Carolina Center for Computational Toxicology

Funded by U.S. EPA Cooperative Agreement STAR RD 833825
Progress Report for Project Period 04/01/11 - 03/31/12
Report Date: 02/19/12
Ivan Rusyn, Principal Investigator

The Carolina Center for Computational Toxicology (CCCT) is comprised of three research projects and an administrative core. The major aims and objectives of the CCCT have not changed from the original application. The content of this progress report is organized according to U.S. EPA guidelines and it summarizes significant activities and accomplishments of all four components of the CCCT.

Preliminary Data and Work Progress

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

We have continued to make substantial progress in the aims of this Project, including toxicogenetic modeling in population-based *in vitro* models (e.g., human lymphoblastoid cells), and developing software and statistical methods for eQTL mapping. The FastMap and Matrix eQTL analysis software formed the major software development activity, and supports much of the activity of this Project. FastMap 2.0, with ability to analyze two-genotype (e.g., mouse inbred) data, as well as three-genotype (e.g., human) data, is now complete and in use by our group when graphical exploration of the data are important. Innovations in this tool include (i) the use of adaptive permutations; (ii) multiple subset summation trees; (iii) automated reading of HapMap and PLINK genotype formats; (iv) association mapping using linear models or one-way ANOVA; and (v) generation of the output that is both graphical and numerical and is also suitable for plotting using LocusZoom tool. The Matrix eQTL package, written in the R language, is intended for more advanced users, and is the fastest eQTL package available, with separate handling of *cis*- vs. *trans*- eQTLs, and automatic control of false discovery rates.

Our *in vitro* population-based toxicity phenotyping data (cell viability and activation of caspase-3/7 endpoints) has been published on 240 EPA-relevant chemicals in a panel of densely genotyped 87 HapMap CEU human lymphoblastoid cell lines. This qHTS screening was collaboration with the Tox21 partner organizations, the National Toxicology Program and the NIH Chemical Genomics Center. A similar, but considerably more ambitious experiment to perform qHTS screening of 180- chemicals at 8 concentrations for cell viability endpoint assay in 1100+ lymphoblastoid cell lines from the 1000 Genomes project was conducted in June-September 2011. The data collection phase is completed and data analysis is underway. This new data will enable more powerful mapping of susceptibility genes to toxicity response, as well as much more sensitive investigation of the effects of rarer variants.

Results to Date

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

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 continue to make substantial progress in all of these goals.

Our fast SNP-correlation method, FastMap 2.0, is now operable for human (i.e. three-genotype) data, and is useful when graphical exploration of the data is important. In addition to speedups implemented for FastMap, our new package Matrix eQTL is a command-line tool written in R, and is extremely fast. Thus Matrix eQTL is well-positioned to be the platform of choice for linear modeling and testing with eQTL data in the future. In addition, Drs. Rusyn, Wright, and Nobel are co-PIs of a methods grant to accompany the GTEx initiative (http://commonfund.nih.gov/GTEx/), which will provide RNA and DNA profiles of ~160 donors in multiple tissues, and these data are serving as an excellent testing ground for matrix eQTL and related software. Dr. Wright also continues to lead the statistical component of a large eQTL project, mapping lymphocyte eQTL variation using ~1300 twins, which is providing important information about eQTL heritability, which in turn provides directly useful information for prioritizing transcripts for potential genetic variation in susceptibility. In a partnership with multiple investigators, we have also developed the seeQTL eQTL browser (Xia et al., 2012) that has more extensive search and display capabilities than competing browsers.

Additional methods work has included fast approximate approaches to eQTL permutation testing (Sun and Wright, 2010), and analysis of sequence-based expression (RNA-Seq) data. For RNA-Seq analysis, we have developed the new R package BBSeq (Zhou et al., 2012), and described a framework for handling RNA-Seq data in eQTL analysis (Sun, 2011).

Another source of computational difficulty arises in performing genetic pathway analysis. We and others have repeatedly shown that methods in which individual expression profiles are permuted relative to clinical phenotypes, as is performed in our SAFE R package, are superior to competing methods. However, the approaches can be time-consuming and memory-intensive. We have recently developed the safeExpress procedure (Zhou et al., submitted) to overcome this obstacle, using mathematical approximations to accurately mimic permutation testing without requiring actual permutation.

