Final Report

Period Covered by the Report: 04/01/2008 - 10/31/2013

Date of Final Report: 10/31/2013

GRANT NUMBER (FAIN): 83382501

Title: Carolina Center for Computational Toxicology

Investigator(s):
Ivan Rusyn, M.D., Ph.D. iir@unc.edu
Shawn Gomez, Ph.D. smgomez@unc.edu
Timothy Elston, Ph.D. telston@amath.unc.edu
Fred Wright, Ph.D. fwright@bios.unc.edu
Andrew Nobel, Ph.D. nobel@email.unc.edu
Wei Sun, Ph.D. wsun@bios.unc.edu
Alexander Tropsha, Ph.D. alex_tropsha@unc.edu

Institution(s): The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

Project Period: 04/01/2008 - 10/31/2013

Project 1
Predictive Modeling of Chemical-Perturbed Regulatory Networks in Systems Toxicology

This project was focused on the creation and development of computational methods and models, with a particular focus on better understanding molecular networks and how these networks respond to chemical perturbations. The results of this work will be used to both define molecular network architectures relevant to toxicology as well as model their dynamical behavior.

Specific Objective 1. Development and application of data-driven methods for the inference and high-level modeling of regulatory network response to chemical perturbation.

Understanding how chemical perturbations may lead to phenotypic changes requires a picture of how the underlying molecular networks are "wired" as well as a means of quantifying the differences in network architectures after this perturbation. Relevant to chemically-induced cancers and developmental disorders, this work has fostered a range of approaches and we have established a variety of methods for the prediction of molecular interaction networks (Choi and Gomez, 2009; Doolittle and Gomez 2010, 2011). As predicting protein interactions is a potentially slightly easier task than inference of chemical-protein interactions, we initially focused on data mining and structural-based approaches for predicting protein network architecture. Analogous to chemical systems, we tested these methods on host-pathogen systems where an "external" molecule exerted its effect on the host (human) molecular network. We have since mapped these approaches to chemical-protein interactions, providing a novel approach for mapping relevant network architectures.

Expanding on a related aspect of this work, the development of novel methods (xCEED) for the comparison of graph/tree structures (Choi and Gomez, 2009) provides a unique way to quantify similarities and/or differences between molecular signaling networks. This family of methods has a wide range of applications, from helping to infer protein interactions to the comparison of pathway/network structures. These comparisons are an important component of future toxicity studies where understanding and comparing the relevance of experimental results across species (e.g. comparison of rat/mouse studies to human data) is essential.

As chemicals can be described structurally in terms of a network, maximum common subgraph algorithms are a mainstay of chemical informatics applications. While the previous xCEED approaches provide some capability in this area, we have since developed a novel graph algorithm approach (in preparation) that provides a powerful new capability to describe and compare network structures, both small-scale chemicals as well as very large-scale molecular networks, and quantify their differences and similarities. These approaches are fundamental in the prediction of where such chemical structures may best "plug in" to the human molecular network within the cell.

The complexity of data sets such as found in ToxCast, require the development of novel approaches for data integration, analysis and interpretation. We have developed dimension reduction and data mining approaches that allow the integration of different data types (assays, chemicals and endpoints) and their downstream interpretation as "functional modules" that define chemical activity networks and provide a novel means of analysis (Choi et al, 2009; Staab et al, 2009; Gomez, 2011). In addition, we have worked with other center members on approaches that utilize known biological pathway information to enhance QSAR predictions (Zhu et al, 2011).

We are working with a number of groups on the collection of data describing metabolic changes in the face of chemical perturbation. Supporting this, we developed a novel method for the alignment of NMR spectra that are typical outputs of systems-level metabolomic analyses (O'Connell et al, 2009; Staab et al, 2010). This approach (termed PCANS – Progressive Consensus Alignment of NMR Spectra) improves the analysis of metabolomic data sets that are relevant in the analysis of toxicity-relevant data sets.

Application of these and related tools to cancer systems having well-defined experimental systems, known chemical interventions, and a wealth of background information is leading to significant insight into how perturbations to cellular molecular networks lead to the presentation of disease as well as potential methods for ameliorating these perturbations through alterations in network wiring (Abell et al. 2011; Duncan et al. 2012).

Specific Objective 2. Mechanistic modeling of nuclear receptors.

The original goal of this aim was to develop mechanistic models describing nuclear receptor function. As originally proposed, the data for this model would be derived from literature data and from associated center members. However, in the course of initial model investigation, we found that the data available was generally inadequate for the level of mechanistic model development we desired. In addition, provided Toxcast generated data could not generally be used to inform such models.

Parallel efforts, however, led to the development of powerful large-scale physiological model of metabolism (Xu al. 2011b). This model describes, at a mechanistic level, metabolism between liver, adipose, muscle and blood compartments. Primary detail is provided at the level of metabolic pathways within the liver. This model has been validated through comparisons to experimental measurements of metabolism in the mouse and rat model animal systems at the cellular and whole-liver levels. Complimentary to PK/PD models, this model provides insight into the dynamics of the metabolic networks at the molecular level; where the action of chemicals of interest exert their effect. In a related capacity, this model helps in the interpretation of fasting effects on gene expression, which is a potentially important confounding variable in toxicity studies. In collaboration with Ingenuity Systems Inc., we established a 10-gene signature that can serve as a diagnostic flag signaling the possible presence of confounding due to fasting effects (Xu et al. 2011a).

