Tox21 Phase III: The S1500 Genes High Throughput Transcriptomics Project

Progress Report

Richard S. Paules, Ph.D.
Biomolecular Screening Branch, DNTP
National Institute of Environmental Health Sciences

NTP Board of Scientific Counselors Meeting
June 18, 2014

Outline

- *The Problem and Need*
- *Our Response*
- *Our Solution*
- *The Future*

The Problem and Need

The Need:

A thorough understanding and evaluation of the adverse effects on humans from exposures to chemicals in the environment in order to protect human health.

The Challenge:

Thousands of chemicals are in use for which there is little or insufficient safety or toxicological information to evaluate the risk of adverse effects on human health from exposures.

The Response:

- NTP 2004 *Vision and Roadmap for the 21st Century*
- NRC 2007 Report on *Toxicity Testing in the 21st Century*
- *Toxicology in the 21st Century* (“Tox21”) Partnership

The Problem and Need (cont.)

The Assumption:

- Global “Omic” (Whole System) approaches can link perturbations with alterations in biological processes that result in toxicity and / or disease.

The Hypothesis:

- Alterations in the **transcriptome** in cells and tissues of humans, as well as model organisms, following exposures can provide linkage between chemicals and human toxicity and / or disease outcomes.

The Need:

- A rapid and low-cost method to measure alterations in the transcriptome in large numbers.

Will it work?

Transcriptomic Compendia: Published studies support transcriptomic linkage of chemical and genetic perturbations with adverse effects or diseases.

Burczynski, ME, *et al*., (2000) Toxicological Sciences. 58(2):399-415. *(HepG2; 162 cites)*
Hughes, TR, *et al*., (2000) Cell. 102(1):109-26. *(Yeast Compendium;* ***1,541 cites****)*
Waring, JF, *et al.,* (2001) Toxicology and Applied Pharmacology. 175(1):28-42 . *(Rat Liver; 238 cites)*
Waring, JF, *et al.,* (2001) Toxicology Letters. 120(1-3):359-368. *(Rat Hepatocytes; 225 cites)*
Hamadeh, HK, *et al*., (2002) Toxicological Sciences. 67(2):232-40. *(Rat Liver; 175 cites)*
Hamadeh, HK, *et al*., (2002) Toxicological Sciences. 67(2):219-31. *(Rat Liver; 276 cites)*
Steiner, G, *et al*., (2004) Environmental Health Perspectives. 112(12):1236-48. *(Rat Liver; 75 cites)*
Heinloth, AN, *et al*., (2004) Toxicological Sciences. 80(1):193-202. *(Rat Liver; 117 cites)*
Ellinger-Ziegelbauer, H, *et al.,* (2005) Mutation Research. 575(1-2):61-84. *(Rat Liver; 115 cites)*

Gene Logic
*** Iconix (Entelos) (DrugMatrix)**

*** TG-GATES** (Japanese Consortia “Toxicogenomics Project – Genomics Assisted Toxicity Evaluation System”

The Problem and Need (cont.)

The Ideal Solution:

A rapid and low-cost High Throughput (HT) method to measure expression levels of ALL GENES for use with:

- multiple cells lines and tissues
- multiple species
- exposures to thousands of perturbagens/chemicals
- multiple exposure levels (dose responses, benchmark doses, point of departure, lowest effect levels, etc.)
- exposures for varying lengths of time (kinetics, etc.).

The Assumption:

- At this time, whole transcriptome technologies are prohibitively expensive for HT applications.
- It will be necessary to focus on a subset of genes to use in a rapid, low-cost technology suitable for HT studies.

Our Response

NIEHS Federal Register notice on July 29, 2013 requesting the nomination and prioritization of environmentally responsive genes for use in screening large numbers of substances using toxicogenomic technologies.

Workshop sponsored by DNTP & DERT of NIEHS with the following goals to:

- **Address the need for identifying environmentally responsive genes** in **humans**, **rats**, **mice**, **zebrafish**, and ***C. elegans*** for use in toxicological studies of large numbers of substances.

- **Address approaches for prioritization of genes** for each species that provide maximal toxicogenomic information concerning both
  1) general responses, independent of cell type and
  2) responses that are specific to an organ or cell type.

- **Discuss criteria for prioritizing genes** in order to identify those potentially most useful in a screening paradigm.

