Evaluating the Boundaries of Toxicity Pathway Space Using High-Throughput Environmental Chemical Data

Holly Mortensen, David Reif, David Dix, Keith Houck, Robert Kavlock, Richard Judson

U.S EPA, ORD, Computational Toxicology Research Program

research&development

Science Question

ToxCast Phase I HTS assays have been used to construct target gene lists linked with public information on genes and proteins, molecular, biological, and cellular pathways/processes, and disease. Currently there is no gold-standard for analysis of available gene-pathway interaction data, and most studies to date have focused on a single data source.

We are performing pathway inference and network analyses, with the aim being understanding the links between chemical exposure and adverse health outcomes at the level of pathways rather than individual targets. This approach permits exploration of disease at a higher level of cellular and organismal organization, focusing on multiple, related disorders, and may aid in the understanding of common disease outcomes (*e.g.* cancer or immune disorders) that are characterized by locus heterogeneity. Through the use of the ToxMiner™ database and the analysis framework presented here, we hope to gain insight into complex relationships between disease states in humans and environmental chemicals.

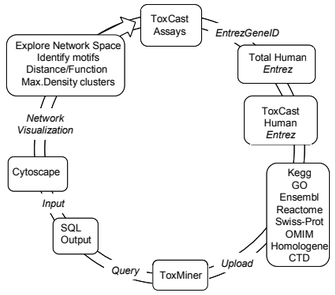
1. Is global pathway "space" made up of smaller "modules" that have biological relevance?
2. How much of that global space is defined by toxicity and toxicity-related pathways that may be perturbed by environmental chemicals?
3. Which toxicity pathways are targeted by ToxCast Phase I assays, and which pathways relevant to human toxicity and disease need to be the focus of future research efforts?

Research Goals

1. To cluster all human genes into a set of functionally consistent modules that can be associated with toxicity pathways.
2. To use statistical and network analyses to build and visualize these clusters, consolidating multiple, publically available pathway resources.
3. To annotate these clusters with disease association and phenotype information.
4. To delineate the current assay coverage across toxicity and disease related pathways, and to propose other targets for future assay development to help probe important regions of pathway space.

Methods/Approach

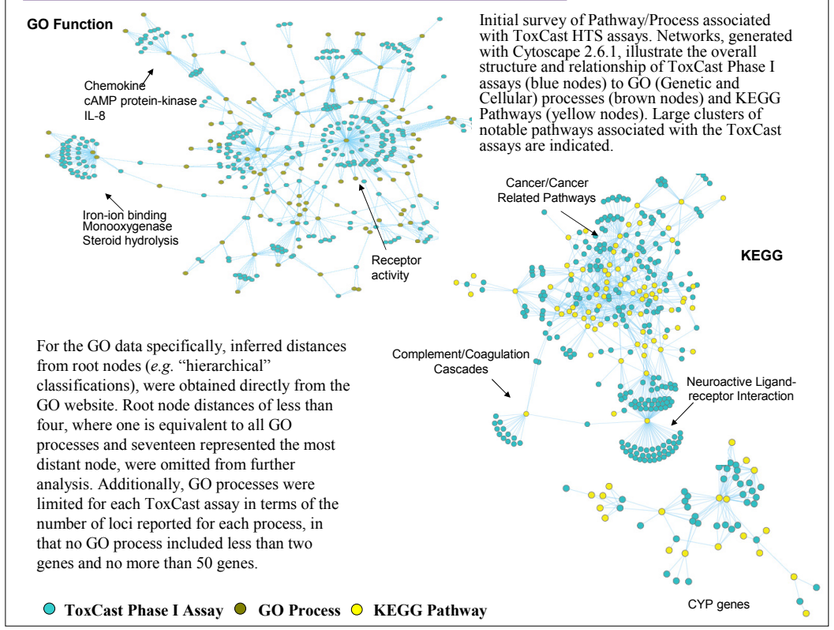
Database Generation



Schematic of the workflow for mapping ToxCast data to publicly available information The ToxMiner database was created in order to link biological, metabolic, and cellular pathway data from multiple sources to genes and *in vitro* assay data for the chemicals screened in Phase I. Also included in ToxMiner is human disease information, which correlates with ToxCast assays that target specific genetic loci. These data are preprocessed and consolidated via scripts written in perl and implemented using MySQL, and can be accessed and queried. The ToxMiner database extends the currently available ACToR (Aggregated Computation Toxicology Resource) database, which captures information on chemicals and assays of chemical-biological effects (Judson, *et al.*, 2008).

Pathway Source	*ToxCast Human*	*Total Human*
GO	1126	7213
KEGG	94	202
Ingenuity	99	110
Total Unique Entrez GeneID	231	18,187

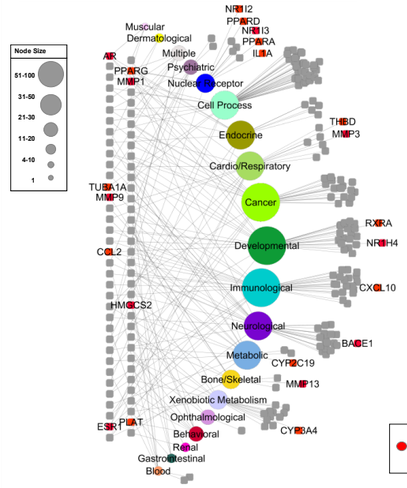
Pathway Network Construction and Analyses



Initial survey of Pathway/Process associated with ToxCast HTS assays. Networks, generated with Cytoscape 2.6.1, illustrate the overall structure and relationship of ToxCast Phase I assays (blue nodes) to GO (Genetic and Cellular) processes (brown nodes) and KEGG Pathways (yellow nodes). Large clusters of notable pathways associated with the ToxCast assays are indicated.

