Preliminary Data and Work Progress

We have made substantial progress in the aims of this Research Project, including software and statistical methods development for eQTL mapping, toxicogenetic modeling in mouse hepatocytes, and toxicity phenotyping in human cells. The FastMap eQTL analysis software forms the major software development activity, and is being used to support all activities of this Research Project. We envision FastMap as the platform of choice upon which to build additional future eQTL analysis capabilities. We have also made substantial progress in enabling p-value based inference in eQTL settings that does not require permutation. This work requires careful handling of the correlation structure among transcripts using matrix decomposition, and may enable very fast and valid inference using a quick "snapshot" of the dataset. More sophisticated and CPU-intensive analyses of the same data can then be performed using FastMap. We have also developed a new approach to estimating permutation p-values that is essentially time-constant, and does not increase with decreasing p-values. Our continued work on analyzing transcription factor binding will help integrate data from eQTL studies with data about transcription regulation, eventually providing a much more complete view of transcriptional regulation in toxicity response. Finally, we have finished collection of phenotyping (cell viability and apoptosis) data on 14 EPA-relevant chemicals in 87 HapMap human lymphoblastoid cell lines and are close to establishing appropriate cell culture conditions for murine hepatocytes from a large panel of inbred strains.

Results to Date

The goals of this project are to (i) develop toxicogenetic expression quantitative trait loci (eQTL) mapping tools, perform transcription factor network inference and integrative pathway assessment; (ii) perform toxicogenetic modeling of liver toxicity in cultured mouse hepatocytes; (iii) discover chemical-induced regulatory networks using population-based toxicity phenotyping in human cells. We have made substantial progress in all of these goals.

The proposed fast-SNP correlation method has been published (Gatti et al., 2009). This tool, called FastMap, currently works for 2-genotype populations (and therefore crosses between inbred rodent strains). The approach increases the speed of eQTL mapping compared to existing methods by several orders of magnitude, enabling the use of permutation-based inference, which can provide superior error control. The initial paper has included three-SNP ANOVA moving window approaches, in addition to standard single SNP analyses. We have already been successfully applying FastMap to mouse toxicogenomic datasets. We have also made considerable progress toward a human version of FastMap, i.e., a version that can analyze three-SNP genotypes. The new version of FastMap will be applicable to our work on toxicity profiling of HapMap cell lines.

Additional eQTL method development includes approaches to handling stratification and outliers, as well as finding proper detection thresholds for so-called eQTL "hotspots." The overarching goal is to obtain accurate and interpretable p-values for transcriptome maps of SNP x transcript correlations, without the need for permutation. The identification of eQTL hotspots is complicated by the presence of strong correlation between transcripts, which can produce seemingly striking results by chance alone. Therefore it is important to model the correlation structure carefully in order to obtain an understanding of the thresholds necessary to apply for eQTL hotspots. Briefly, for each SNP we consider the sum of genotype-transcript correlation statistics across the transcripts, and approximate the null distribution of this sum as a mixture of independent chi-square densities. Then we consider the distribution of the maximum of this modeled sum across the genome, using simple Gaussian process approximations to model the SNP correlation structure. Together the two approximations provide a reasonably accurate threshold for hotspot detection, greatly enhancing the ability to quickly examine datasets for eQTL evidence.

We have also been working on multiple testing correction approaches, which are among the most important statistical problems in eQTL analysis. We have proposed a geometric interpretation of permutation p-values, and have developed an efficient permutation p-value estimation method based on this geometric interpretation (Sun and Wright, submitted). Application to eQTL studies has demonstrated the efficacy of our method.

A major emphasis of our work has been on the integrated study of eQTL and transcription regulation. In addition to genetic variation, gene expression is regulated by many other factors, such as nucleosome occupation and transcription factor binding. Incorporation of these factors would lead to better understanding of the genetic basis of gene expression, which is an important issue in toxicogenomics. However, efficient and accurate quantification of nucleosome occupation and transcription factor binding are not trivial. We have developed two methods for these purposes (Sun et al. 2009a, Sun et al. 2009b), respectively.