- **Discuss potential uses** such as in **biomarker development** and in **basic research efforts.**

"High Throughput Transcriptomics Workshop: Gene Prioritization Criteria"

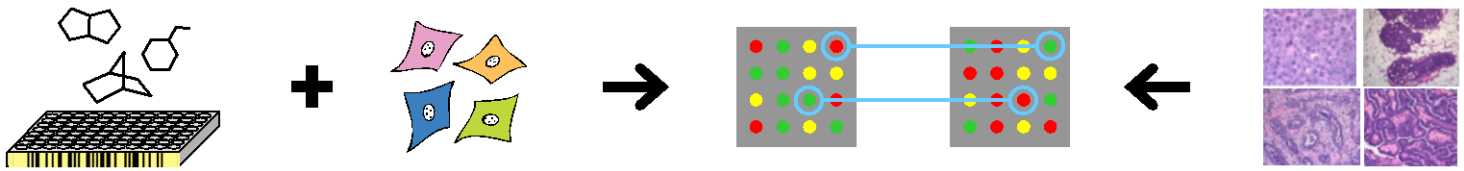
September 16-17, 2013; NIEHS

12 Invited Presentations Followed by Discussions on:

- **What is the best approach:**
  Data Driven (L-1000) vs. Knowledge-Based Selection or a Hybrid
- **Pathway-centric or Agnostic Gene Selection?**
- **What are Disease-Centric, Chemical-Responsive, Toxicology-centric genes?**
  17 Nominated Gene Sets submitted in response to Federal Register Notice published July 29th, 2013
- **Which are cells of most interest?**
  - ✓ Liver – metabolic activation, detoxification
  - ✓ Metabolically competent hepatocyte, renal proximal tubule, lung epithelium (Clara), intestinal epithelium, etc.; What about cardiac, neuro, muscle?
  - ✓ Stem cells; iPSC, ESC
  - ✓ Differentiated vs. dividing cells
  - ✓ Primary vs. immortalized (transformed) cells

Expression-Based Connectivity for Screening

Todd Golub, MIT, & Justin Lamb, GENOMETRY



chemical library **cultured cells** **transcriptional changes** **disease tissues**

"GE-HTS"

- assay expression of small number of genes
- conventional cell-based screening format

Stegmaier *et al*, Nature Genetics (2004); Peck *et al*, Genome Biology (2006); Hieronymus *et al*, Cancer Cell (2006)

"Connectivity Map"

- whole-genome expression-profile database
- universal *in silico* screen

Lamb *et al*, Science (2006); Wei *et al*, Cancer Cell (2006); Lamb, Nature Reviews Cancer (2007)

Landmark Genes – the L1000

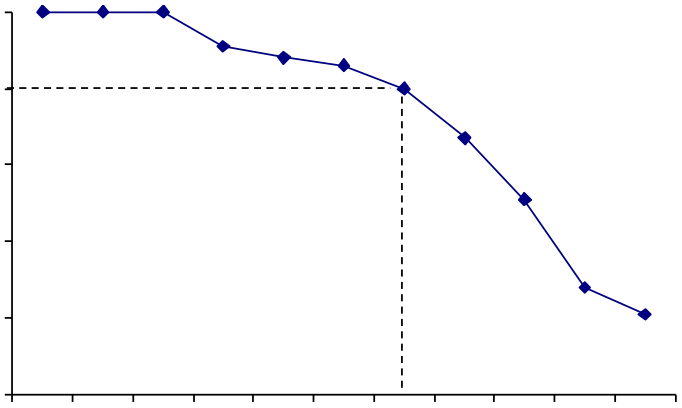
Justin Lamb, GENOMETRY

- **dimensionality of transcriptome ≪ total number of transcripts**
  - **co-expression highly prevalent**
- **learn the relationship between expression of all transcripts**
  - **large, diverse, high-quality, microarray-derived training set**
- **explicitly measure *landmark* genes, infer the remainder**
  - **multiple context-specific dependency matrices possible**

Landmark Genes – Is 1000 Enough?

Justin Lamb, GENOMETRY

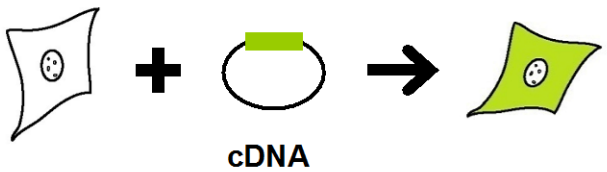
ation) performance at each level



Library of Integrated Network-based Cellular Signatures (LINCS) - NIH Common Fund Project

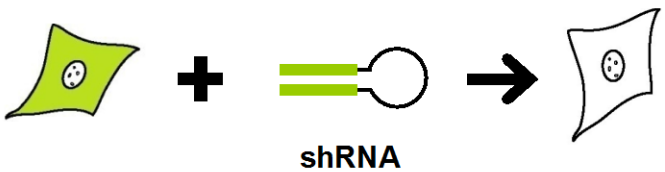
L1000 Data *– Over 1.4 M Gene-Expression Signatures*

ectopic expression



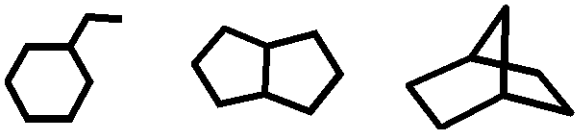
3,000× ORFs

RNAi-mediated ablation



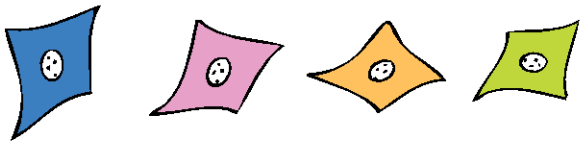
9,000× hairpins (3 per transcript)

small molecules



4,000× tool compounds
8-point dose ranges

cellular contexts



8× cancer cell lines
6× primary cell cultures
4× post-mitotic cell systems
2× hTERT-immortalized lines

Our Response (cont.)

