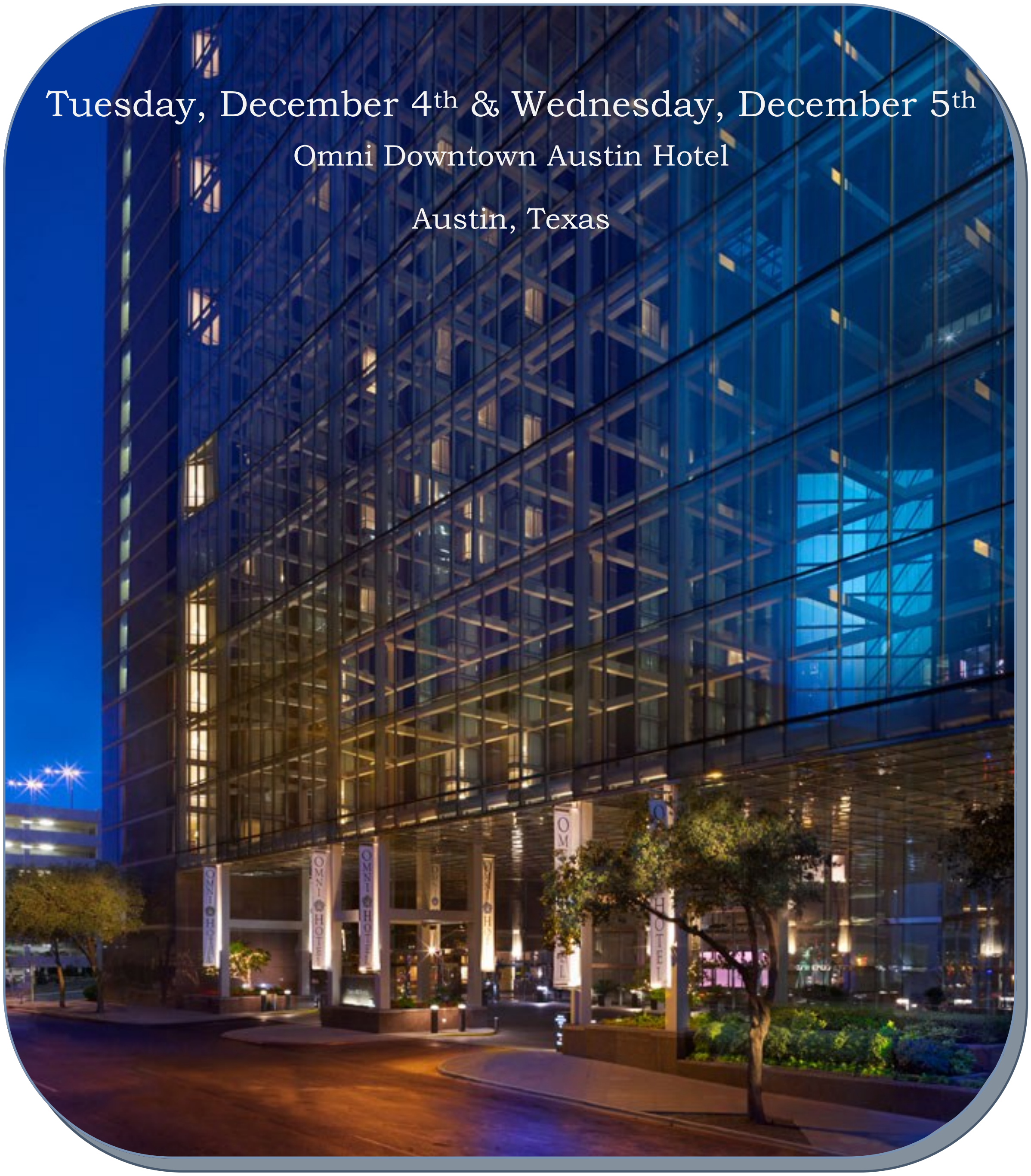
P-STAR Analysis Workshop 2012



Tuesday, December 4th & Wednesday, December 5th

Omni Downtown Austin Hotel

Austin, Texas

www.pgrn-star.org

P-STAR Analysis Workshop – 2012

Tuesday, December 4th & Wednesday, December 5th
Austin, Texas

Tuesday Agenda		
7:00-8:00	*Registration and Breakfast*	
8:00-8:30	*Welcome and Introductions*	Marylyn Ritchie, P-STAR
8:30-9:10	Integration of cell line and clinical trial genome-wide analyses implicates multiple loci in paclitaxel-induced peripheral neuropathy	Heather Wheeler, PAAR
9:10-9:50	ATHENA: A method for integrating genome-wide gene expression and genotype data to generate meta-dimensional prediction models in CAP	Emily Holzinger, P-STAR
9:50-10:20	*Break*	
10:20-11:00	The Convergence of Functional Genomics, Heritability Estimation and Polygenic Modeling	Eric Gamazon, PAAR
11:00-11:40	Phenotype-Specific Genomic Network Discovery	Cheng Cheng, PAAR4Kids
11:40-12:20	Polygenic Inheritance of Paclitaxel-Induced Peripheral Neuropathy Driven by Axon Outgrowth Gene Sets	Aparna Chhibber, PMT
12:20-1:20	*Lunch*	
1:20-2:00	Estimating Heritability Of Drug Induced Liver Injury From Genome-wide Common Variants	Casey Overby, Affiliate
2:00-2:40	Polygenic Prediction of complex phenotypes and drug response	Hae Kyung Im, PAAR
2:40-3:10	Break	
3:10-3:50	Using BioBin to Explore Rare Variant Population Stratification Using 1000 Genomes Project Data	Carrie Moore, Affiliate
3:50-4:50	Statistical methods for association testing with multiple outcomes and (pharmacological) responses	Matthew Stephens, Department of Human Genetics and Department of Statistics, The University of Chicago
4:50-5:00	Closing notes for the day	
5:00-6:00	*Cocktail Hour*	
6:00	*Dinner – See Sign-up sheets at the reception table*	

Wednesday Agenda		
7:00-8:00	*Breakfast*	
8:00-9:00	Pharmacogenetic sequencing at the BCM-HGSC	Dr. Steve Scherer, Human Genome Sequencing Center, Baylor College of Medicine
9:00-9:40	PGRN RNA-seq Pilot Analysis	Xiang Qin, BCM-HGSC
9:40-10:10	*Break*	
10:10-10:50	Continuing challenges in analysis of RNA-seq data	Courtney French, PMT
10:50-11:30	A Study of Asthma Pharmacogenomics Using RNA-Seq	Blanca Himes, PHAT
