Carolina Center for Computational Toxicology

Funded by U.S. EPA Cooperative Agreement STAR RD 833825

Progress Report for Project Period 04/01/09 - 03/31/10 (Year 2)

Report Date: 04/21/10

Ivan Rusyn, Principal Investigator

Project 2:
Toxico-Genetic Modeling: Population-Wide Predictions from Toxicity Profiling

Preliminary Data and Work Progress

We have made substantial progress in the aims of this Research Project, including toxicogenetic modeling in population-based in vitro models (e.g., mouse hepatocytes, and human lymphoblastoid cells), and software and statistical methods development for eQTL mapping. The FastMap eQTL analysis software forms the major software development activity, and is being used to support all activities of this Research Project. FastMap 2.0, with ability to analyze three-genotype (human) data, is in beta form and already proving useful for this project. We have also substantially improved our methods for conducting p-value based inference in eQTL settings without need for permutation. This approach has strong connections to related work by our group in which permutation-based p-values are expressed in a geometric sense. We have also completed the analyses of our *in vitro* population-based toxicity phenotyping data (cell viability and apoptosis) on 14 EPA-relevant chemicals (pesticide actives, plasticizers, polychlorinated biphenyls, etc.) in a panel of densely genotyped 87 HapMap CEU human lymphoblastoid cell lines. We have established the degree of inter-individual variability in responses to these agents, examined whether such variability is heritable, and conducted genome-wide association analysis of the individual chemical-assay phenotypes. In addition, we successfully completed data collection on 240 chemicals (10 concentrations) in two assays in 80+ HapMap cell lines through an ambitious collaboration with the Tox21 partner organizations, the National Toxicology Program and the NIH Chemical Genomics Center. Additional progress is being made in our experiments that aim at establishing an *in vitro* primary hepatocyte population-based model using a panel of inbred mouse strains. We have completed experiments aimed at establishing appropriate cell culture conditions for murine hepatocytes from a large panel of inbred strains. Furthermore, we have tested the reproducibility of the biochemical and molecular function of the cultured cells isolated from several strains, as well as experiments that tested the dose-response to 3 model toxicants in cells from various 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.

In order to perform fast eQTL mapping in human cells, we have largely finished the extension of our fast SNP-correlation method FastMap to handle the three-genotype setting. We expect that the extension, which still operates using tree-based methods, will be very widely used, and we are gearing up to apply the FastMap 2.0 to our toxicity profiling of HapMap cell lines. Additional extensions of FastMap under development include bootstrapping approaches, which, in contrast to the permutation analyses we currently perform, can allow the simultaneous discovery of SNP-transcript correlations as well as correlations between SNP and outcomes such as toxicity susceptibility. The bootstrap approach, as with permutations, can re-use the SNP tree, and thus is extremely fast.

In addition to computational advances in eQTL mapping, our approaches to handling stratification and outliers for eQTL analysis have been refined considerably. Through extensive analysis of HapMap cell lines, we have explored the limits and necessity of performing population stratification control in eQTL studies. In addition, we have developed a further extension to our method for finding a proper detection threshold for so-called eQTL “hotspots” (transbands). Our original approach uses a decomposition of the transcript-transcript correlation matrix in order to save computation vs. using permutation. However, we have recently shown that essentially the same result can be achieved by performing a decomposition of the covariance matrix of individual expression *samples*, which is of much lower dimension, for a further reduction in computation burden. Another example of an improvement which provides computational savings is our approach to view permutation p-values in a geometric manner has been published. We have already applied this to eQTL data, and are eager to expand the approach further.

The large-scale testing inherent in eQTL analysis has limitations in the biological interpretability of significant associations. At the other extreme, existing specific pathways and signaling networks can be investigated in a systematic and predictive manner using mathematical approaches employed in Project 1. However, there is a great need for methods that address the intermediate investigations of ensembles of genes that jointly affect toxicity susceptibility. Accordingly, we are developing methods to fill in the gap using Bayesian iterative adaptive Lasso regression approaches for eQTL analysis, to further refine lists of candidates and to gain potential power from groups of weakly correlated transcripts. These approaches, based on recent general work by Dr. Sun and colleagues, are being adapted for eQTL analysis with toxicity applications for the Center.