“S1500 Genes” Selection Project Workgroup Members:

- **Scott Auerbach**, Biomolecular Screening Branch, DNTP, NIEHS
- **Pierre Bushe**l, Biostatistics Branch, DIR, NIEHS
- **Jennifer Collins**, Exposure, Response & Technology Branch, DERT, NIEHS
- **Agnes Forgacs**, National Center for Computational Toxicology, US EPA
- **David Gerhold**, Genomic Toxicology Group, National Center for Advancing Translational Sciences (NCATS)
- **Richard Judson**, National Center for Computational Toxicology, US EPA
- **Elizabeth Maull**, Biomolecular Screening Branch, DNTP, NIEHS
- **Deepak Mav**, Social & Scientific Systems, Inc.
- **Alex Merrick**, Biomolecular Screening Branch, DNTP, NIEHS
- **Rick Paules**, Biomolecular Screening Branch, DNTP, NIEHS
- **Ruchir Shah**, Social & Scientific Systems, Inc.
- **Dan Svoboda**, Social & Scientific Systems, Inc.

- **Donna Mendrick**, National Center for Toxicological Research, US FDA
- **Rusty Thomas**, National Center for Computational Toxicology, US EPA

Our Response (cont.)

"S1500 Genes" Selection Project Workgroup Deliberations

Solution 1: Use the L1000 Platform, working with GENOMETRY

Advantages:

- Well characterized Luminex-based assay ready to use now
- Being utilized by members of the LINCS project
- Huge amounts of human gene expression data

Disadvantages:

- *A*vailable only for human
- Costs are still higher than what is needed for true HTS
- Bioinformatics approaches are not published and thus the performance is not transparent.

Solution 2: Develop a Tox21 "Sentinel" Gene Set and HT Assay

"S1500 Genes" Selection Project Workgroup Priorities

- Pursue **Hybrid** Data Driven and Knowledge-Based Selection Approach
- Focus efforts towards **Human** gene set first
  - Provides linkage with Tox21 HTS efforts focusing on human health
- Develop robust bioinformatic modules to provide a gene set that:
  - Maximizes biological **Diversity** (Diversity Importance Score (i_D))
  - Maximizes **Co-Expression** information (Co-Expression Importance Score (i_C))
  - Optimizes **Pathway Coverage**
  - Captures some if not all nominated Disease-Centric, Chemical-Responsive, Toxicology-Centric genes, as well as **L1000 genes**
  - Extrapolates from subset to full transcriptome (**"Extrapolatability"**)
- Use robust **rat** toxicogenomics data sets to develop bioinformatic approach and then apply approach to human Affymetrix data in GEO
  - **Train** with TG-GATES rat liver data sets
  - **Test** performance with the independent DrugMatrix rat data sets

Our Response (cont.)

Training Data Set

TG-GATES

Rat Affymetrix GeneChip Arrays

- **Strain**: Sprague Dawley (male)
- **Tissue:** Liver
- **Chemicals**: 131
- **Dose levels**: 3 plus vehicle control
- **Study Duration**:
  - Single dose: 3, 6, 9, 24 hrs
  - Repeat dose: 3, 7, 14, 28 days
- **Biological replicates**: 3
- **Experiments**: 3127

Test Data Set

DrugMatrix

Rat Affymetrix GeneChip Arrays

- **Strain**: Sprague Dawley (male)
- **Tissues:** Cell Cultures, Heart, Kidney, Liver, Thigh Muscle
- **Chemicals**: 376
- **Dose levels**: 1 or 2 plus vehicle control
- **Study Duration**:
  - Single dose: 6, 24 hrs
  - Repeat dose: 3, 4, 5, 7 days
- **Biological replicates**: 3
- **Experiments**: 1540

The S1500 gene set should have the following attributes:

1. **Diversity:** Capture the maximal expression variability and dynamics.

2. **Co-Expression:** Capture the **Sentinel** genes with maximal co-expression information to represent members of nodes or networks.

1. **Maximal Pathway Coverage:** Genes are included to ensure maximal biological pathway coverage.

2. **Inclusion of toxicity and disease related genes:** Specific genes will be selected for their reported roles in toxicity-related and disease-related processes.

3. **Capture the L1000 gene set as a component of S1500 genes.** Facilitate linkage with LINCS data as much as possible.

4. **"Extrapolatability":** This property refers to the ability to extrapolate or **infer** or **impute** with some accuracy the expression changes in all genes from those observed in this reduced set of sentinel genes.