11:30-12:10	Drug Metabolism and Drug Interactions	Joseph Kitzmiller, XGEN
12:10-1:30	*Lunch and General Discussion*	

Dining & Entertainment in Austin!

A huge thank you to Steve Scherer for putting together this list of dining, entertainment and sightseeing options

# - sign-up sheet at the registration desk! * = recommended by real live Austinites!

Dining:

Nearby:

# * Mekong River - 215 East 6th Street (1.5 blocks – right down the street from the hotel) Thai and Vietnamese cuisine.

# * Bess Bistro – 500 W. Sixth St. (7 blocks; 0.5 miles) Sandra Bullock's place; American Contemporary, French, Southern

Walton's Fanncy and Staple – 609 W. Sixth St. (7 blocks; 0.5 miles); From the same folks that run Bess; More of a brunch/lunch place.

# Hut's Hamburgers - 807 W 6th St. 9 blocks; 0.6 miles); Old fashioned hamburgers.

# Moonshine Patio Bar & Grill - 303 Red River St. (7 blocks; 3rd Ave. and Red River St.; 0.5 miles) Texas cuisine like chicken fried steak, etc.

# Max's Wine Dive - 207 San Jacinto St. (4 blocks straight downhill south toward the lake); Comfort food and wine; famous for their egg sandwich.

A little further to walk:

Papi Tinos - 1306 East 6th St. (12 blocks; 0.8 miles walking east along 6th St. under the freeway and a few blocks beyond); Contemporary Mexican - good atmosphere. Good tequilas and mescals. For further drinks, check out Rio Rita across the street.

* Clay Pit – 1601 Guadalupe St. (13 blocks; 1 mile; on the other side of the capitol Building) Contemporary Indian.

* Texas Chili Parlor - 1409 Lavaca St. (10 blocks; 0.8 miles); Real Texas chili – not exactly health food, but delicious!

* Green Mesquite – 1400 Barton Springs Rd. (1.7 miles; other side of the lake, but a nice walk) Texas BBQ and more.

Probably want to taxi but probably worth it:

* Ruby’s BBQ - 512 West 29th St. (2.4 miles; it’s a taxi ride up to the north side of the UT campus) Probably the best local BBQ in Austin. The best of the best is 30 minutes away in Lockhart

* Perla - 1400 South Congress Ave. (1.6 miles; another taxi, but really good!) Seafood and Oyster Bar; if you’re an oyster lover, Gulf Coast oysters cannot be beat.