We also continue to work closely with Project 3 on predictive QSAR modeling of the toxicity data, including Tox21 *in vitro* qHTS data. Project 2 investigators work closely with Project 3 investigators that resulted in several poster presentations and publications. Progress towards the goals of Project 3 is detailed below.

Our ambitious collaboration with the Tox21 partner organizations, the National Toxicology Program and the NIH Chemical Genomics Center has been completed with a publication to appear in *Toxicological Sciences*. In collaboration with NCGC and NTP, we have completed screening of 240 chemicals (in 12 concentrations) in 81 HapMap CEU human lymphoblastoid cell lines for 2 phenotypes (cell viability and apoptosis). This dataset remained a major nexus of the computational analyses in Year 4 of the project. Specifically, the quantitative high-throughput screening in a population-based human in vitro model system has several unique aspects that are of utility for toxicity testing, chemical prioritization, and high-throughput risk assessment. First, standardized and high-quality concentration-response profiling, with reproducibility confirmed by comparison with previous experiments, enables prioritization of chemicals for variability in inter-individual range in cytotoxicity. Second, genome-wide association analysis of cytotoxicity phenotypes allows exploration of the potential genetic determinants of inter-individual variability in toxicity. Furthermore, highly significant associations identified through the analysis of population-level correlations between basal gene expression variability and chemical-induced toxicity suggest plausible mode of action hypotheses for follow up analyses. We conclude that as the improved resolution of genetic profiling can now be matched with high-quality in vitro screening data, the evaluation of the toxicity pathways and the effects of genetic diversity are now feasible through the use of human lymphoblast cell lines.

A further extension of these in vitro studies using a genetically-diverse and –defined in vitro population-based qHTS screening approach is underway. In partnership with NTP and NCGC we have collected data on 1000+ lymphoblast cell lines from HapMap and 1000 Genomes projects. These cell lines are from 9 populations representing 5 continents: Utah

residents with Northern & Western European ancestry; Han Chinese in Beijing, China; Japanese in Tokyo, Japan; Luhya in Webuye, Kenya; Mexican ancestry in Los Angeles, CA; Tuscan in Italy; Yoruban in Ibadan, Nigeria; British from England and Scotland; and Columbian in Medellin, Colombia. Of these, 1095 cell lines (99.2%) were screened in quantitative high-throughput screening format with 180 chemical substances at 8 concentrations (0.3 nM-92 μM) in a cell viability (CellTiter-Glo®) assay. Duplicate 1536-well plates were screened for ~70% of the cell lines, revealing excellent intra- and inter-experimental reproducibility in both concentration-response curve class and relative cell viability (r=0.977 for EC_{50} values). Among the 180 substances screened, the variability in cytotoxicity induced by individual compounds (from 20% to 68%) across the different human cell lines demonstrates the presence of considerable inter-individual variability in sensitivity to chemicals. As part of this project, we are also exploring robust EC_{10} estimation procedures, in order to best describe variation in individual susceptibility (Lock et al., 2012). The 1000 Genomes-based *in vitro* screening model offers exceptional opportunities for identifying variations in response at the DNA sequence level, filling gaps in high-throughput risk assessments by establishing population-based confidence intervals in toxicity, and probing candidate susceptibility pathways by exploring the correlations with mRNA levels.

Activities for Subsequent Reporting Period

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

In the next year we will continue to refine our software in order to meet the analysis needs for the extensive profiling data developed in Project 2. FastMap 2.0, Matrix eQTL and additional software developed by our group are relatively mature and are applied to the toxicity profiling studies. We expect that our safeExpress fast approximations to pathway significance testing will be extremely useful for future dose-response expression studies.