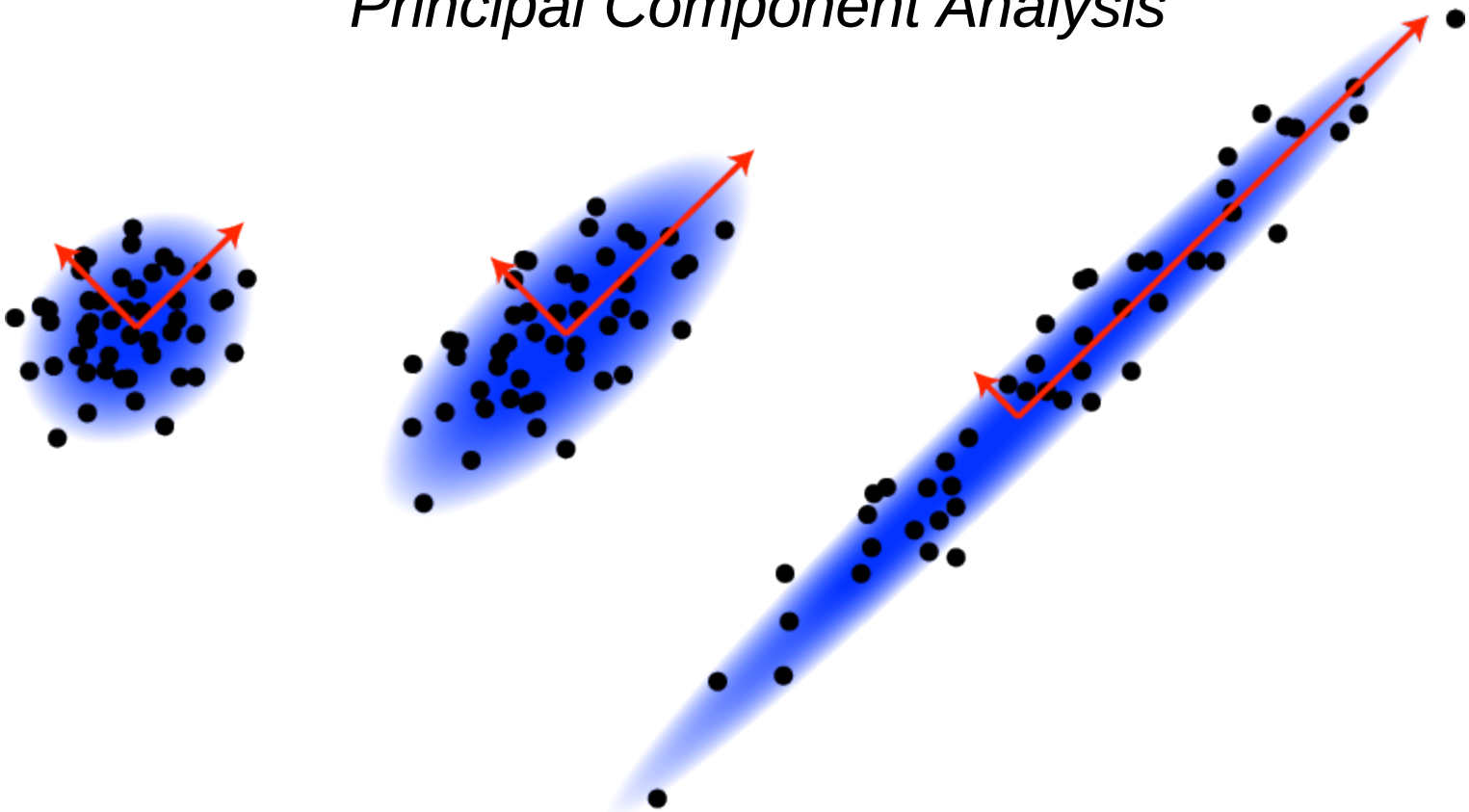
Our Solution

The S1500 gene set should have the following attributes:

1. **Diversity:** Capture maximal expression variability and dynamics:

Diversity Importance Score (i_D)

Principal Component Analysis



How Do They Do That? – An Evolving Approach

Ruchir Shah, Deepak Mav, Richard Judson, Scott Auerbach, Pierre Bushel

- For each partition (i),
  - Perform Principal Component Analysis using probeset level log2 fold change values
  - Compute partition specific probeset importance score (z_{ip}) using weighted average of squared PC loading vectors of first K PCs that collectively retain 90% of total variability as following

$$z_{ip} = \sum_{j=1}^{K} \alpha_j L_{pj}^2; \quad \alpha_j = \frac{\lambda_j}{\sum_{m=1}^{K} \lambda_m};$$

where λ_j denotes variability retained by j^{th} PC.