* Vespaio - 1610 South Congress Ave. (1.8 miles; taxi) Italian

Various Food Trucks – Austin is a city of food trucks and trailers; many of the best can be found on South First St. (not far from Perla or Vespaio above; next busy street to the west – taxi recommended) or on Barton Springs Road (a little further west of Green Mesquite above).

Music:

check out austinlivemusic.com for complete listings. Austin trails just NY, LA, Nashville and possibly New Orleans as an American music hub and recording mecca.

Stubb’s Barbecue - 801 Red River St.; Mostly Rock

Emo’s - 2015 East Riverside Dr.; Progressive

ACL Live - 310 West Willie Nelson Boulevard; Texas singer/songwriter

Antones - 213 West 5th St.; Blues

Cactus Café - 2247 Guadalupe St.; up near the UT campus; various artists

Hole in the Wall - 2538 Guadalupe St.; also up north of the UT campus on “the drag”; various artists

Continental Club - 1315 South Congress Ave.; The “Granddaddy of Austin Music Clubs”.

Sightseeing:

Lady Bird Lake (Town Lake) Just walk six blocks downhill to the lake and go left to Congress Ave. to find the trail. Great running/walking trail on both sides of the lake – goes forever.

Congress Ave. Bridge Bats – see above for directions; Every evening at dusk, about 1.5 million Mexican free-tail bats emerge to hunt. December may be a little late in the season, so I’ll check.

Texas State Capitol Building – go two blocks west to Congress and turn right. The building is generally open for a look around.

Integration of cell line and clinical trial genome-wide analyses implicates multiple loci in paclitaxel-induced peripheral neuropathy

Heather E. Wheeler[1], Eric R. Gamazon[2], Claudia Wing[1], M. Eileen Dolan[1], Nancy J. Cox[2]

[1]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL
[2]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL

Our goal is to understand to what extent cell-based models can capture the overall genetic architecture of patient chemotherapy response. When comparing modestly sized genome-wide association studies from patients and lymphoblastoid cell lines (LCLs) treated with the same drug, SNPs rarely overlap at stringent thresholds such as $P < 10^{-6}$, but significant overlap of SNPs at more relaxed thresholds determined by enrichment analysis through random sampling are possible. Under this cumulative hypothesis, large numbers of common variants with small effects account for substantial heritability. As a first approach, we asked whether SNPs with nominally significant associations with paclitaxel-induced cytotoxicity or apoptosis in studies in HapMap LCLs (n=247) are enriched among the top SNPs associated with paclitaxel-induced peripheral neuropathy in patients. We observed an enrichment of LCL cytotoxicity-associated SNPs in the peripheral neuropathy-associated SNPs from breast cancer patients (n=855) treated with paclitaxel in the Cancer and Leukemia Group B (CALGB) 40101 trial (empirical P = 0.007). In addition, we observed an enrichment of LCL apoptosis-associated SNPs in the peripheral neuropathy-associated SNPs from ovarian and lung cancer patients (n=143) treated with paclitaxel and carboplatin (empirical P = 0.028). Importantly, permutations were conditioned on concordant SNP direction of effect, minor allele frequency and patient genotyping platform. Both sets of overlap SNPs were enriched in expression quantitative trait loci (eQTLs), demonstrating the potential importance of this functional class in chemotherapeutic response. One of these eQTLs is located in *RFX2* and decreased expression of this gene by siRNA resulted in increased sensitivity of NS-1 cells to paclitaxel measured by reduced neurite outgrowth and increased cytotoxicity, functionally validating the involvement of *RFX2* in paclitaxel sensitivity and supporting our multi-gene hypothesis. This robust enrichment demonstrates that susceptibilities to increased cytotoxicity/apoptosis in LCLs and increased sensory peripheral neuropathy in cancer patients likely have some genetic mechanisms in common and supports the role of LCLs as a preclinical model for paclitaxel toxicity studies.

ATHENA: A method for integrating genome-wide gene expression and genotype data to generate meta-dimensional prediction models in CAP

E. Holzinger[1,2], S. Dudek[1,2], A. Frase[2], M. Medina[3], R. Krauss[3], M. Ritchie[2]

1) Ctr Human Gen Res, Vanderbilt Univ, Nashville, TN; 2) Pennsylvania State University, University Park, PA; 3) Children's Hospital Oakland Research Institute, Oakland, CA

Technology is driving the field of human genetics research with advances in the ability to generate high-throughput data that interrogate various levels of biological regulation (genomic, transcriptomic, proteomic, etc). With this massive amount of data comes the important task of using powerful and creative bioinformatics techniques to sift through the noise to find true, meaningful signals that predict various human traits. A popular analytic method thus far has been the genome-wide association study (GWAS), which assesses the association of each DNA variation with the trait of interest. Unfortunately, GWAS has not been able to explain a substantial proportion of the estimated heritability for most common, complex traits. Due to the inherently complex nature of biology, this phenomenon could be a factor of the simplistic GWAS study design. A more powerful study design for this data may be a systems biology approach that integrates different types of data, or a meta-dimensional analysis.