We are also working to establish a toxicogenetic model of liver toxicity in cultured mouse hepatocytes cultures. This year we have focused our efforts on standardization of cell isolation and culture conditions, and determination of whether the near-physiological maintenance of the cells isolated from different mouse inbred strains can be achieved and assess whether the reproducibility of functionality can be attained within a given strain over subsequent isolations. Hepatocytes were isolated from 15 strains of mice and cultured for up to 7 days in traditional 2D culture and cell viability and functionality were assessed. Our data shows that high yield (48-87 million hepatocytes/mouse) and viability (86-98%) can be achieved across a panel of strains. We observed that cell function of hepatocytes isolated from different strains and cultured under standardized conditions is comparable and cells remain viable and metabolically active. These experiments open new opportunities for high-throughput and low-cost *in vitro* assays that may be used for studies of toxicity in a genetically diverse population.

In addition, we aim to extend the application of toxico-genetic studies to investigative toxicology by assessing inter-individual variability and heritability of chemical-induced toxicity phenotypes in cell lines from the Centre d'Etude du Polymorphisme Humain (CEPH) trios assembled by the HapMap Consortium. We have completed an initial screen whereby we treated 87 cell lines from the CEPH trios with 14 environmental chemicals. We applied each chemical in three doses and measured cell viability and apoptosis 24 hours after treatment. Our data demonstrates that variability of response across the chemicals exists for some, but not all agents, with perfluorooctanoic acid and phenobarbital exhibiting the greatest degree of inter-individual variability. At the same time, an appreciable degree of inter-individual variability in susceptibility of cell lines to the chemicals was also observed. While our preliminary assessment of the data shows no significant heritability of toxicity response phenotypes across these cell lines, genetic factors controlling wide variability in response to some agents needs to be addressed. The approach of screening chemicals for toxicity in a genetically-defined, yet variable *in vitro* system, is potentially useful for identification of both agents and individuals that may be at highest risk.

Finally, we work closely with Project 3 on predictive QSAR modeling of the toxicity data, including ToxCast Phase I data. Progress towards the goals of Project 3 is detailed below.

Activities for Subsequent Reporting Period

In Year 2 we will add the described extensions to FastMap, and apply the software to the human lymphocyte toxicity profiling studies. We expect our work on fast eQTL p-value inference to mature rapidly, and be applied to several toxicity eQTL datasets in Year 2. We will investigate the possibility of incorporating our geometric p-value approach into FastMap, although there are substantial challenges in doing so. For FastMap we are already devising approaches to reduce permutation computation by focusing mainly on those genotype-transcript correlations that are likely to be significant, with fewer permutations expended on those correlations that are not significant. For our fast p-value based inference approaches, we are currently handling

challenges posed in analytic approximations to sums of chisquare random variables. We expect to resolve these challenges in Year 2 and have a workable analysis procedure.

For our integrated genomic approaches, we will incorporate the eQTL data, nucleosome occupancy, and transcription regulation information, possibly including microRNA expression to construct larger transcription regulation networks in the Bayesian network framework. Full characterization of a huge Bayesian network using -omics data may be difficult. We will attack this problem by a bottom-up strategy, starting with small networks with only three nodes. Such small networks are easier to construct by the likelihood-based method Dr. Sun developed in previous work. Larger networks can be constructed based on those small networks. We would be interested in the hub genes in the constructed network, especially in their response to toxic exposures.

For our *in vitro* toxico-genomic experiments we will complete characterization of the mouse hepatocyte cultures and perform experiments with several key toxicants for which multi-strain panel *in vivo* data is available: trichloroethylene, WY-14,643, acetaminophen and ethyl alcohol. For experiments with human lymphoblasts, we will complete GWAS analyses of the cell viability and apoptosis data and will correlate the toxicity endpoints with basal gene expression profiles collected from these cell lines.

Finally, we will continue collaborations with Project 1 on the interpretation of the toxicant-perturbed networks from ToxCast Phase I data and other toxicity datasets; and with Project 3 on the QSAR-based analysis of the ToxCast Phase I data.