For the GO data specifically, inferred distances from root nodes (*e.g.* "hierarchical" classifications), were obtained directly from the GO website. Root node distances of less than four, where one is equivalent to all GO processes and seventeen represented the most distant node, were omitted from further analysis. Additionally, GO processes were limited for each ToxCast assay in terms of the number of loci reported for each process, in that no GO process included less than two genes and no more than 50 genes.

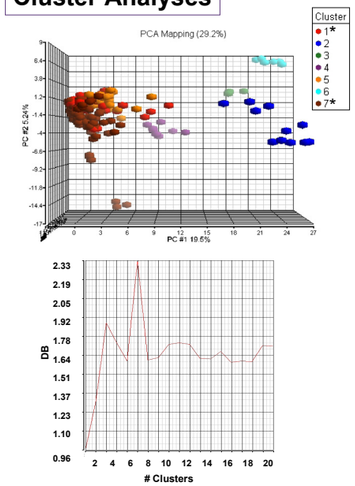
Gene-Disease Associations



Network diagram of the genetic loci and corresponding disorder classes probed by ToxCast Phase I HTS Assays. Colors denote disease types defined by OMIM disease-gene associations and based loosely on disease classes used by Goh *et al*. PNAS (2007). Node size corresponds to the number of genes present in each disorder class. All genes are illustrated in gray, with the exception of those shown in red that are discussed in the recent article by Judson *et al*. (*2009*).

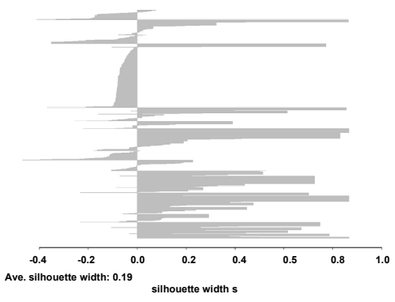
Genetic Loci Associated with Liver Proliferative Lesions in Rat

Cluster Analyses



The Davies-Bouldin (DB) index was used for cluster validation. This index is a function of the ratio of the sum of within cluster scatter to between cluster separation (Davies and Bouldin, 1979). We confirm the presence of seven distinct clusters in the Ingenuity dataset, where cluster 1* and 7* include 3 and 8 subclusters, respectively (results not shown).

The Principle Components Analysis was applied to a gene x gene distance matrix generated from Ingenuity pathway data, using the Partek Discovery Suite. We observe seven main clusters in the Ingenuity pathway dataset, using all 24,763 annotated loci (3300 genes/110 pathways). Of those seven, two clusters, indicated with *, are "complex" and contain multiple subclusters.



The plot of cluster identity (Ingenuity n=2952) was generated using the Silhouette validation technique (Rousseeuw, 1987) within the function pam, using the statistical package R. The pam algorithm is based on the search for k representative objects or medoids among the observations of the dataset. Each bar represents a gene, where increasing positive values for a group of genes indicates strong cluster identity. Small and negative values indicate increased fuzziness of classification.

Results/Conclusions

The ToxMiner database, which links biological, metabolic, and cellular pathway data from multiple sources to genes and *in vitro* assay data for the chemicals screened in Phase I, as well as complementary data for all annotated genes, is currently in place and being used to generate hypotheses regarding the structure of "toxicity pathway space". Current efforts are focused on the application of undirected clustering optimization of this global pathway dataset into biologically relevant modules.

Pathway sources have been consolidated for each gene in terms of its presence or absence in all pathways, creating a binary matrix. We have estimated the similarity/dissimilarity between all gene pairs, initially using the Ingenuity pathway-gene information, where proximity between genes implies biological/functional similarity. Upon validation of biological relevance of each cluster, we will expand this workflow to include other pathway sources.

Impact and Outcomes

The initial survey of the pathway data and gene-disease associations has allowed visualization of particular pathway/processes affected by genes targeted by ToxCast Phase I assays, placing these data into their biological context. We have observed that the majority of the current assays probe pathways that are associated with immunological, developmental, and cancer related pathways.

Future Directions

Current efforts are focused on: describing global pathway space for all annotated human genes (currently 24,763 loci), defining coverage of toxicity and disease related pathways across that space, and determining the targets of current, as well as future assays.

References

1. D.L. Davies, D.W. Bouldin. A cluster separation measure. 1979. IEEE Trans. Pattern Anal. Machine Intell. I (4). 224-227.
2. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabasi, A-L (2007) The human disease network. *Proc Natl Acad Sci* 104: 8685-8690
3. Judson, R., Richard, A., Dix, D., *et al.* (2008). ACToR Aggregated Computational Toxicology Resource. *Toxicol. Appl. Pharmacol.* 233, 7-13.
4. Judson, R., Houck,K., Kavlock,R.,Knudsen, T., Martin, M., Mortensen, H., Reif, D., Richard, A., Rotroff, D., Shah, I., Dix, D."Predictive *In Vitro* Screening of Environmental Chemicals – The ToxCast Project", *Environmental Health Perspectives* (*Submitted 2009*)
5. P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. 1987. Journal of Computational and Applied Mathematics. 20. 53-65.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY COMPUTATIONAL TOXICOLOGY