Extension of this work has led to the establishment of a model for glutathione metabolism (Jeffries et al. 2012). Glutathione is a key metabolic player in toxicity metabolism, acting as an antioxidant as well as within the detoxification processes. The model we have developed has been supported by experimental data derived from a NMR bioreactor housing human hepatocytes, with metabolite data samples capable of being collected once every minute. Exposing these cells to bromobimane, the model has now undergone significant parameter fitting and validation and provides a power capability for the modeling of a key detoxification pathway that is described within the context of whole human body metabolism. Models of airway epithelia, DNA damage and/or involving complex boundaries have similarly been developed (Herschlag et al. 2013; Strychalski et al. 2010; Garcia et al. 2011; Kesseler et al. 2013).

Together, these models form the foundation for the modeling of key detoxification pathways and their effect on metabolism and physiology. They are now being applied in a recent pilot study in Diversity Outbred (DO) mice where we have collected 50 liver and serum samples from this genetically heterogeneous population and identified metabolites using the Metabolon metabolomics platform. Over 300 metabolites have been identified in both tissue types and these

measurements are now being used in model development as well as a proof-of-concept for understanding metabolic network perturbation in the context of genetic variation.

Specific Objective 3. Integration and deployment of high- (Specific Objective 1) and low- (Specific Objective 2) modeling tools.

As described in the original proposal, a major challenge is in making modeling tools practical for investigators with a range of backgrounds and areas of expertise. To enhance the broad use and effectiveness of the tools developed in this work, we have provided them in an open-source format and provided extensive support for their further use. Longer-term, we are working on translating some of these tools into web applications, which the goal of further enhancing their broader usability.

Our whole-body physiological model of liver/hepatocyte-focused metabolism is being provided as MATLAB code that is distributed, along with the publication, on the *PLoS Computational Biology* website. In addition, the simulation code is available as an SBML (Systems Biology Markup Language) representation to facility use by a wider range of investigators. We have consulted with several groups on its use and implementation. As an example, we have provided this code to Gary Churchill of the Jackson Laboratory and have consulted with him and his group on its use as both a research tool as well as its use in a potential educational application. This is a continuing collaboration and is being further developed in the context of studying genetic variation (utilizing mouse models) and its role in metabolism and linkage to chemical perturbation.

The xCEED tools for network/graph comparison have been distributed both on our website as well as a bitbucket repository (https://bitbucket.org/gomezlab/xceed). In collaboration with downstream users, we have improved aspects of usability and presentation of results since its initial publication and these changes are part of the bitbucket distribution.

The PCANS software for improved analysis of NMR spectra commonly acquired from toxicity studies, is provided on our website. This code and a prototype web application led to feedback from users that led to a new version with an improved ability to pick peaks from spectra. In addition, due to potentially high computational demands, the ability to run on Amazon Web Services has been developed. This new version will be distributed through a public data repository such as *bitbucket* or *github*.

Additional Outcomes.

While not an original objective of this project, we found that a significant limitation in the development of relevant mechanistic toxicity models was the lack of dynamic measurements - i.e. those that explicitly quantify at how these systems change as a function of time (in parallel with factors such as dose). We have developed a set of image analysis tools that that allow the quantification of cell/molecular behaviors captured during live cell microscopy studies (Berginski et al. 2011; Chen et al. 2013; Chaki et al. 2013). Longer-term, these technology platforms provide the ability to perform high-throughput quantification of the effects of chemical exposure on the dynamic behaviors of molecular networks within individual cells and cell populations.

Project 2
Toxico-genetic modeling: Population-wide predictions from toxicity profiling

This project was focused exclusively on the development of tools for the analysis of the toxicological data in genetically diverse subjects by taking into account the information on high-density (up to millions of SNPs) genetic maps of the test populations. The results will be used to provide toxico-genetic context to the development of mechanistic, predictive, and statistical models for computational toxicology (Rusyn and Daston 2010).

Specific Objective 1: Develop toxico-genetic expression quantitative trait loci (eQTL) mapping tools, perform transcription factor network inference and integrative pathway assessment.

1.1. Fast SNP-phenotype correlation methods

Expression quantitative trait locus (eQTL) analysis has moved from a cutting-edge concept in genomics to a mature area of investigation, with important connections to genome-wide association studies for human disease, pharmacogenomics and toxicogenomics. Despite the importance of the topic, many investigators must develop their own code or use tools not specifically suited for eQTL analysis. This project was developing convenient computational tools that can be used for discovery of eQTL, or to interpret genome-wide association study results (Wright et al. 2012).

The fast-SNP correlation method, called FastMap (Gatti et al. 2009b), works for 2-genotype populations (i.e., 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. We have also created a human version of FastMap, i.e., a version that can analyze three-SNP genotypes (available from http://comptox.unc.edu/fastmap.php). 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 2010).

In genome-wide association studies, population stratification is recognized as producing inflated type I error due to the inflation of test statistics. We showed that principal component-based methods applied to genotypes provide information about population structure, and can be used to control for stratification (Zou et al. 2010). To address this challenge with a computational tool, we developed EigenCorr, which selects principal components based on both their eigenvalues and their correlation with the (disease) phenotype. Our approach tends to select fewer principal components for stratification control than does testing of eigenvalues alone, providing substantial computational savings and improvements in power. Analyses of simulated and real data demonstrate the usefulness of the proposed strategy (Lee et al. 2011).

In order to incorporate sequence-based expression (RNA-Seq) data into our analysis pipelines we have been developing more powerful approaches to handle transcript count data. A hallmark of such datasets is that they show strong evidence of overdispersion, and must be modeled using beta-binomial or similar distributions. We have developed a new package, BBSeq (Zhou et al. 2011), which performs this modeling, and also attempts to maximize power for small samples by modeling the relationship between the mean and the overdispersion. We also described a framework for handling RNA-Seq data in eQTL analysis (Sun 2012; Sun and Hu 2013; Sun et al. 2009b).

eQTL heritability 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.