Our method is a two-step, filtering-modeling process. First, to reduce the noise in the data set, we use a statistical filtering method that allows for main and interaction effects. Next, to generate more parsimonious meta-dimensional models, we analyze the filtered set of variables using the Analysis Tool for Heritable and Environmental Network Associations (ATHENA). We performed a proof-of-concept analysis to demonstrate that our method can detect "known" loci associated with HDL cholesterol and triglyceride levels using a data set with only SNP predictor variables. Next, we applied this method to the CAP data set which consists of ~2.8 million SNPs and ~25,000 gene expression variables from 480 individuals that participated a simvastatin clinical trial to generate meta-dimensional models that predict changes in LDL cholesterol before and after simvastatin treatment.

With this systems biology approach, we were able to integrate different types of high-throughput data to generate meta-dimensional models that are predictive for the various lipid trait outcomes in our data sets. Importantly, these models should be tested in independent data sets to show that they replicate and to obtain more reliable prediction estimates.

The Convergence of Functional Genomics, Heritability Estimation and Polygenic Modeling

E. R. Gamazon[1], H. K. Im[1], C. Liu[2], D. L. Nicolae[1], N. J. Cox[1]

1) University of Chicago, Chicago, IL, USA; 2) University of Illinois, Chicago, IL, USA.

It is widely held that a substantial genetic component underlies Bipolar Disorder and other neuropsychiatric disease traits. Recent efforts have been aimed at understanding the genetic basis of disease susceptibility, with genome-wide association studies (GWAS) unveiling some promising associations. Nevertheless, the genetic etiology of Bipolar Disorder remains elusive with a substantial proportion of the heritability - which has been estimated by ourselves and others to be 80% (congruent with previous reports based on twin and family studies) - unaccounted for by the specific genetic variants identified by large-scale GWAS. Furthermore, functional understanding of associated loci generally lags discovery. Studies we report here provide considerable support to the claim that substantially more remains to be gained from GWAS on the genetic mechanisms underlying Bipolar Disorder susceptibility, and that a large proportion of the variation in disease risk may be uncovered through integrative functional genomic approaches. We set out to combine recent analytic advances in heritability estimation and polygenic modeling and leverage recent technological advances in the generation of -omics data to evaluate the nature and scale of the contribution of functional classes of genetic variation to a relatively intractable disorder. We identified 2,375 cis-acting eQTLs in cerebellum that account for 36% (s.e. = 0.024) of the heritability, which represents 65% of the total heritability attributable to SNPs interrogated through GWAS. Our findings show that a much greater resolution may be attained than has been reported thus far on the number of common loci that capture a substantial proportion of the heritability to disease risk and that, importantly, the functional nature of contributory loci may be clarified *en masse*.

Phenotype-Specific Genomic Network Discovery

Cheng Cheng[1]

[1]St. Jude Children's Research Hospital, Memphis, TN, USA

Ultra-high dimensionality is an immediate challenge in detection of molecular association networks involving several types of genomic factors, all measured genome-wide. Searching for all possible relationships among the measured genomic factors is an NP-hard problem; therefore in practice proper dimension reduction is necessary. In a cancer genomic study there is often a specific biological context defined by one or more phenotypes of interest; for example, to search for inter-related genomic factors jointly affecting the "response to remission induction" (phenotype) in a cohort of uniformly treated patients. The specific biological context provides an opportunity to perform the biologically meaningful, computationally effective Phenotype-Driven Dimension Reduction (PhDDR). This presentation will describe the development of the PhDDR approach in the context of detecting genomic networks associating gene co-expressions and single nucleotide polymorphisms (SNPs), and discuss the extensions to integration of other types of genetic and epigenetic factors. Briefly, the PhDDR approach starts with one type of genomic factors such as gene (mRNA) expressions that are significantly associated (by appropriate statistical significance criteria) with the phenotype(s) of interest, builds local gene co-expression clusters, and then extends the co-expression clusters to incorporate other types of genomic factors. In this way the difficulties of ultra-high dimensionality in integrating massive numbers of different types of genomic factors can be effectively circumvented, and the resulting networks are all naturally pertinent to the biological problem under investigation, or to the generation biological hypotheses about the process underlying the phenotype(s). This approach is illustrated by an application to a genomic association analysis of treatment response of childhood leukemia involving gene expressions and SNPs.