We are also moving forward quickly in methods for next-generation sequence analysis. This work is especially relevant for 1000 Genomes LCL project. Incorporating published expression data into this project will be extremely important, but creates new analysis and computational challenges. In this year we expect to perform extensive integrative analyses, in which the major genotype, sequence, and RNA expression data are synthesized and combined to better understand variation in cytotoxicity variability. Accordingly, for the experimental aims of this project, we will be focusing on data analysis from the 1000 genomes qHTS screening experiment. 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. This work will also be combined with results from the twin-study heritability project, to refine the plausible transcriptional candidates underlying apparent variation in susceptibility.

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

Publications Arising From this Project in Year 4

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

Papers:

1. Gatti, D.M., Lu, L., Williams, R.W., Sun, W., Wright, F.A., Threadgill, D.W., and Rusyn, I. (2011) MicroRNA expression in the livers of inbred mice. Mutat Res 714:126-133 (*from in press to published*).

2. Low Y., Uehara T., Minowa Y., Yamada H., Ohno Y., Urushidani T., Sedykh A., Muratov E., Fourches D., Zhu H., Rusyn I., Tropsha A. (2011) Predicting Drug-induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. Chem. Res. Tox. 24:1251-1262 (*from submitted to published*).
3. Zhou, Y., Xia, K., and Wright, F.A. (2011) A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics 27:2672-2678 (*from submitted to published*).
4. Lee S, Wright FA, Zou F. (2011) Control of population stratification by correlation-selected principal components. Biometrics. 67:967-974.
5. Wright, F.A., Shabalin, A., and Rusyn, I. (2012) Computational tools for discovery and interpretation of expression quantitative trait loci. Pharmacogenomics 13:343-352, 2012.
6. Lock, E.F., Abdo, N. Huang, R., Xia, M., Kosyk, O., O'Shea, S.H., Zhou, Y.H., Sedykh, A., Tropsha, A., Austin, C.P., Tice, R.R., Wright, F.A., and Rusyn, I. (2012) Quantitative high-throughput screening for chemical toxicity in a population-based *in vitro* model. Toxicol Sci (in press).
7. Sun, W. A. (2012) Statistical Framework for eQTL Mapping Using RNA-seq Data. Biometrics (in press).
8. Rusyn, I., Sedykh, A., Low, Y., Guyton, K.Z., Tropsha, A. (2012) Predictive modeling of chemical hazard by integrating numerical descriptors of chemical structures and short-term toxicity assay data. Toxicol. Sci. (*in press*)
9. Xia, K., Shabalin, A.A., Huang, S., Madar, V., Zhou, Y.H., Wang, W., Zou, F., Sun, W., Sullivan, P.F., Wright, F.A. (2012) seeQTL: a searchable database for human eQTLs. Bioinformatics. 28:451-452.
10. Zhou, Y., Barry, W.T., and Wright, F.A. (2012) Empirical pathway analysis, without permutation (submitted).

Posters/Abstracts:

1. Abdo, N., Xia, M., Kosyk, O., Huang, R., Sakamuru, S., Austin, C., Tice, R., Wright, F., and Rusyn, I. The 1000 genomes toxicity screening project: Utilizing the power of human genome variation for population-scale *in vitro* testing. Society of Toxicology Annual Meeting, San Francisco, CA. 2012.
2. Low, Y., Fourches, D., Sedykh, A., Rusyn, I., and Tropsha, A. Multi-space *k*-Nearest Neighbors as a Novel Hybrid Approach Integrating Chemical and Toxicogenomic Descriptors for Improved Toxicity Prediction. Society of Toxicology Annual Meeting, San Francisco, CA. 2012.
3. Sedykh, A., Low, Y., Lock, E., Rusyn, I., and Tropsha, A. Using population-based dose-response cytotoxicity data in computational modeling of *in vivo* rat acute toxicity. Society of Toxicology Annual Meeting, San Francisco, CA. 2012.
4. Wignall, J., Sedykh, A., Tropsha, A., Woodruff, T., Zeise, L., Rusyn, I., Cogliano, V., Chiu, W., and Guyton, K. Modeling toxicity values using chemical structure, *in vitro* screening, and *in vivo* toxicity data. Society of Toxicology Annual Meeting, San Francisco, CA. 2012.