1.2. Integration of eQTL and transcriptional network regulation

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. 2009c).

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 by our SAFE tool (Gatti et al. 2010a; Gatti et al. 2009c), are superior to competing methods. However, the approaches can be time-consuming and memory intensive. We have developed the safeExpress procedure (Zhou et al. 2013) to overcome this obstacle, using mathematical approximations to accurately mimic permutation testing without requiring actual permutation.

Specific Objective 2: Perform toxico-genetic modeling of liver toxicity in cultured mouse hepatocytes.

In the past 5 years, genetically defined mouse models have been combined with the limited data from human studies to not only discover the genetic determinants of susceptibility, but to also understand the molecular underpinnings of toxicity (Rusyn et al. 2010). While our i*n vivo* studies have been exploring the role that genetic variability may play in the mechanisms of chemical toxicity (Bradford et al. 2011; Gatti et al. 2009a; Gatti et al. 2011; Gatti et al. 2010b; Harrill et al. 2009), the throughput of the studies in experimental animals is low.

Over the years, various liver-derived *in vitro* model systems have been developed to enable investigation of the potential adverse effects of chemicals and drugs (Soldatow et al. 2013). We have been working to establish a toxico-genetic model of liver toxicity in cultured mouse hepatocytes (Martinez et al. 2009). We 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. In a study of primary mouse hepatocytes from 15 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 (Martinez et al. 2010). 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.

Specific Objective 3: Discover chemical-induced regulatory networks using population-based toxicity phenotyping in human cells.

We have been working to establish a human *in vitro* population-wide toxicity testing system by using lymphoblast cell lines from the 1000 Genomes and HapMap projects. In our

first study, we collected *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 (Hatcher et al. 2009; O'Shea et al. 2011a; O'Shea et al. 2010; O'Shea et al. 2011b). This work has 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 served 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 this novel experimental approach has led to a much more ambitious collaboration with the Tox21 partner organizations, the National Toxicology Program and the NIH Chemical Genomics Center. In our second study, the population-based qHTS screening paradigm was explored in which hundreds of compounds may be profiled rapidly in dozens of cell lines and multiple biological targets. Our goal was to obtain data for predictive *in vitro* models of chemical-induced toxicity anchored on inter-individual genetic variability. Eighty one human lymphoblast cell lines from 27 CEPH trios were exposed to 240 chemicals at 12 concentrations and cell viability and apoptosis were evaluated (Lock et al. 2012).

In our third study, also in collaboration with NCGC and NTP, we successfully screened 1086 human lymphoblast cell lines, representing 9 populations from 5 continents, in a cell viability assay with 179 diverse environmental chemicals at 8 concentrations (Abdo et al. 2012; Abdo et al. 2013). The extent of inter-individual variability was less than 10 fold for about 2/3 of the 179 compounds; however, some compounds showed more than a 100-fold range. Across all 179 compounds, clustering of EC10 profiles revealed similarity of responses in cell lines according to the genetic ancestry. For individual chemicals, population differences were evident for about 40% of the compounds. Heritability analysis, using populations with parent-child trios, showed that cytotoxicity responses of 22 chemicals were significantly heritable (~20-40%). Because cell lines used in this study have been densely genotyped or sequenced, analyses of genome-wide association, and association pathway analyses, were performed to identify significant associations between variants/genes/pathways and cytotoxicity. Across the 179 compounds, 142 polymorphisms were identified as suggestive of the genetic association. Of the 1086 screened cell lines, basal gene expression (RNA sequencing-based) data is publicly available for 344 cell lines and we performed correlation analysis of this data with cytotoxicity across the 179 chemicals to identify additional genes and pathways that may also be associated with inter-individual differences in toxicity.

Overall, we conclude that genomic-anchored *in vitro* screening model using 1000 Genomes lymphoblast cell lines offers unique advantages for identifying variations in toxicity responses at the DNA sequence level and for defining experimentally-based confidence intervals for population variability.

Project 3
Development of validated and predictive Quantitative Structure-Toxicity Relationship models that employ both chemical and biological descriptors of molecular structures and take into account genetic diversity between individuals

This project was testing a hypothesis that the accuracy and interpretability of a computational model to predict chemical toxicity will be improved by including the results of short term biological assays as additional, "biological" descriptors of chemicals in combination with the traditional chemical descriptors (Zhu et al. 2008). Based on our expertise in cheminformatics we have expanded the concept of *in vitro – in vivo* extrapolation to include the explicit knowledge of chemical structure formulating the challenge of chemical toxicology as "structure – *in vitro* – *in vivo* extrapolation." During the course of this project, we have implemented several types of computational workflows that express this concept in the form of computational algorithms (Rusyn et al. 2012) as well as conducted and published multiple studies in support of this hypothesis.

Specific Objective 1: Develop rigorous end point toxicity predictors based on the QSAR modeling workflow and conventional chemical descriptors.

1.1. Models of mutagenicity

We have complied and curated the largest in the public domain database of more than 7,000 molecules tested in the Ames genotoxicity assay. We have formed international virtual collaboratory incorporating 14 QSAR modeling groups from eight countries which established that consensus QSAR model integrating individual predictions from individual laboratories achieved the highest statistical significance and external predictive power with the prediction accuracy exceeding 80% (Sushko et al. 2010). In a recent follow-up of this study, our group has developed QSAR models of mutation-site specific genotoxicity. Large data sets for each mutation target DNA sequence were compiled and curated by including data from additional, less common test strains. The resulting validated sequence-specific models have higher accuracy (65-79%) than previous models for individual bacterial assays (51-81%) (Low et al. 2013). In addition, we used carcinogenicity potency database to model carcinogen hazard (Wang et al. 2009; Zhu et al. 2008).