- Aggregate importance scores across partitions using Tukey (weighted) mean and rank probesets accordingly.
- Select probe sets with ranks <= 500

(note: 500 is arbitrarily selected as a place holder, we can replace it with a number we decide as a group)

Our Solution

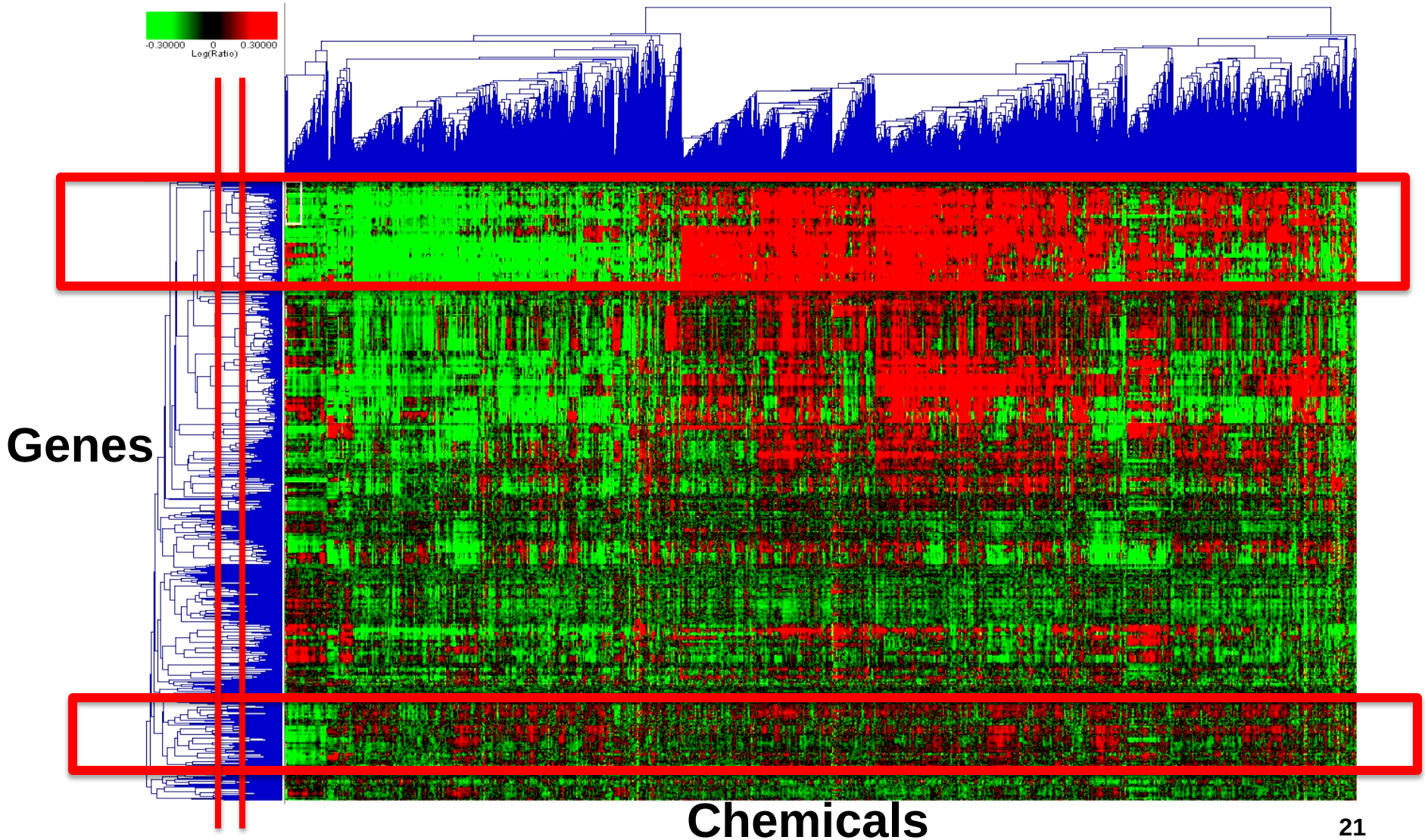
The S1500 gene set should have the following attributes:

1. **Diversity:** Capture the maximal expression variability and dynamics.

2. **Co-Expression:** Capture the **Sentinel** genes with maximal co-expression information to represent members of nodes or networks.

Our Solution
Co-Expression Importance Score (i_C)

Gene Modules Identified by Pruning of Unsupervised Clustering Dendrogram
(Spearman's Correlation Coefficient + Ward's Linkage)