Polygenic Inheritance of Paclitaxel-Induced Peripheral Neuropathy Driven by Axon Outgrowth Gene Sets

A. Chhibber[1,2], J. Mefford[3], S.A. Pendergrass[7], R.M. Plenge[4,5,6], M.D. Ritchie[7], E.A. Stahl[4,5,6], J.S. Witte[2,3], D.L. Kroetz[1,2].

1) Bioeng. & Therapeutic Sci, Univ Cal San Francisco, San Francisco, CA; 2) Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA; 3) Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA; 4) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA; 5) Division of Genetics, Brigham and Women's Hospital, Boston, MA, USA; 6) Division of Rheumatology Immunology and Allergy, Brigham and Women's Hospital, Boston, MA, USA; 7) Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN , USA.

Peripheral neuropathy (PN) is a common and often dose-limiting toxicity for patients treated with paclitaxel. While diabetes or previous exposure to other chemotherapeutics with similar toxicity increase susceptibility to paclitaxel-induced neuropathy, for the vast majority of individuals there are no known risk factors that predispose patients to the adverse event. Further, pathogenesis for paclitaxel-induced neuropathy is unknown, though studies in animal and cell models have provided some potential mechanisms. Determining whether there is a heritable component to paclitaxel induced PN would be valuable in guiding clinical decisions and may provide insight into treatment of and mechanisms for the toxicity. Using genotype and patient information for the paclitaxel arm of CALGB 40101, a Phase III clinical trial comparing efficacy of single-agent paclitaxel with the current standard regimen as adjuvant therapy for breast cancer in women, we estimated the variance in maximum grade of PN and dose at first instance of PN explained by all autosomal SNPs, SNPs selected based on functional effect or location, and SNPs in gene sets selected based on prior knowledge regarding possible mechanisms of the pathogenesis of paclitaxel-induced PN. DNA was isolated and genotyped using the Illumina 610-Quad platform. Following QC and Principal Component analysis, a total of 849 genetic Europeans were included in our analyses. The whole genome, SNP function/location, and pathway-specific heritability analyses were conducted using the GCTA software tool. While whole genome estimates of heritability were not significant, estimates captured by genic SNPs suggest that paclitaxel-induced neuropathy does indeed have a significant heritable component. Further, of the seven gene sets evaluated, the Axonogenesis GO Term (GO: 0007409) set had significant estimates of heritability close to 20%. Further evaluation of GO sets included within the Axonogenesis set suggests a large portion of this heritability is driven by genes involved in the regulation of axon extension. These results suggest that paclitaxel-induced neuropathy does in fact have a significant heritable component, and that this heritability is driven in part by genes involved in axon outgrowth. Disruption of axon outgrowth may be one of the primary mechanisms by which paclitaxel treatment results in PN in susceptible patients.

Estimating Heritability Of Drug Induced Liver Injury From Genome-wide Common Variants

Casey Lynnette Overby[1], George Hripcsak[1], Yufeng Shen[1]

[1]Department of Biomedical Informatics, Columbia University

Background: The heritability of complex traits is traditionally estimated through twin and family studies. In the context of low prevalence pharmacological traits, however, it is difficult to recruit and obtain clinical outcome data in families. This work investigates for a low prevalence pharmacological traits of moderate sample size, our ability to estimate the heritability (h^2) from genome-wide single nucleotide polymorphism (SNP) data. The trait of focus in this work was drug-induced liver injury (DILI).

Methods: DILI datasets were from case-control studies of individuals taking flucloxacillin (77 cases and 288 controls)[1] and taking co-amoxiclav (201 cases and 532 controls)[2]. We used the GCTA algorithm[3] to estimate h^2 from iSAEC DILI genome-wide data (all chromosomes) and chromosome 6 data. We also estimated h^2 for all individuals (co-amoxiclav or flucloxacillin induced DILI), for individuals with co-amoxiclav induced DILI, and for individuals with flucloxacillin induced DILI. For individuals with co-amoxiclav induced DILI we evaluated northwest Europeans only and southern Europeans both together and separately. For each population and dataset, we calculated h^2 adjusting for prediction errors due to global structure and local structure.