1.2. Models of environmental toxicity

We have modeled toxicity of 95 nitroaromatic compounds against the ciliate *T. pyriformis.* Multiple models have been generated with leave-one-out cross-validated R^2=0.82–0.95 for the training sets and R^2=0.57-0.86 for the test sets. The resulting models were employed for predicting toxicity of an external set of 63 nitroaromatics and R^2_{ext}=0.65 was obtained. Validated models were explored to (i) elucidate the effects of different substituents in nitroaromatic compounds on toxicity; (ii) differentiate compounds by probable mechanisms of toxicity based on their structural descriptors; (iii) analyze the role of various physical-chemical factors (*e.g.*, electrostatics, hydrophobicity, hydrogen bonding, atomic identity, etc.) in compounds' toxicity. Resulting models (Artemenko et al. 2011) can be used in assessment of environmental hazard of nitroaromatic chemicals.

1.3. Models of skin sensitization

We have recently developed QSAR models of skin sensitization (tested in local lymph node assay) and compared their external predictivity with that of the OECD QSAR toolbox

models. We applied kNN and RF algorithms to a dataset of 471 compounds, which was obtained from the 2009 annual report of Interagency Coordinating Committee on the Validation of Alternative Methods at NIEHS. In the course of external validation, our models demonstrated good sensitivity, specificity, and overall balanced accuracy (81-89%, 69-73%, and 77-79% respectively). Likewise evaluation by OECD QSAR toolbox showed prediction accuracy of only around 50%. We posit that our models will be useful in assessing skin sensitization potential of chemicals (Alves et al. 2013; Pu et al. 2011). We have made these models freely available for public use on our Chembench portal (chembench.mml.unc.edu).

1.4. Models of acute toxicity

In a collaborative study with EPA (Easterling et al. 2009; Zhu et al. 2009a; Zhu et al. 2009b; Zhu et al. 2009c; Zhu et al. 2009d; Zhu and Tropsha 2009) we have modeled the largest publicly available dataset of rat acute LD_{50}. The dataset of 7,385 compounds with their most conservative lethal dose (LD_{50}) values has been compiled by EPA. The prediction accuracy for the external validation set ranged from 0.24 to 0.70 (linear regression coefficient R2) depending on the use of applicability domain thresholds. The best model is publicly available on the Chembench portal (chembench.mml.unc.edu).

Specific Objective 2: Develop novel computational toxicogenomic models based on combined chemical and biological descriptors through QSAR modeling workflow.

2.1. Development of models that combine qHTS screening and chemical descriptors

We have explored *in vitro* IC_{50} cytotoxicity values and *in vivo* rodent LD_{50} values for over 300 chemicals from the ZEBET database (Zhu et al. 2009e). Conventional Quantitative Structure-Activity Relationship (QSAR) modeling of mouse or rat acute LD_{50} for ZEBET compounds yielded no statistically significant models. A novel two-step modeling approach was developed as follows. First, all chemicals are partitioned into two groups based on the relationship between IC_{50} and LD_{50} values. Second, conventional binary classification QSAR models are built to predict the group affiliation based on chemical descriptors only. Third, k-Nearest Neighbor (kNN) continuous QSAR models are developed for each sub-class to predict LD_{50} from chemical descriptors. Importantly, models resulting from this approach employ chemical descriptors only for external prediction of acute rodent toxicity.

To extend our previous successful QSAR modeling efforts using *in vitro* qHTS data, we aimed at incorporating dose-response data into the modeling workflow (Sedykh et al. 2010, 2011). Cell viability qHTS data for 1,408 compounds in 13 cell lines have been deposited in PubChem providing the opportunity to study the relationship between *in vitro* and *in vivo* effects. Hybrid QSAR models developed with chemical and noise-filtered qHTS descriptors outperformed conventional QSAR models. This study provides another confirmation that adding biological data, such as dose-response qHTS profiles, to conventional chemical descriptors can improve predictive power of toxicity prediction models.

2.2. ToxCast™ data analysis and modeling

We have been actively engaged in the analysis of the ToxCast™ Phase I data. Utilizing chemical descriptors did not yield statistically significant and externally predictive QSAR models resulted from these efforts. We then applied hierarchical and hybrid QSAR approaches (Rusyn et al. 2012) that combine the results of selected ToxCast *in vitro* bioassays with a subset of chemical descriptors. This, however, gave only limited improvement, indicating that Phase I data is lacking both chemical (structurally too diverse) and biological (reproducibility/noise

issues) dimensions. Nevertheless, we showed (Tang et al. 2010; Zhang et al. 2009a; Zhang et al. 2009b) that Toxcast *in vitro* assays can be used as biological profiles for establishing biological similarity between compounds, as well as for systematic mining of those *in vitro* assays with high correlation to ToxRefDB *in vivo* endpoints.

In (Golbraikh et al. 2013a; Golbraikh et al. 2013b; Golbraikh et al. 2012a; Golbraikh et al. 2012b), we have modeled ToxCast data, one example being *C. elegans* toxicity for 212 compounds from ToxCast™ Phase I. We have built variety of models using chemical and biological (ToxCast assays) descriptors: individually and in combination (as hybrid descriptors). The best performing was a hybrid kNN model that had externally validated accuracy of 73% (at 100% coverage). Using stricter thresholds for prediction confidence improves performance further (but for a subset of chemicals): to 78% (at 55% coverage), and to 83% (at 35% coverage).