Although much of our work has followed non-biological testing approaches for eQTL analysis, any serious follow-up of eQTLs must also consider that gene expression is regulated by many factors other than genetic variation. Accordingly, we have continued to refine our methodology for estimating nucleosome occupation (Sun et al. 2010).

We continue to work closely with Project 3 on predictive QSAR modeling of the toxicity data, including ToxCast Phase I data. Project 2 investigators were actively involved in the ToxCast Data Analysis Summit in May 2009, and a number of manuscripts with Project 3 investigators are being prepared. Progress towards the goals of Project 3 is detailed below.

We have also completed the analyses of our *in vitro* population-based toxicity phenotyping data (cell viability and apoptosis) on 14 EPA-relevant chemicals (pesticide actives, plasticizers, polychlorinated biphenyls, etc.) in a panel of densely genotyped 87 HapMap CEU human lymphoblastoid cell lines. We have established the degree of inter-individual variability in responses to these agents, examined whether such variability is heritable, and conducted genome-wide association analysis of the individual chemical-assay phenotypes. This work serves as an important proof of principle in establishing the utility of *in vitro* toxicity screening in a genetically-defined population model, as well serving as a testing ground for the combination of toxicity phenotypes with baseline expression and genotype data.

Our initial success with these novel experimental approaches has led to a much more ambitious collaboration with the Tox21 partner organizations, the National Toxicology Program and the NIH Chemical Genomics Center. First, we are working with NTP researchers on the analysis of the data from toxicity screening of 1,408 chemicals (assayed for 2 endpoints in 12+ concentrations) in lymphoblasts from 20 human twin pairs. Second, in collaboration with NCGC and NTP, we have just completed the data collection phase on a very ambitious experiment which assayed HapMap CEU human lymphoblastoid cell lines from 27 parent-child trios for 2 phenotypes (cell viability and apoptosis) with 240 chemicals in 12 doses. This dataset will serve as a major nexus of the computational analyses in Year 3 of the project.

Additional progress is being made in our experiments that aim at establishing an *in vitro* primary hepatocyte population-based model using a panel of inbred mouse strains. We have completed experiments aimed at establishing appropriate cell culture conditions for murine hepatocytes from a large panel of inbred strains. Furthermore, we have tested the reproducibility of the biochemical and molecular function of the cultured cells isolated from several strains, as well as experiments that tested the dose-response to 3 model toxicants in cells from various strains. The data from this *in vitro* experiment is being analyzed in relationship to the data on inter-individual variability of response to these agents from both *in vivo* mouse experiments, and *in vitro* human lymphoblast experiments detailed above.

Activities for Subsequent Reporting Period

In Year 3 we will finish the rollout of HapMap 2.0 and apply it to the toxicity profiling studies described, including the NCGC 1408 chemical screening collaboration for HapMap cell lines. In addition, this tool will be applicable to additional toxicological datasets that have been collected with independent NIH funding. The need for fast computation is truly becoming critical, as the large scale of the studies underway creates challenges for both bench researchers and computationally sophisticated users. In addition to extensions to the approaches described in the “Results to Date” section, we have the development of several novel methods underway. One such method is a new approach to trans-band eQTL analysis in which the apparent effect of a SNP on numerous transcripts is decomposed into a number of principal components (PCs). The value of this approach is to discover whether a single highly correlated set of transcripts (a single PC) is under the influence of the SNP, or whether such transcripts are only weakly correlated. If the transcripts are highly correlated, a single expression module may be implicated. If the transcripts are weakly correlated, the transband may be consistent with a broader influence, such as might be seen for a transcription factor. Although carefully elucidating the underlying biology will be challenging, the ability to quickly distinguish between trans-bands of very different types will be an important advance.

In Year 3 we will also extend our Bayesian Iterative Adaptive Lasso procedures to handle the current computational challenges. Markov Chain Monte Carlo methods can be used, but are highly computationally intensive. We are exploring an expectation-conditional maximization approach to find the posterior mode, in order to provide approximations to the posterior distributions of eQTL effects and hyper-parameters. In our formulation, gene expression of a limited number of transcripts is the response and genotypes of a limited (though still large) number of SNPs are used to predict response, with fixed effects for experimental condition (for example toxicity condition or other covariates) included. Substantial challenges remain in applying such data to real toxicity eQTL datasets, even after the initial discovery phase employed by HapMap. In addition, there are difficulties posed by the so-called “winner’s curse” of over-estimation of the magnitude of effect sizes if the same data are used to discover the vetted list of transcripts and SNPs. We will explore the possibility of a unified discovery and analysis approach in Year 3 that will overcome this difficulty.