Publications Arising From this Project

Papers:

1. Gatti DM, Shabalin AA, Lam TC, Wright FA, Rusyn I, and Nobel A. (2009) FastMap: Fast eQTL mapping in homozygous populations. Bioinformatics 4: 482-489.
2. Sun W, Xie W, Xu F, Grunstein M, and Li KC. (2009) Pattern recognition in tiling array ChIP-chip data by segmental semi-Markov model, with application in nucleosome free region study. PLoS ONE, 4(3) e4721.
3. Zhu H, Ye L, Richard A, Golbraikh A, Wright FA, Rusyn I, and Tropsha A. (2009) A novel two-step hierarchical quantitative structure activity relationship modeling workflow for predicting acute toxicity of chemicals in rodents. Envr Health Persp (In Press).
4. Gatti, DM, Harrill, AH, Wright FA, Threadgill DW, and Rusyn I. (2009) Replication and Narrowing of Gene Expression Quantitative Trait Loci using Inbred Mice. Mammalian Genome (*In revision*).
5. Harrill AH, Gatti DM, Threadgill DW, and Rusyn I. (2009) Population-Based Discovery of Toxicogenomics Biomarkers for Hepatotoxicity Using a Laboratory Strain Diversity Panel. Toxicological Sciences (*In revision*).
6. Sun W, and Wright FA (2009) A geometric interpretation of the permutation p-value and its application in eQTL studies. Annals of Applied Statistics (Submitted).
7. Sun W, Buck MJ, Patel M, and Davis IJ (2009) Improved ChIP-chip analysis by mixture model approach. BMC Bioinformatics (*Submitted*).

Posters/Abstracts

1. Li Z, Rusyn I, and Wright FA. (2009) Dose-response pathway analysis for gene expression microarrays. National Academy of Sciences Symposium on Toxicity Pathway- Based Risk Assessment, May 11-13, 2009.
2. Wright FA, Li Z, Huang H, Ghosh A, Sun W, Zou F, and Rusyn I (2009). Prediction of in vivo toxicity endpoints from ToxCast Phase I data using a variety of machine learning approaches. EPA ToxCast Data Analysis Summit, May 14-15, 2009 (abstract and oral presentation).

3. Martinez S, Bradford B, Kaiser R, Soldatow V, Amaral K, Ferguson S, Black C, LecLuyse E, and Rusyn I. (2009) Development of *in vitro* toxicogenetic models for hepatotoxicity. EPA ToxCast Data Analysis Summit, May 14-15, 2009.
4. Hatcher S, Kosyk O, Ross P, Wright F, Schwartz J, Dix D, and Rusyn I. (2009) *In vitro* screening for chemical toxicity in a genetically-diverse human model system. EPA ToxCast Data Analysis Summit, May 14-15, 2009.
5. Zhu H, Sedykh A, Zhang L, Rusyn I, and Tropsha A. (2009). Using ToxCast cell-viability and gene-expression assays as biological descriptors in QSAR modeling of animal toxicity endpoints. EPA ToxCast Data Analysis Summit, Durham, USA.
6. Martinez S, Bradford B, Kaiser R, Soldatow V, Amaral K, Ferguson S, Black C, LecLuyse E, and Rusyn I. (2009) Development of the *in vitro* toxicogenetic models for hepatotoxicity. Society of Toxicology Annual Meeting, Baltimore, MD.
7. Romanov S, Gatti D, Tsuchiya M, Zeng M, Medvedev A, and Rusyn I. (2009) Profiling of multiple transcription factors activity in mouse liver. Society of Toxicology Annual Meeting, Baltimore, MD.
8. Zhao N, Gatti DM, and Rusyn I. (2009) Genome-level analysis of genetic regulation of sex-specific gene expression in mouse liver. Society of Toxicology Annual Meeting, Baltimore, MD.
9. Zhang L, Zhu H, Rusyn I, Judson R, Dix D, Houck K, Martin M, Richard A, Kavlock R, and Tropsha A. (2009) Cheminformatics Analysis of EPA ToxCast chemical libraries to identify domains of applicability for predictive toxicity models and prioritize compounds for toxicity testing. Society of Toxicology Annual Meeting, Baltimore, MD.