2.3. Using toxicogenomics data as biological predictors.

QSAR modeling and toxicogenomics are used independently as predictive tools in toxicology. In a study by (Low et al. 2011a; Low et al. 2011b), we compared several models predicting drug hepatotoxicity in rats based on different descriptors of drug molecules, namely their chemical descriptors and toxicogenomic profiles. The target property was hepatotoxicity level following 28 days of treatment, established by histopathology and serum chemistry. First, we developed multiple conventional QSAR classification models using chemical descriptors and several classification methods. Next, we built models with the same classification methods using only toxicogenomic data treated as biological descriptors. Finally, hybrid models combining both chemical descriptors and transcripts were developed. Although the accuracy of hybrid models did not exceed that of models based on toxicogenomic data, the use of both types of descriptors enriched the interpretation of the models. We concluded that concurrent exploration of chemical features and acute treatment-induced changes in gene transcript levels will (i) enrich the mechanistic understanding of subchronic liver injury and (ii) afford models capable of accurate prediction of hepatotoxicity from chemical structure and short-term assay results.

Specific Objective 3: Develop novel computational toxicogeneitic models based on combined genetic, chemical and toxicity descriptors through QSAR-like modeling workflow.

Population-based toxicity screening has been conducted against 81 cell lines to test a dataset comprised of 240 compounds (in the previous studies, only 13 cell lines were explored). These compounds were selected from the 1,408 substances of the NTP screening library. Each compound was tested in each cell line using cell viability and caspase assays. Using this data, we have applied our standard modeling workflow to develop models of rat acute toxicity, Ames mutagenicity, and rodent carcinogenicity, using both conventional chemical descriptors as well as noise-treated qHTS data employed as biological descriptors (Sedykh et al. 2012). Interestingly, biological descriptors alone afforded acceptable QSAR models (accuracy of 55-63%), although hybrid models were still the best performing (61-72%). Importantly, the benefit from using *in vitro* data was varying and corresponded to the complexity of *in vivo* endpoint (most benefit - for acute toxicity, carcinogenicity; least - for mutagenicity). Our population-based hybrid model of acute toxicity shows higher accuracy (72±2%) than the previously published models (68±2% on the same external set).

Administrative Core

The Administrative Core Unit was responsible for the overall planning and coordination of the Center's research activities, fostering interdisciplinary interaction, providing leadership on public outreach/translation, managing the Quality Assurance, scheduling meetings of the Executive Committee and Internal and External Advisory Committees, as well as fiscal and resource management and planning. The Core worked closely with the Research Projects to ensure effective communication with the EPA, government agencies, the scientific community, and potential users of the technology developed by our research team.

With regards to administrative and public outreach functions, the core has been instrumental in maintaining very high level of interactions between the Research Projects by coordinating regular Center meetings. 7-8 Center-wide meetings were held yearly. In addition, every year the Center held annual meetings with the External Advisory Board.

With regards to the communication and public outreach functions, the Center has established a web-site (http://comptox.unc.edu) which contains information about all components and an up-to-date listing of news items, PDFs and lists of all publications and abstracts, as well as the links to software tools developed. Furthermore, regular meetings with EPA NCCT personnel were held (every 3-6 months) by each of the Research Projects and the work of the Center was disseminated through scientific publications, as well as oral and poster presentations at the national and international meetings, including those organized by the EPA.

Quality Assurance:

All Research Projects have developed Quality Assurance Project Plans (QAPP) and Dr. Karin Yeatts, the Quality Assurance Manager, approved these and was performing annual audits of all projects. The audits identified some deficiencies and minor corrective actions were recommended, as needed, to projects. The results of the technical audits were distributed to the principal investigators as well as the EPA Project Officers annually. The adequacy of corrective actions was verified annually and the revised QAPPs, where necessary, were approved by the QA Manager. The outcomes of the post-correction action verification process were communicated with EPA/NCER through memos on the updates on the annual audits and corrective actions. The QA process, outcome of the annual audits and corrective actions, if needed, were communicated to the External Advisory Board at the annual meetings.

Relatively few environmental measurements or primary data were generated by the Center, with the exception of experiments in Project 2. SOPs for the data generation and the associated QAPP for Project 2 covered data generation and analyses. Data generated at EPA as part of the ToxCast Phase I and DSSTox databases was already under EPA quality management oversight, and used as part of the research of the Center. Beyond existing data used by the Center, derived analyses, analysis scripts, and developed software posed quality management challenges that were addressed by each project's QAPPs and the Center QMP.

Publications (#)/Presentations (*):

*Abdo N, Xia M, Kosyk O, Huang R, Sakamuru S, Austin C, et al. 2012. The 1000 genomes toxicity screening project: Utilizing the power of human genome variation for population-scale in vitro testing. In: Society of Toxicology Annual Meeting. San Francisco, CA.

*Abdo N, Xia M, Kosyk O, Huang R, Sakamuru S, Brown C, et al. 2013. The 1000 genomes toxicity screening project: Utilizing the power of human genome variation for population-scale in vitro testing. In: Society of Toxicology Annual Meeting. San Antonio, TX.

#Abell AN, Jordan NV, Huang W, Prat A, Midland AA, Johnson NL, Granger DA, Mieczkowski PA, Perou CM, Gomez SM, Li L, Johnson GL. 2011. MAP3K4/CBP-Regulated H2B Acetylation Controls Epithelial-Mesenchymal Transition in Trophoblast Stem Cells. Cell Stem Cell. 8:525-37.

#Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer NC, Andrade CH, et al. 2013. QSAR models of skin sensitization and skin permeability and their application to identifying potentially hazardous chemicals. Chem Res Toxicol: submitted.

#Artemenko AG, Muratov EN, Kuz'min VE, Muratov NN, Varlamova EV, Kuz'mina AV, et al. 2011. QSAR analysis of the toxicity of nitroaromatics in Tetrahymena pyriformis: structural factors and possible modes of action. SAR QSAR Environ Res 22(5-6): 575-601.