To evaluate the robustness of the GCTA algorithm with datasets of moderate sample sizes, we estimated h^2 for subsets of cases and controls from a Type I Diabetes (T1D) dataset. We used the Welcome Trust Case Control Consortium T1D dataset (1963 cases and 3004 controls)[4]. We estimated h^2 from genome-wide data (all chromosomes) and chromosome 6 data for cases and controls with a 1:1 ratio and up to a 1:15 ratio, where the number of cases ranged from 50 to 500. Estimates for h^2 were averaged over five random selections and two random selections of cases and controls for sample sizes ranging from 50 to 75 and 150 to 500, respectively.

Results: We estimate the proportion of h^2 captured by common SNPs for DILI to be between 0.331 and 0.477 after stratifying by population (northwest Europeans vs. southern Europeans). Estimates depend on the drug implicated. For flucloxacillin induced DILI patients, chromosome 6 explained almost all of the heritability ($h^2_{\text{all}} = 0.477$, $h^2_{\text{chr6}} = 0.477$). Where with co-amoxiclav induced DILI patients, chromosome 6 explained part of the heritability ($h^2_{\text{all}} = 0.399$, $h^2_{\text{chr6}} = 0.169$), indicating additional contributions from common variants yet to be found. Varying the number of cases from 50 to 500, estimates for h^2 captured by common SNPs for T1D ranged from 0.334 and 0.414. The estimated proportion of h^2 captured by chromosome 6 SNPs ranged from 0.048 and 0.164, which is likely underestimated due to inadequate SNP density in the MHC region from that data set.

Conclusion: Findings suggests that continuous collection of DILI cases is valuable for the potential of discovering additional associations. Our assessment of the impact of sample size on heritability estimates indicated that estimates remain relatively stable with varying sample size. While further investigation is required to confirm the robustness of the GCTA algorithm for low prevalence traits such as DILI, this work highlights the potential value of its application. Here we were able to apply the algorithm to investigate the contribution of genome-wide variants on the chromosome level and within different populations. Such investigations can provide insight into disease mechanism and into inter-individual variation. Moreover, for adverse drug reactions in particular, we are able to provide previously unobtainable estimates of heritability. Such estimates will help optimize designs of future studies for identifying additional genetic contributions to these conditions.

Acknowledgments: This work is supported in part by the Columbia Training in Biomedical Informatics (NIH NLM #T15 LM007079) and NIH NIDDK 1D3DK094400-01. This study makes use of data generated by the Wellcome Trust Case Control Consortium and by the International Serious Adverse Event Consortium.

1. Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A, Daly MJ, Goldstein DB, John S, Nelson MR *et al*: HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat Genet* 2009, 41(7):816-819.

2. Lucena MI, Molokhia M, Shen Y, Urban TJ, Aithal GP, Andrade RJ, Day CP, Ruiz-Cabello F, Donaldson PT, Stephens C *et al*: Susceptibility to amoxicillin-clavulanate-induced liver injury is influenced by multiple HLA class I and II alleles. *Gastroenterology* 2011, 141(1):338-347.

3. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al*: Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010, 42(7):565-569.

4. Nejentsev S, Howson JM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J *et al*: Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature* 2007, 450(7171):887-892.

Polygenic Prediction of complex phenotypes and drug response

Hae Kyung Im[1], Eric R Gamazon[2], R Stephanie Huang[2], Nancy J Cox[2]

1 Department of Health Studies, University of Chicago 2 Department of Medicine, University of Chicago

The set of common genetic variations interrogated by genome-wide association studies (GWASs) have been shown to explain a substantial proportion of the heritability of complex traits. The linear mixed modeling approach (and bayesian generalizations of it) can provide relatively accurate estimates of the combined effects of many common variants even though the individual effect sizes cannot be determined with the same level of accuracy. Our goal is to utilize this framework to predict complex phenotypes. We use a training set to estimate the parameters needed for prediction, such as the relatedness matrix and the heritability explained by an additive genetic model. Makowski et al. have shown that prediction performance is lower than expected from the estimated heritability because of estimation errors. In order to make the prediction algorithm clinically relevant, we need to further improve our estimation procedures. In general, all typed variants from GWASs are used to estimate heritability. It is clear that using only causative variants would be ideal; consequently the use of variants that are more likely to be functional should improve the estimation. In fact, we have reported that using only genetic variants associated with gene expression levels (eQTLs) substantially improves the estimation (Gamazon et al. under review). Other functional prior information can be used to filter out uninformative variants. We will show application of our prediction algorithm to cellular drug sensitivity and clinical drug response.