With regards to the experimental aims of this project, we will be focusing on both our work with human lymphoblastoid cell lines and primary mouse hepatocytes. In case of the former, we will be analyzing the data from the experiment performed in partnership with Tox21. Specifically, we will perform both traditional analyses of variability and heritability, and a genome-wide association study to not only characterize, but also mechanistically interpret the inter-individual differences in responses to a large number of chemicals across the panel of cells. With regards to the work with primary mouse hepatocytes isolated from a panel of mouse inbred strains, we will be aiming to screen a set of 15-30 model chemicals in both 2D culture conditions, and on 3D mouse liver bioreactors available in Dr. Rusyn's lab through a collaboration with MIT on an NIEHS-funded project.

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. Ross, P.K., Woods, C.G., Bradford, B.U., Kosyk, O., Gatti, D.M., Cunningham, M.L., and Rusyn, I. (2009) Time-course comparison of xenobiotic activators of CAR and PPARalpha in mouse liver. Toxicol Appl Pharmacol 235:199-207.
2. Shabalin, A.A., Weigman, V.J., Perou, C.M., and Nobel, A.B. (2009) Finding Large Average Submatrices in High Dimensional Data. Annals of Applied Statistics 3:985-1012.
3. Zhu, H., Ye, L., Richard, A., Golbraikh, A., Wright, F.A., 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 117:1257-1264 (*from in press to published*).
4. Gatti, D.M., Harrill, A.H., Wright, F.A., Threadgill, D.W., and Rusyn, I. (2009) Replication and Narrowing of Gene Expression Quantitative Trait Loci using Inbred Mice. Mammalian Genome 20:437-446 (*from submitted to published*).
5. Harrill, A.H., Gatti, D.M., Threadgill, D.W., and Rusyn, I. (2009) Population-Based Discovery of Toxicogenomics Biomarkers for Hepatotoxicity Using a Laboratory Strain Diversity Panel. Toxicological Sciences 110:235-243 (*from submitted to published*).
6. Sun, W., and Wright, F.A. (2010) A geometric interpretation of the permutation p-value and its application in eQTL studies. Annals of Applied Statistics (*from submitted to in press*).
7. Sun, W., Buck, M.J., Patel, M., and Davis, I.J. (2010) Improved ChIP-chip analysis by mixture model approach. BMC Bioinformatics 10:173. (*from submitted to published*).
8. Rodgers, A.D., Zhu, H., Fourches, D., Rusyn, I., and Tropsha, A. (2010) Modeling Liver-Related Adverse Effects of Drugs Using kNN QSAR Method. Chem Res Toxicol 23:724-732 (*from submitted to published*).
9. Adams, T.M., and Nobel, A.B. (2010) Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. Annals of Probability (*in press*).
10. Gatti, D., Barry, W.T., Nobel, A., Rusyn, I., and Wright, F.A. (*submitted*) Heading Down the Wrong Pathway. (*submitted to BMC Genomics*).

Posters/Abstracts

1. Li, Z., Rusyn, I., and Wright, F.A. (2010) Dose-response pathway analysis for gene expression microarrays. National Society of Toxicology Annual Meeting, March 7-11, 2010.

2. O'Shea, S. H., Schwarz, J., Kosyk, O., Ross, P.K., Wright, F.A., Tice, R.R., Dix, D.J., and Rusyn, I. (2010) *In vitro* screening for population variability in chemical toxicity. Society of Toxicology Annual Meeting, Salt Lake City, UT.
3. Sedykh, A., Zhu, H., Tang, H., Zhang, L., Richard, A., Rusyn, I., and Tropsha, A. (2010) Using *in vitro* Dose-Response Profiles to Enhance QSAR Modeling of *in vivo* Toxicity. Society of Toxicology Annual Meeting, Salt Lake City, UT.
4. Zhu, H., Sedykh, A., Wright, F., Rusyn, I., Tropsha, A. (2009) Using quantitative High Throughput Screening (q-HTS) results as biological descriptors to assist modeling of acute rat toxicity, American Chemical Society 238th National Meeting, Washington, DC, August, 2009