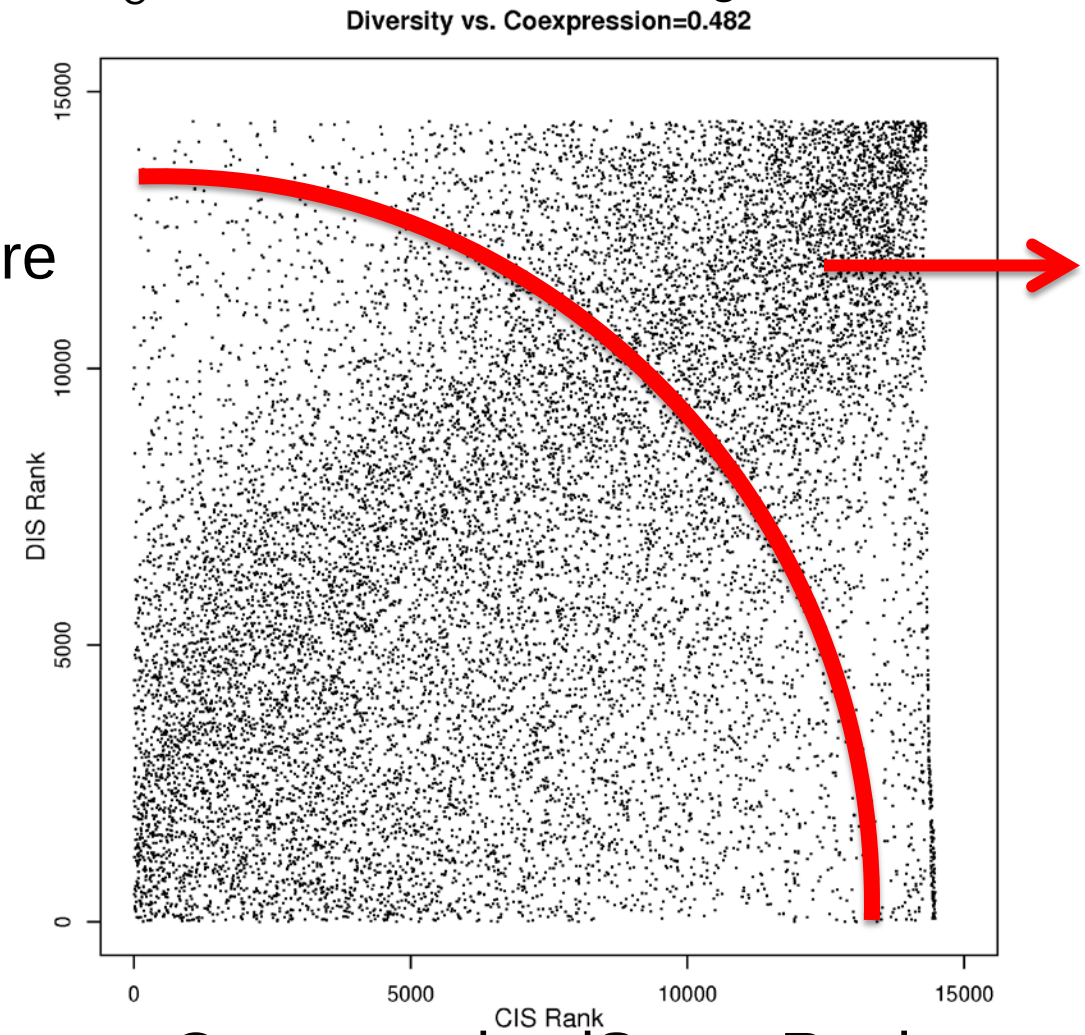


Our Solution

Aggregate Diversity and Coexpression Importance (i) Scores

(Square the i_D and i_C values, Sum and Average for each Gene, then Rank)

Diversity iScore Rank



1500 Genes

Coexpression iScore Rank

Our Solution

The S1500 gene set should have the following attributes:

1. **Diversity:** Capture the maximal expression variability and dynamics.

2. **Co-Expression:** Capture the **Sentinel** genes with maximal co-expression information to represent members of nodes or networks.

3. **Maximal Pathway Coverage:** Genes are included to ensure maximal biological pathway coverage.

Our Solution

Pathway Coverage Optimization

(Broad GSEA Molecular Signature DataBase [MSigDB])

Curated Pathway / Gene Set	Number of Pathways	Pathways Covered (>= 3 genes)	Mean Pathway Multiplicity per Gene
Affymetrix GeneChip Platform Coverage			
Canonical Pathways (CP)	243	243	1.1
CP-BioCarta	217	217	1.4
CP-KEGG	186	186	1.8
CP-Reactome	673	673	3.0
Data Driven Initial S1500 Coverage			
Canonical Pathways (CP)	243	135	1.3
CP-BioCarta	217	75	1.5
CP-KEGG	186	137	1.9
CP-Reactome	673	345	3.3
Pathway Optimized S1500 Coverage			
Canonical Pathways (CP)	243	243	1.8
CP-BioCarta	217	217	2.1
CP-KEGG	186	186	2.5
CP-Reactome	673	673	4.5