#Berginski ME, Vitriol EA, Hahn KM, Gomez SM. 2011. High-resolution quantification of focal adhesion spatiotemporal dynamics in living cells. PLoS ONE 6(7): e22025.

#Bradford BU, Lock EF, Kosyk O, Kim S, Uehara T, Harbourt D, et al. 2011. Interstrain differences in the liver effects of trichloroethylene in a multistrain panel of inbred mice. Toxicol Sci 120(1): 206-217.

#Chaki SP, Barhoumi R, Berginksi ME, Sreenivasappa H, Trache A, Gomez SM, Rivera GM. 2013. Nck enables directional cell migration through the coordination of polarized membrane protrusion with adhesion dynamics. J Cell Sci. 126:1637-1649.

#Chen Z, Lessey E, Berginski ME, Cao L, Li J, Trepat X, Itano M, Gomez SM, Kapustina M, Huang C, Burridge K, Truskey G, Jacobson K. 2013. Gleevec, an Abl Family Inhibitor, Produces a Profound Change in Cell Shape and Migration. PLoS ONE. 8(1): e52233.

#Choi K, Gomez SM. 2009. Comparison of phylogenetic trees through alignment of embedded evolutionary distances. BMC Bioinformatics 10:423.

*Choi K, Staab J, Elston TC, Gomez SM. 2009. Dimension reduction and functional relevance in ToxCast chemical toxicity data. EPA ToxCast Data Analysis Summit, Durham, USA.

#Doolittle JM, Gomez SM. 2010. Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. Virology J, 7:82.

#Doolittle JM, Gomez SM. 2011. Mapping protein interactions between Dengue virus and its human and insect hosts. PLoS Neglected Tropical Diseases, 5(2):e954.

#Duncan JS, Whittle MC, Nkamura K, Abell AN, Midland AA, Zawistowksi JS, Johnson NL, Granger DA, Jordan NV, Darr D, Usary J, Major B, He X, Hoadley K, Sharpless NE, Perou CM, Gomez SM, Jin J, Frye SV, Earp HS, Graves LM, Johnson GL. 2012. Dynamic

reprogramming of the kinome in response to targeted MEK inhibition in triple negative breast cancer. Cell. 149(2);307-21.

*Easterling M, Tropsha A, Ye L, Pirone J, Zhu H, Martin TM, et al. 2009. Quantitative Structure Toxicity Relationships (QSTR) models for predicting acute/sub-acute and sub-chronic/chronic adverse effect levels. In: Society of Toxicology Annual Meeting. Baltimore, MD.

#Garcia GJM, Picher M, Zuo P, Okada SF, Lazarowski ER, Button B, Boucher RC, Elston TC. 2011. Computational model for the regulation of extracellular ATP and adenosine in airway epithelia. In: Purinergic Regulation of Respiratory Diseases (M. Picher and R.C. Boucher, eds.), Chapter 3, Springer Publishing.

#Gatti DM, Barry WT, Nobel AB, Rusyn I, Wright FA. 2010a. Heading down the wrong pathway: on the influence of correlation within gene sets. BMC Genomics 11: 574.

#Gatti DM, Harrill AH, Wright FA, Threadgill DW, Rusyn I. 2009a. Replication and narrowing of gene expression quantitative trait loci using inbred mice. Mamm Genome 20(7): 437-446.

#Gatti DM, Lu L, Williams RW, Sun W, Wright FA, Threadgill DW, et al. 2011. MicroRNA expression in the livers of inbred mice. Mutat Res 714(1-2): 126-133.

#Gatti DM, Shabalin AA, Lam TC, Wright FA, Rusyn I, Nobel AB. 2009b. FastMap: fast eQTL mapping in homozygous populations. Bioinformatics 25(4): 482-489.

#Gatti DM, Sypa M, Rusyn I, Wright FA, Barry WT. 2009c. SAFEGUI: resampling-based tests of categorical significance in gene expression data made easy. Bioinformatics 25(4): 541-542.

#Gatti DM, Zhao N, Chesler EJ, Bradford BU, Shabalin AA, Yordanova R, et al. 2010b. Sex-specific gene expression in the BXD mouse liver. Physiol Genomics 42(3): 456-468.

#Golbraikh A, Muratov E, Fourches D, Sedykh A, Tropsha A. 2013a. QSAR modelability index for chemical datasets. J Chem Inf Model: submitted.

*Golbraikh A, Sedykh A, Muratov E, Shah R, Boyd W, Smith M, et al. 2013b. Modelability of ToxcastTM Phase I datasets. In: Society of Toxicology Annual meeting. San Antonio, TX.

*Golbraikh A, Shah R, Sedykh A, Boyd W, Smith M, Zhu H, et al. 2012a. Predictive in silico models of toxicity using C.elegans growth assay along with ToxCast Phase I short-term assay data. In: Society of Toxicology Annual meeting. San Francisco, CA.

#Golbraikh A, Wang XS, Zhu H, Tropsha A. 2012b. Predictive QSAR Modeling: Methods and applications in drug discovery and chemical risk assessment. In: Handbook of Computational Chemistry, (Leszczynski J, ed). New York, NY:Springer, 1311-1342.

*Gomez, SM. 2011. Predictive modeling of chemical-perturbed regulatory networks in systems toxicology. In: Third Toxicogenomics Integrated with Environmental Sciences (TIES) international conference. Chapel Hill, NC.

#Harrill AH, Ross PK, Gatti DM, Threadgill DW, Rusyn I. 2009. Population-based discovery of toxicogenomics biomarkers for hepatotoxicity using a laboratory strain diversity panel. Toxicol Sci 110(1): 235-243.

