Identifying Functionally Linked Gene Modules Within Biological Pathways Assessed by ToxCast *In Vitro* Assays

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

U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology

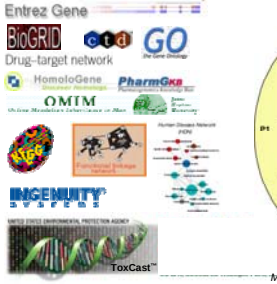
Introduction

The new toxicity-testing paradigm envisioned in the National Research Council's report on Toxicity Testing in the 21st Century (2007) relies on "understanding 'toxicity pathways'-defined as the cellular response pathways that can result in adverse health effects when sufficiently perturbed". Still "toxicity pathways" remain elusive. This is in part because we are faced with incongruence between the toxicological process, how it relates to disease (*i.e.* at what point does toxicity equate with disease), and how we can effectively understand toxicological processes on a global scale. With incoming data from HTS of environmental chemicals, there is a need to simultaneously characterize toxicity pathways for multiple, integrated and diverse biological processes. Currently, our knowledge of pathways that are affected by environmental chemicals, and the role of perturbation of those pathways in human disease, is limited. To understand toxicity and subsequently disease etiology, we must define toxicity space. In order to address the main goals of the EPA of predicting toxicological endpoints and mode of action, phasing out costly animal studies, as well as near-term goals of selecting *in vitro* assays, it is necessary to obtain a mechanistic understanding of the process of toxicity in relation to environmental chemicals. Herein, assay target selection and coverage across biological space becomes critical.

Primary Aims

1. **Characterize pathway information on a global scale, in order to inform future assay target selection**
2. **Define biologically coherent pathway modules**
3. **Map the location of chemical binding, disease phenotype, and drug target information, to improve understanding of the regions of pathway space occupied by toxicological processes**

Current ToxMiner Data Sources



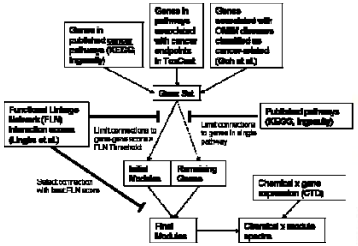
ToxMiner Database Resource: We implement publically available data from multiple sources, as well as ToxCast Phase I assay targets, to define biologically coherent modules on a global scale

Pathway Module Creation

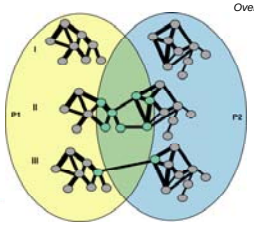
Human Genes and Pathways Represented in ToxMiner

	n Genes	*Pathways*	*ToxCast*
Entrez	24,763		
GO	17,891	7,213	1,104
KEGG	4,215	201	166
Ingenuity	3,300	110	180
Pathway Commons	3,266	1050	
ToxCast	231		
CTD	9,317		
Disease Genes[a]	4,135		
Drug Targets[b]	1,288		

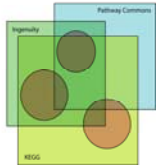
[a] Goh *et al.* (2007)
[b] Yildirim *et al.* (2007);Klein, *et al.* (2001)



Overview of workflow for module creation



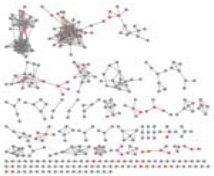
Module merge/split. I: *non-overlap*; II: *do not merge*(<intersection redundancy); III: *merge* (>intersection redundancy)



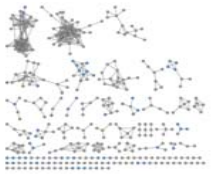
Intuitive visualization: Modules represent coordinated regulation of genes within pathways