Using BioBin to Explore Rare Variant Population Stratification Using 1000 Genomes Project Data

Carrie B. Moore, John R. Wallace, Alex T. Frase, Daniel Wolfe, Sarah A. Pendergrass, Ken Weiss, Marylyn D. Ritchie

Rare variants (RVs) will likely explain additional heritability of many common complex diseases; however, the natural frequencies of rare variation across and between human populations are largely unknown. We have developed a powerful, flexible collapsing method called BioBinthat utilizes prior biological knowledge using multiple publicly available database sources to direct analyses. Variants can be collapsed according to functional regions, evolutionary conserved regions, regulatory regions, genes, and/or pathways without the need for external files. We conducted an extensive comparison of rare variant burden differences (MAF < 0.03) between fourteen ancestry groups from 1000 Genomes Project data, we found that on average 24.46% of gene bins, 32% of intergenic bins, 42.23% of pathway bins, 12.93% of ORegAnno annotated bins, and 4.72% of evolutionary conserved regions (shared with primates) have statistically significant differences in RV burden. Ongoing efforts include examining additional regional characteristics using regulatory regions and protein binding domains. Our results show interesting variant differences between two ancestral populations and demonstrate that population stratification is a pervasive concern for sequence analyses.

PGRN RNA-seq Pilot Analysis

Xiang Qin[1], Steve Scherer[1], Michael Metzker[1], Hsu Chao[1], Harsha Doddapaneni[1], Donna Muzny[1], and Richard Gibbs[1].

[1]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030

One of the primary goals of the Pharmagenomics Research Network (PGRN) is to identify the correlation of genomic variants with differences in drug response. Advances in sequencing technologies (massively parallel sequencing) provides an affordable approach to generate deep sequences derived from total RNA (RNA-seq) with far greater dynamic range than what is possible with more traditional solid-phase chip systems. RNA-seq allows us to comprehensively annotate and quantify the expression of all genes and isoforms from samples, and ultimately correlate expression profiles with genome variants and drug response, thus identifying genetic factors and mechanisms contributing to individual drug effects. The Human Genome Sequencing Center at Baylor College of Medicine (BCM-HGSC), in collaboration with multiple members of the PGRN has completed the sequencing of 25 samples from each of 4 human tissues (heart, liver, kidney and adipose). These efforts were undertaken to pilot RNA-seq protocols, establish quality control metrics standards, provide baseline expression profiles and spur development of downstream analysis routines. Using RNA-seq data from these samples, we established RNA-seq quality control standards from sample intake through sequencing. Various tools for read mapping as well as gene and isoform expression analysis were evaluated to assess their performance and accuracy. Further evaluation of tools for transcriptome reconstruction is ongoing. The evaluations provide insights into the possible strengths and weaknesses of these tools, which will facilitate the development of protocols for continuing RNA-seq analysis within the PGRN.

Continuing challenges in analysis of RNA-seq data

Courtney E. French[1], Steven E. Brenner[1]

[1]University of California, Berkeley

The PGRN RNA-seq Project is generating and analyzing RNA-seq data for dozens of samples from multiple human tissues. For gene expression analyses, RNA-seq has many advantages over microarray experiments, including a larger dynamic range, greater linearity, and the ability to discover novel genes and splicing events. However, there are inherent biases in RNA-seq data and non-trivial challenges in analysis, many of which have yet to be adequately addressed. There exist many analysis tools, and choosing the best one depends on the purpose of the experiment, but ultimately all have limitations.

Over the course of analyzing RNA-seq data, including data from the PGRN RNA-seq Project, we have identified continuing issues that broadly affect RNA-seq analysis and should be borne in mind when interpreting the results. Using the Cufflinks suite [1] for gene- and transcript-level expression analysis and a tool developed in our group, JuncBASE [2], for splice junction-based analysis, we are able to discover tissue-specific gene expression and alternative splicing events. Determining transcript level expression continues to be difficult due to the uncertainty in assigning reads to a particular isoform from an alternatively spliced gene. We have also found that not all transcript assemblies reported by Cufflinks are confidently supported by the underlying reads. Additionally, because analysis tools continue to be rapidly developed, different versions produce radically different results (up to a 5x difference in the fraction of up-regulated isoforms in one study). Thus, in order to compare across experiments, the same version should be used for all analysis in a given project. Finally, we present the idea of using a ‘virtual’ reference as a potential solution, when suitable, for situations where there is no appropriate control for changes in gene expression, i.e., when comparing across multiple tissues.

