypic variance quantitative trait loci (or vQTL) are of particular interest as they may indicate undetected genetic or environmental interaction. With this in mind, we tested three loci for lipids segregating in the Old Order Amish (OOA) for the presence of vQTL effects. All analyses were conducted using our mixed models analysis for pedigrees and populations (MMAP) software with Levene's Median F-test implemented to model vQTL while accounting for polygenic effects, age and sex. The first variant tested was the *APOC3* null mutation (R19X) associated with markedly lower mean triglycerides ($p=2x10^{-23}$). In addition to the large effect on the mean there was a significant vQTL effect ($p=4.6x10^{-6}$). The second variant was the common, promoter variant in *CETP* (rs3764261), which was highly significant for mean differences in high-density lipoprotein cholesterol ($p=2.6x10^{-10}$) but had only a modest vQTL effect (p=0.02). Due to a founder effect ~10% of Amish are carriers of the APOB3500Q allele which is associated with ~60mg/dL higher low-density lipoprotein (LDL) cholesterol ($p=1.2x10^{-94}$). We observed a significant vQTL (p=1.3x10-6) for LDL with this variant. To our knowledge, the vQTL effects for R19X and R3500Q are the first reported for lipids. The lack of vQTL effect for CETP could indicate that highly-penetrant, rare variants are more likely to be vQTLs but additional modeling is needed.

Harnessing a hefty set of whole genome sequencing data with an automated workflow

Zhe Zhang, Ariella Sasson, Deanne Taylor, Elizabeth Goldmuntz

Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, and the Pediatric Cardiac Genomics Consortium

Congenital heart disease (CHD) is the most common birth defect and the leading cause of death in infants and young children. The Pediatric Cardiac Genomics Consortium has generated a data set of ~6,700 WES samples (1,243 case trios and 900 control control trios), which have been sequenced and processed into a wide VCF file of approximately two million variants. Manipulating VCF in such size is computationally challenging. We implemented an automated workflow to facilitate the processing, quality control, and statistical analysis of this data set. This workflow is composed of three major stages that can be run separately or together. Stage 1 loads the whole VCF file and converts genotype calls to an integer matrix after which the variants were annotated and filtered by call rate, sequencing depth, etc. The loaded data was then saved in formats supporting fast access, such as R integer matrix and indexed database table. Stage 2 analyzes the samples to identify potential quality issues, sample mislabeling, and unknown confounders. We used methods such as PCA and IBD to identify un-reported kinship, gender swapping, etc. Stage 3 implements several statistical tests to compare case and control groups at variant, gene, and pathway levels. It supports parallel computing to facilitate the process. This workflow has been successfully applied to several sub-projects, demonstrating much improved computational efficiency. It laid the groundwork for efficient processing and analysis of large WES cohorts we will continuously encounter in the future.

Test of Genotypic Association Allowing for Sequencing Misclassification

*Lisheng Zhou

Department of Genetics, Rutgers University, Piscataway, NJ 08854

Genome-wide association studies (GWAS) have led to identification of an ever-increasing number of single nucleotide polymorphisms (SNPs) for further studies. However, the presence of genotype sequencing misclassification, differential or non-differential, among cases and controls may cause either an increase in the false positive rate or a decrease in the power of statistical tests. A commonly used statistic to test for association between SNPs in cases and controls is the chi-square test of independence. This statistic tests whether genotype frequencies (single or multi-locus) differ between case and control groups. This is commonly referred to as a test of association. Regions of the genome where frequencies significantly differ are regions that may harbor a disease locus or loci. Several researchers have documented that genotype misclassifications may alter either the null or alternative distribution of the chi-square test.

Our overall objective is the development of a statistical test of association that uses NGS data and is robust to random sequencing misclassifications, both non-differential and differential. More concretely, we develop a chi-square test of independence that uses parameters such as the observed alternative variant reads at a given polymorphic site, the coverage per individual, the individual's phenotype, and error model parameters. This statistic is developed in a log-likelihood framework and the Bootstrapping approach is applied to approximate the type I error.

* Eligible for pre-doc poster award