Our Solution

Tox21 Pathway Tool – BioPlanet

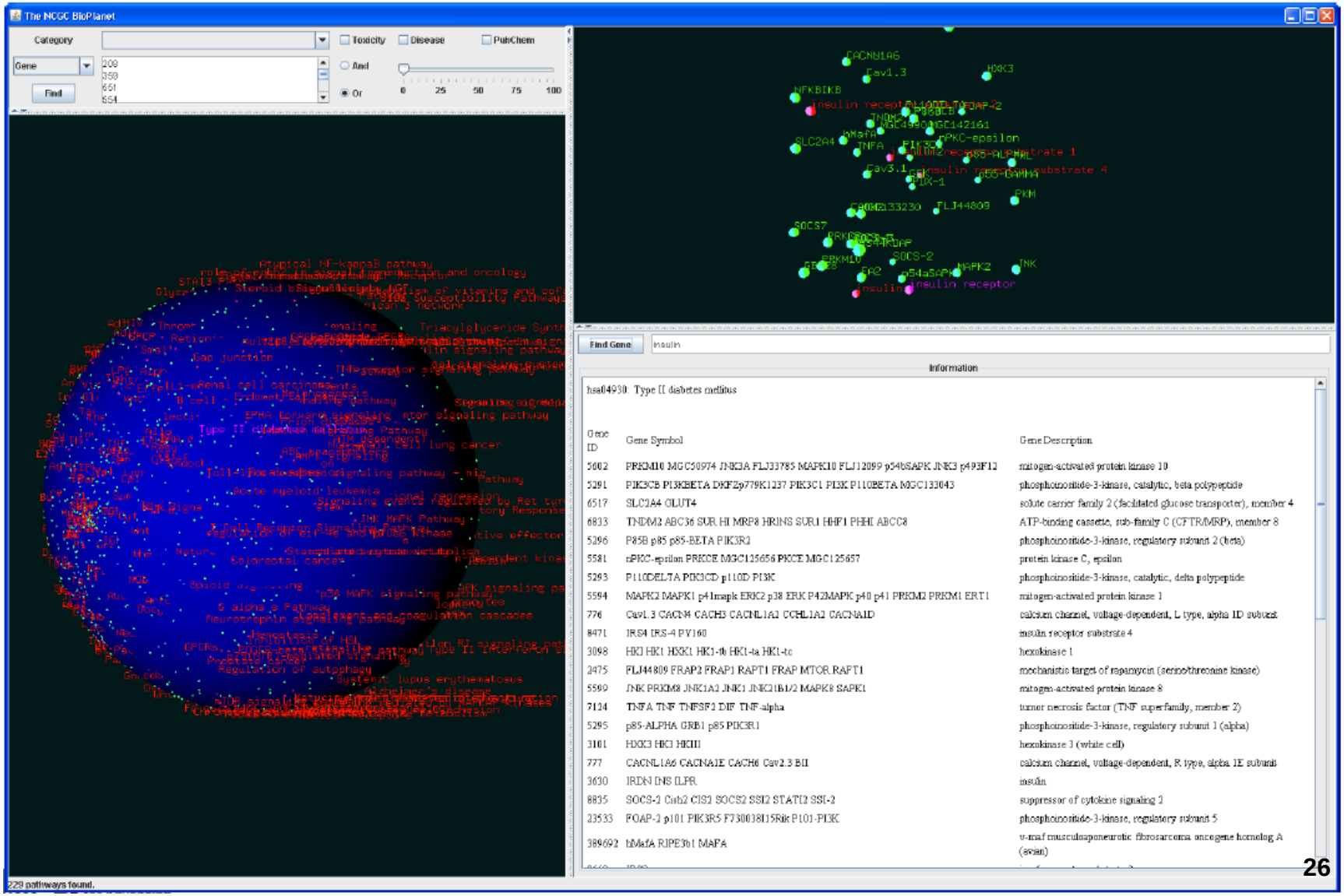
Developed by Ruili Huang at NCATS

- Goal – To host the universe of well-documented pathways

  Focus on Human pathways (> 2000 unique)
- All pathway annotations are from manually curated, public source

  e.g. KEGG, WikiPathways, Reactome, Science Signaling, etc.
- Integrates pathways from > 10 different data sources
- Annotates pathways by source, species, biological function, processes, disease/toxicity, assay, etc.
- Easy visualization for browsing and analysis of pathways
- Facilitates assay interpretation and prioritization of future assays for Tox21
- Web version in development for public release

Our Solution
Pathways containing diabetes genes



Our Solution

The S1500 gene set should have the following attributes:

1. **Diversity:** Capture the maximal expression variability and dynamics.

2. **Co-Expression:** Capture the **Sentinel** genes with maximal co-expression information to represent members of nodes or networks.

3. **Maximal Pathway Coverage:** Genes are included to ensure maximal biological pathway coverage.

4. **Inclusion of toxicity and disease related genes:** Specific genes will be selected for their reported roles in toxicity-related and disease-related processes.

In Progress

Performance Check: “Extrapolatability” or Imputation

Pathway Level Reproducibility

TG-GATES Rat Data

Complete GeneChip Data

↓

Derive S1500 Gene Set → **Extract S1500 Gene Set Values from DrugMatrix Data Sets**

↑

DrugMatrix Rat Data

↔

Imputed GeneChip Data | **Actual GeneChip Data**

"Extrapolatability" or Imputation Performance

Pathway Level Reproducibility

Curated Pathway / Gene Set	Number of Pathways	Concordance Rate	Enrichment Score (Pearson Correlation)

Data Driven Initial S1500 Gene Set

Canonical Pathways (CP)	243	86%	0.70
CP-BioCarta	217	77%	0.66
CP-KEGG	186	87%	0.71
CP-Reactome	673	82%	0.72

Pathway Optimized S1500 Gene Set

Canonical Pathways (CP)	243	86%	0.75
CP-BioCarta	217	79%	0.75
CP-KEGG	186	87%	0.75
CP-Reactome	673	83%	0.77