*Hatcher S, Kosyk O, Ross PK, Wright F, Schwartz J, Dix D, et al. 2009. In vitro screening for chemical toxicity in a genetically-diverse human model system. In: EPA ToxCast Data Analysis Summit. Durham, NC.

#Herschlag G, Garcia GJ, Button B, Tarran R, Lindley B, Reinhardt B, Elston TC, Forest MG. 2013. A mechanochemical model for auto-regulation of lung airway surface layer volume. J Theor Biol. May 21; 325:42-51.

*Jeffries RE, Xu K, Gomez S, Macdonald J, Gamcsik M. 2012. Modeling and experimental correlation of the glutathione metabolic network. In: BMES Annual Conference, Atlanta, GA.

#Kesseler KJ, Blinov ML, Elston TC, Kaufmann WK, Simpson DA. 2013. A predictive mathematical model of the DNA damage G2 checkpoint. J Theor Biol. 320:159-69.

*O'Connell TM, Staab J, Gomez SM. 2009. Enhancing metabolomic data analysis with Progressive Consensus Alignment of NMR Spectra (PCANS). In: Fifth International Conference of the Metabolomics Society, Edmonton, Canada.

#Lee S, Wright FA, Zou F. 2011. Control of Population Stratification by Correlation-Selected Principal Components. Biometrics 67(3): 967-974.

#Lock EF, Abdo N, Huang R, Xia M, Kosyk O, O'Shea SH, et al. 2012. Quantitative high-throughput screening for chemical toxicity in a population-based in vitro model. Toxicol Sci 126(2): 578-588.

*Low Y, Sedykh A, Chakravarti S, Saiakhov R, Tropsha A. 2013. In silico models of sequence-specific mutagenicity: Exploring chemical-mutation site associations. In: Society of Toxicology Annual meeting. San Antonio, TX.

*Low Y, Uehara T, Minowa Y, Yamada H, Ohno Y, Urushidani T, et al. 2011a. Predictive value of chemical and toxicogenomic descriptors for drug-induced hepatotoxicity. In: Society of Toxicology Annual Meeting. Washington, DC.

#Low Y, Uehara T, Minowa Y, Yamada H, Ohno Y, Urushidani T, et al. 2011b. Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. Chem Res Toxicol 24(8): 1251-1262.

*Martinez SM, Bradford B, Kaiser R, Soldatow VY, Amaral K, Ferguson SS, et al. 2009. Development of the in vitro toxicogenetic models for hepatotoxicity. In: Society of Toxicology Annual Meeting. Washington, DC.

#Martinez SM, Bradford BU, Soldatow VY, Kosyk O, Sandot A, Witek R, et al. 2010. Evaluation of an in vitro toxicogenetic mouse model for hepatotoxicity. Toxicol Appl Pharmacol 249(3): 208-216.

*O'Shea SH, Kosyk O, Abdo N, Lock EF, Wright F, Huang R, et al. 2011a. Population-based quantitative high throughput screening (qHTS) for chemical toxicity. In: Society of Toxicology Annual Meeting. Washington, DC.

*O'Shea SH, Schwartz J, Kosyk O, Ross PK, Wright F, Tice R, et al. 2010. In vitro screening for population variability in chemical toxicity. In: Society of Toxicology Annual Meeting. Salt lake City, UT.

[#]O'Shea SH, Schwarz J, Kosyk O, Ross PK, Ha MJ, Wright FA, et al. 2011b. In vitro screening for population variability in chemical toxicity. Toxicol Sci 119(2): 398-407.

*Pu D, Zhu H, Zhao G, Strickland J, Salicru E, Tropsha A. 2011. Quantitative Structure-Activity Relationship Modeling of Skin Sensitizers Tested in Local Lymph Node. In: Society of Toxicology Annual Meeting. Washington, DC.

[#]Rusyn I, Daston GP. 2010. Computational toxicology: realizing the promise of the toxicity testing in the 21st century. Environ Health Perspect 118(8): 1047-1050.

[#]Rusyn I, Gatti DM, Wiltshire T, Kleeberger SR, Threadgill DW. 2010. Toxicogenetics: population-based testing of drug and chemical safety in mouse models. Pharmacogenomics 11(8): 1127-1136.

[#]Rusyn I, Sedykh A, Low Y, Guyton KZ, Tropsha A. 2012. Predictive modeling of chemical hazard by integrating numerical descriptors of chemical structures and short-term toxicity assay data. Toxicol Sci 127(1): 1-9.

*Sedykh A, Low Y, Lock EA, Rusyn I, Tropsha A. 2012. Using population-based dose-response cytotoxicity data for in silico prediction of rat acute toxicity. In: Society of Toxicology Annual Meeting. San Francisco, CA.

*Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, et al. 2010. Using in vitro Dose-Response Profiles to Enhance QSAR Modeling of in vivo Toxicity. In: Society of Toxicology Annual Meeting. Salt Lake City, UT.

[#]Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, et al. 2011. Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. Environ Health Perspect 119(3): 364-370.

[#]Soldatow VY, LeCluyse EL, Griffith LG, Rusyn I. 2013. In vitro models for liver toxicity testing. Toxicol Res 2(1): 23-39.

*Staab J, Choi K, Elston TC, Gomez SM. 2009. Establishing a biological context for ToxCast chemical toxicity data. In: EPA ToxCast Data Analysis Summit, Durham, NC.

[#]Staab J, O'Connell TM, Gomez SM. 2010. Enhancing metabolomic data analysis with Progressive Consensus Alignment of NMR Spectra (PCANS). BMC Bioinformatics. 11: 123.