[1] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28:511–515 (2010)

[2] Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, Brenner SE, Graveley BR. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research.* 2:193-202 (2011) http://compbio.berkeley.edu/proj/juncbase/

A Study of Asthma Pharmacogenomics Using RNA-Seq

Blanca E Himes[1,2,3], Peter Wagner[4], Ruoxi Hu[4], Barbara Klanderman[2], William Jester[5], Martin Johnson[5], Reynold A Panettieri Jr[5], Kelan G Tantisira[1], Scott T Weiss[1,2], Quan Lu[4]

[1]Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; [2]Partners Center for Personalized Genetic Medicine, Boston, MA [3]Children's Hospital Informatics Program, Boston, MA; [4]Program in Molecular and Integrative Physiological Sciences, Department of Environmental Health, Harvard School of Public Health, Boston, MA; [5]Pulmonary, Allergy and Critical Care Division, University of Pennsylvania, Philadelphia, PA, USA

Asthma is a chronic inflammatory airway disease with well-established heritability that affects over 300 million people worldwide. The most common medications used for the treatment of asthma are β_2-agonists and glucocorticosteroids, and one of the primary tissues that these drugs target in the treatment of asthma is the airway smooth muscle. We used RNA-Seq, a high-throughput sequencing method that provides comprehensive expression analysis, including discovery of non-coding transcripts and splice variants, to characterize the human airway smooth muscle (HASM) transcriptome at baseline and under three asthma treatment conditions. The Illumina TruSeq assay was used to prepare libraries for HASM cells from four white male donors under four conditions: 1) no treatment; 2) treatment with a β_2-agonist (i.e. albuterol, 1μM for 18h); 3) treatment with a glucocorticosteroid (i.e. dexamethasone, 1μM for 18h); 4) simultaneous treatment with a β_2-agonist and glucocorticoid, and the libraries were sequenced with an Illumina Hi-Seq 2000 instrument. The Tuxedo Suite Tools were used to align reads to the hg19 reference genome, assemble transcripts, and perform gene-based and transcript-based differential expression analysis. Based on a Benjamini-Hochberg corrected p-value <0.05, there were 205 differentially expressed genes for the Dex vs. Untreated comparison, 16 for the Albuterol vs. Untreated comparison, and 37 for the Dex+Albuterol vs. Untreated comparison. Most of the significant genes in the latter two groups were also significant in the Dex vs. Untreated group. Some of the significant differentially expressed genes (e.g., Dex vs. Untreated: *DUSP1, KLF15* q-values <1.0E-12) have been reported in previous asthma pharmacogenomics studies, while others are novel (e.g., Dex vs. Untreated: *C7, CCDC69,* and *SPARCL1* q-values <1.0E-12). Our results provide a transcriptomic snapshot of the effects of the most common asthma medications in HASM cells and have the potential to improve our understanding of asthma pharmacogenomics.

Drug Metabolism and Drug Interactions

Joseph Kitzmiller

ABSTRACT

Background: Stains are the most-commonly prescribed pharmacotherapy for prevention of atherosclerotic cardiovascular disease. Metabolism of atorvastatin, simvastatin, and lovastatin involves the CYP3A metabolizing enzymes, and *CYP3A4*22* significantly influences the dose needed for achieving optimal lipid control with those stains. Recent reports have described a CYP3A4/5 combined genotype demonstrating significant influence on the pharmacokinetics of various CYP3A substrates.

AIMS

We intend to describe the characteristics of a study population grouped by the CYP3A4/5 combined genotype, to compare the combined CYP3A4/5 analysis to a single-gene CYP3Af analysis, and to evaluate whether the combined genotype analysis approach holds potential utility for guiding statin-dose selection.

METHODS

Two hundred thirty-five patients receiving stable doses of atorvastatin, simvastatin or lovastatin for optimal lipid control were genotyped and grouped according to the CYP3A4/5 combined genotype. Combined-genotype group characteristics were determined and evaluated using standards statistical methods. Results of the CYP3A4/5 combined-genotype analyses were compared with those of the single-gene CYP3A4-analysis in order to determine whether the combined approach better predicts statin-dose requirements

RESULTS AND CONCLUSIONS

The number and demographic composition of patients categorized into CYP3A4/5 combined-genotype groups were consistent with those reported for other cohorts. Although statin-dose requirement was significantly associated with the order CYP3A4/5 combined-genotype grouping, this combined consideration of CYP3A4 and CYP3A5 did not improve the genotype-dosage correlation – suggesting that CYP3A4*22 was the primary variable in this cohort. A larger cohort may be necessary to fully elucidate whether a CYP3A4/5 combined-genotype approach holds promise for better predicting statin-dose requirements compared to CYP3A*22 alone.