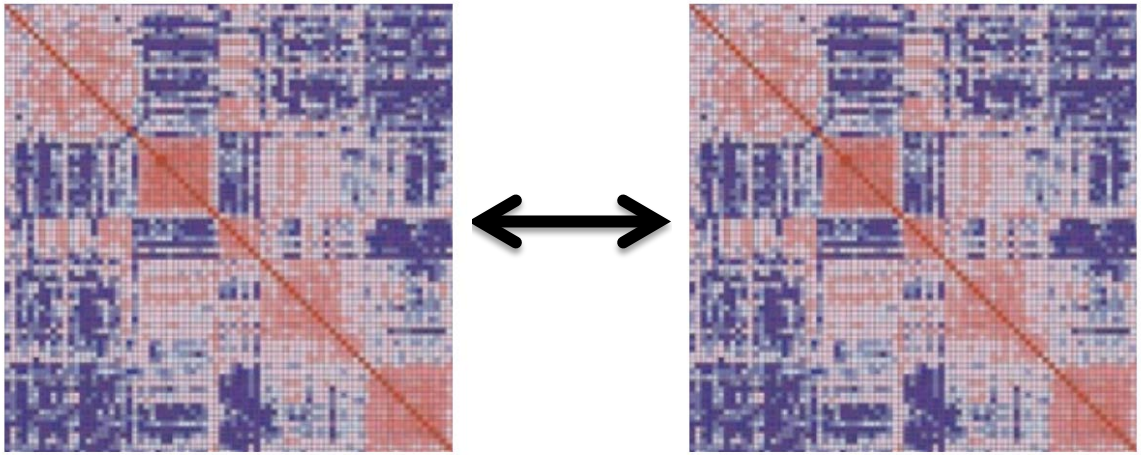
Affymetrix Pathway Enrichment Value	Imputed Pathway Enrichment Value	Concordance	Correlation
0.9	0.85	1	0.7
0.2	0.4	0	
- 0.9	- 0.7	1	
- 0.5	0.6	0	

"Extrapolatability" or Imputation Performance

Chemical Response Level Reproducibility

TG-GATES Rat Data

DrugMatrix Rat Data



Complete GeneChip Data

Impute Chemical Responses

Derive S1500 Gene Set → **Extract S1500 Gene Set Values from DrugMatrix DataSets**

Paules_2014

Summary

Rat

Bioinformatic modules have been developed that appear to be performing as desired

Optimization - Continuing bioinformatic modifications to improve performance

Evaluation - Connectivity with Chemicals and Adverse Endpoints

Human

Identify highest quality Affymetrix Human datasets in GEO

Build Ratios in those datasets in order to work in ratio space – minimize batch effects

Apply bioinformatic approach to human GEO Data

Perform Pilot test experiments to evaluate performance

Other

Extend to other species (rat, mouse, zebrafish, *C. elegans*, etc.)

Evaluate Technological Advancements (Gerhold/NCATS Lead) –

Existing Options – Luminex Beads, RASL-Seq, Illumina NextGen Seq

Advances may eliminate the need to focus on a subset of the transcriptome and thus eliminate the need to “Imput” values

Where Are We Going?

Short Term

Application of HT Transcriptomics to Human Cells that have already been used in Tox21 Phase II assays with all or a portion of the 10k set of chemicals

Application of HT Transcriptomics to Metabolically-Competent Human Cells (HepaRG, etc.)

Application of HT Transcriptomics to Human iPS and ES Cells undifferentiated and induced to differentiate along specific lineages

Mid Term

Develop similar Gene Sets and HT Transcriptomics Platforms for other species (rat, mouse, zebrafish, *C. elegans*, etc.)

Application of HT Transcriptomics to NTP archived material from rat and mouse studies and Tox21 Phase III alternative species studies

Longer Term

Application of HT Transcriptomics to Human samples from molecular epidemiological studies and clinical studies

Acknowledgements

"S1500 Genes" Selection Project Workgroup Members:

- **Scott Auerbach**, Biomolecular Screening Branch, DNTP, NIEHS
- **Pierre Bushe**l, Biostatistics Branch, DIR, NIEHS
- **Jennifer Collins**, Exposure, Response & Technology Branch, DERT, NIEHS
- **Agnes Forgacs**, National Center for Computational Toxicology, US EPA
- **David Gerhold**, Genomic Toxicology Group, National Center for Advancing Translational Sciences (NCATS)
- **Richard Judson**, National Center for Computational Toxicology, US EPA
- **Elizabeth Maull**, Biomolecular Screening Branch, DNTP, NIEHS
- **Deepak Mav**, Social & Scientific Systems, Inc.
- **Alex Merrick**, Biomolecular Screening Branch, DNTP, NIEHS
- **Rick Paules**, Biomolecular Screening Branch, DNTP, NIEHS
- **Ruchir Shah**, Social & Scientific Systems, Inc.
- **Dan Svoboda**, Social & Scientific Systems, Inc.

- **Donna Mendrick**, National Center for Toxicological Research, US FDA
- **Rusty Thomas**, National Center for Computational Toxicology, US EPA