[#]Strychalski W, Adalsteinsson D, Elston TC. 2010a. A cut-cell method form modeling spatial sytems of biochemical reactions in arbitrary cell geometries. Comm Appl Math Comp Sci 5: 31-53.

[#]Sun W. 2012. A statistical framework for eQTL mapping using RNA-seq data. Biometrics 68(1): 1-11.

[#]Sun W, Buck MJ, Patel M, Davis IJ. 2009a. Improved ChIP-chip analysis by a mixture model approach. BMC Bioinformatics 10: 173.

[#]Sun W, Hu Y. 2013. eQTL Mapping Using RNA-seq Data. Stat Biosci 5(1): 198-219.

[#]Sun W, Wright FA. 2010. A geometric interpretation of the permutation p-value and its application in eQTL studies. Ann Appl Stat 4(2): 1014-1033.

[#]Sun W, Wright FA, Tang Z, Nordgard SH, Loo PV, Yu T, et al. 2009b. Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic Acids Res. 37(16): 5365-5377.

[#]Sun W, Xie W, Xu F, Grunstein M, Li KC. 2009c. Dissecting nucleosome free regions by a segmental semi-Markov model. PLoS One 4(3): e4721.

[#]Sushko I, Novotarskyi S, Korner R, Pandey AK, Cherkasov A, Li J, et al. 2010. Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. J Chem Inf Model 50(12): 2094-2111.

*Tang H, Zhu H, Zhang L, Sedykh A, Richard A, Rusyn I, et al. 2010. Toxicity reference database (ToxRefDB) to develop predictive toxicity models and prioritize compounds for future toxicity testing. In: 239th ACS National Meeting. San Francisco, CA.

*Wang K, Golbraikh A, Tropsha A. 2009. Optimization of pattern recognition and classification by combinatorial QSAR modeling of the carcinogenic potency database. In: 237th ACS National Meeting. Salt Lake City, UT.

[#]Wright FA, Shabalin AA, Rusyn I. 2012. Computational tools for discovery and interpretation of expression quantitative trait loci. Pharmacogenomics 13(3): 343-352.

[#]Xia K, Shabalin AA, Huang S, Madar V, Zhou YH, Wang W, et al. 2012. seeQTL: a searchable database for human eQTLs. Bioinformatics 28(3): 451-452.

*Xu K, Kirchberg C, Gomez SM, Morgan K. 2011a. Development of a transcript profile for fasting as a confounding variable in toxicogenomic studies. In: Society of Toxicology Annual Meeting, Washington DC.

[#]Xu K, Morgan KT, Elston TC, Gomez SM. 2011b. A whole-body model for glycogen regulation reveals a critical role for substrate cycling in maintaining blood glucose homeostasis. PLoS Computational Biology 7: e1002272.

*Zhang L, Judson R, Dix D, Houck DR, Martin M, Richard A, et al. 2009a. Cheminformatics Analysis of EPA ToxCast chemical libraries to develop predictive toxicity models and prioritize compounds for in vivo toxicity testing. In: EPA ToxCast Data Analysis Summit. Durham, NC.

*Zhang L, Zhu H, Rusyn I, Judson R, Dix D, Houck K, et al. 2009b. Cheminformatics Analysis of EPA ToxCast chemical libraries to identify domains of applicability for predictive toxicity models and prioritize compounds for toxicity testing. In: Society of Toxicology Annual Meeting. Baltimore, MD.

[#]Zhou YH, Barry WT, Wright FA. 2013. Empirical pathway analysis, without permutation. Biostatistics 14(3): 573-585.

[#]Zhou YH, Xia K, Wright FA. 2011. A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics 27(19): 2672-2678.

[#]Zhu H, Martin TM, Ye L, Sedykh A, Young DM, Tropsha A. 2009a. Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. Chem Res Toxicol 22(12): 1913-1921.

*Zhu H, Martin TM, Ye L, Young DM, Tropsha A. 2009b. Combinatorial QSAR Modeling of Rat Acute Toxicity by Oral Exposure. In: Society of Toxicology Annual Meeting. Baltimore, MD.

#Zhu H, Rusyn I, Richard A, Tropsha A. 2008. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. Environ Health Perspect 116(4): 506-513.

*Zhu H, Sedykh A, Wright FA, Rusyn I, Tropsha A. 2009c. Using quantitative High Throughput Screening (q-HTS) results as biological descriptors to assist modeling of acute rat toxicity. In: 238th ACS National Meeting. Washington, DC.

*Zhu H, Sedykh A, Zhang L, Rusyn I, Tropsha A. 2009d. Using ToxCast cell-viability and gene-expression assays as biological descriptors in QSAR modeling of animal toxicity endpoints. In: EPA ToxCast Data Analysis Summit. Durham, NC.

*Zhu H, Tropsha A. 2009. The use of hybrid chemical/biological descriptors in QSAR modeling improves the accuracy of in vivo chemical toxicity prediction. In: The 32nd Annual Midwest Biopharmaceutical Statistics Workshop. Muncie, IN.

#Zhu H, Ye L, Richard A, Golbraikh A, Wright FA, Rusyn I, et al. 2009e. A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. Environ Health Perspect 117(8): 1257-1264.

#Zou F, Lee S, Knowles MR, Wright FA. 2010. Quantification of population structure using correlated SNPs by shrinkage principal components. Hum Hered 70(1): 9-22.

Supplemental Keywords: modeling, mechanistic models, hybrid models, bioinformatics, metabolism, *in vitro*, eQTL, genetics, genomics, screening, toxicity, QSAR, machine learning, short term assays,

Relevant Web Sites:
http://comptox.unc.edu/resources.php
https://bitbucket.org/gomezlab/xceed
http://chembench.mml.unc.edu