Global Module Construction

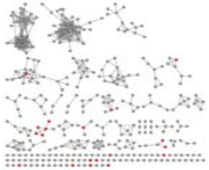
CTD Binding



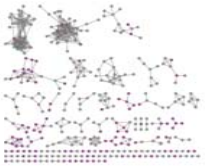
Druggable Genome



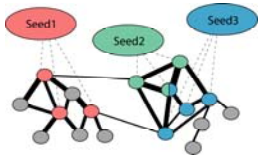
Cancer Associations



Disease Genome

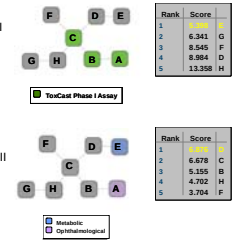


Intelligent Assay Prioritization



Neighborhood Decision Rule

$$S_i = \sum_{j=1}^{J_i} W_j \qquad (1)$$

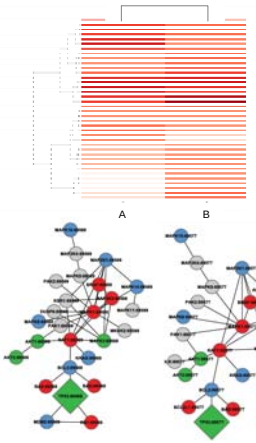


Rank	Score	
1	6.341	G
2	8.545	F
3	8.984	D
4	13.358	H

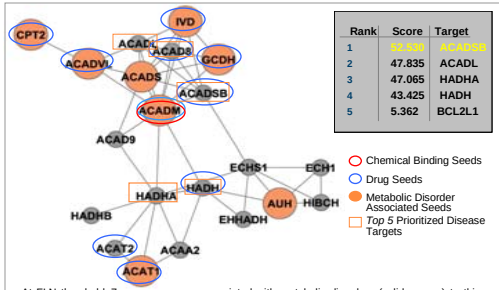
Rank	Score	
1	6.678	C
2	5.155	B
3	4.702	H
4	3.704	F

Module interrogation for future assay selection: We have implemented *Intelligent Assay Selection* to select assay targets in a hypothesis driven manner. Here we present two examples of seed identification (*right*: I-II) and subsequent target ranking, focusing on maximization of assay coverage and disease gene relationship. We have applied a nearest neighbor decision rule (Eq.1) to rank genes (*i*) within modules according to biological relationship to seed genes (*j*), using FLN likelihood scores between gene pairs.

Global modules were constructed according to the workflow presented in 1A, using 24,763 annotated human genes. Module growth was limited using published pathways from KEGG, Ingenuity and Pathway Commons, as well as an FLN threshold of 10. Seed genes were identified for global pathway modules and cancer specific modules, using gene association information from *Chemical-Gene Associations:* Comparative Toxicogenomics Database-CTD; *Drug Target Sources:* PharmGKB, The Druggable Genome (Yildirim, 2007); *Disease Genome Sources:* OMIM, The Human Disease Network (Goh, 2007); *Cancer Gene Inventories*: Memorial, Sloan-Kettering, Sanger Center, UCSF, Oncogene; *EPA*: ToxCast. *A seed gene* is considered as any node that has chemical, drug, disease, or assay association.

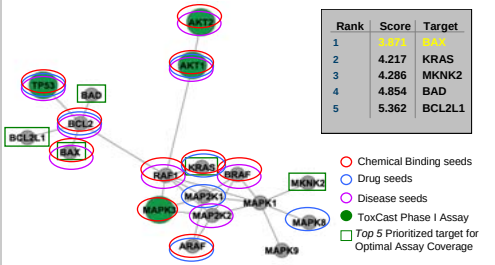


At FLN threshold=7, we identify two cancer-associated modules (A and B) that contain TP53. Chemicals that hit modules A and B most frequently are identified as cancer therapeutics. Green=ToxCast Phase I Assay; Red=Chemical Binding seeds; Blue=Drug seeds



Rank	Score	Target
1	52.530	ACADSB
2	47.835	ACADL
3	47.065	HADHA
4	43.425	HADH
5	5.362	BCL2L1

At FLN threshold=7, we map genes associated with metabolic disorders (solid orange) to this module, defined by cancer and cancer-related pathways, and list the top 5 potential targets. Also shown are the loci that have chemical, drug, and disease association information.



Rank	Score	Target
1	3.871	BAX
2	4.217	KRAS
3	4.286	MKNK2
4	4.854	BAD
5	5.362	BCL2L1

Illustrated are the ToxCast Phase I assay seeds (solid green) for this TP53 associated module, and the top five, ranked targets (green boxes). Also shown are the loci that have chemical, drug, and disease association information.

Conclusions/Future Directions

We have presented an approach to integrate gene and pathway information on a global scale, in order to create a set of pathway modules. The modules are made up of functionally-related genes, where relatedness includes co-expression, molecular function, protein domain sharing, etc. The modules are then linked with chemical, drug and disease association information. The utility of this approach is evidenced by an increased scientific understanding of the coverage of toxicological processes across biological space, as well as the more pragmatic objective of assay selection. The immediate use of these modules are in algorithms that can be used to select optimal candidate genes and proteins for HTS assay development to be used in screening of environmental chemicals. Current efforts are focused on refinement of biological coherence and module identification measures.

*References available upon request

Building a scientific foundation for sound environmental decisions