

---

The *Carolina* Environmental  
Bioinformatics Research Center:  
What's Ahead in the Information  
Series

**Fred Wright, Ph.D.**

University of North Carolina at Chapel Hill



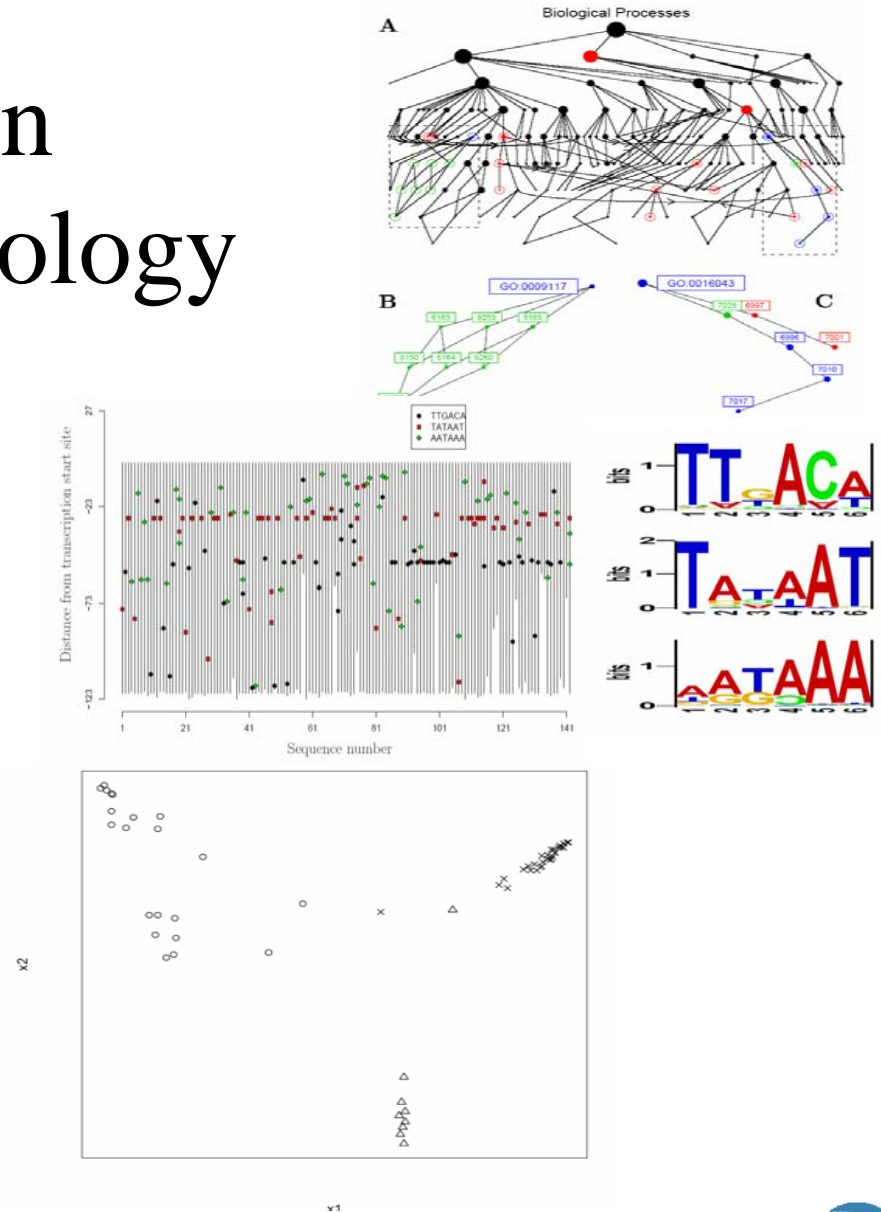
# Organization of Center

- Three major Research Projects: (1) Biostatistics, (2) Cheminformatics, and (3) Computational Infrastructure for Systems Toxicology
- Administrative Unit
- Public Outreach and Training Activity (POTA)
- “Functional areas” of *Analysis, Methods* Development and *Tools* Development overseen by a panel of experienced investigators



# (1) Biostatistics in Computational Toxicology

- Emphasis on strengths in microarray analysis, elucidation of networks/pathways, Bayesian approaches
- Stresses existing capabilities



## (2) Chem-informatics

- seeks to establish a universally applicable and robust predictive toxicology modeling framework
- Focuses on Quantitative Structure Activity/Property Relationships (QSAR)
- Establishes a modeling workflow, toxicity prediction scheme and plan for software development

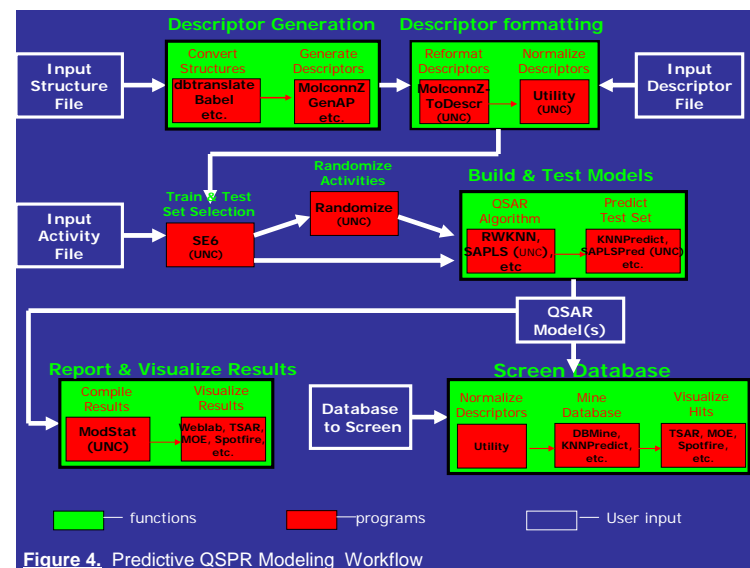


Figure 4. Predictive QSPR Modeling Workflow

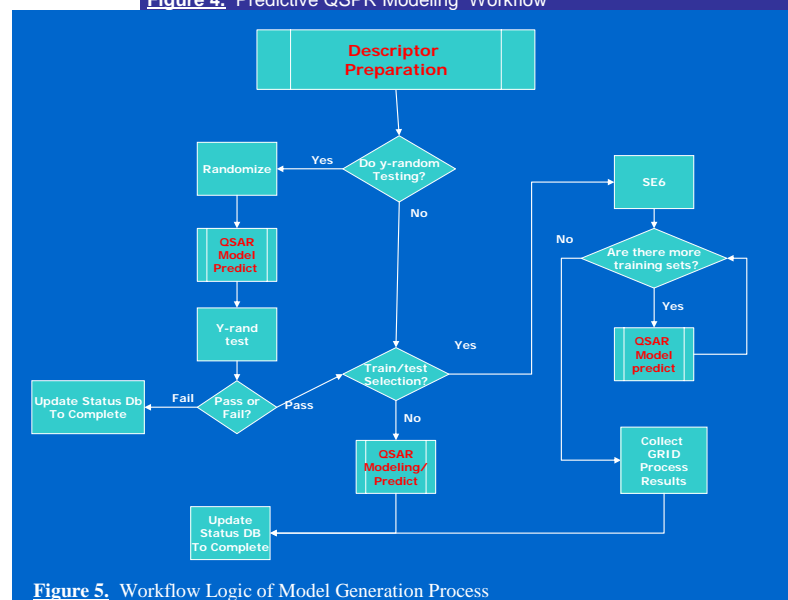
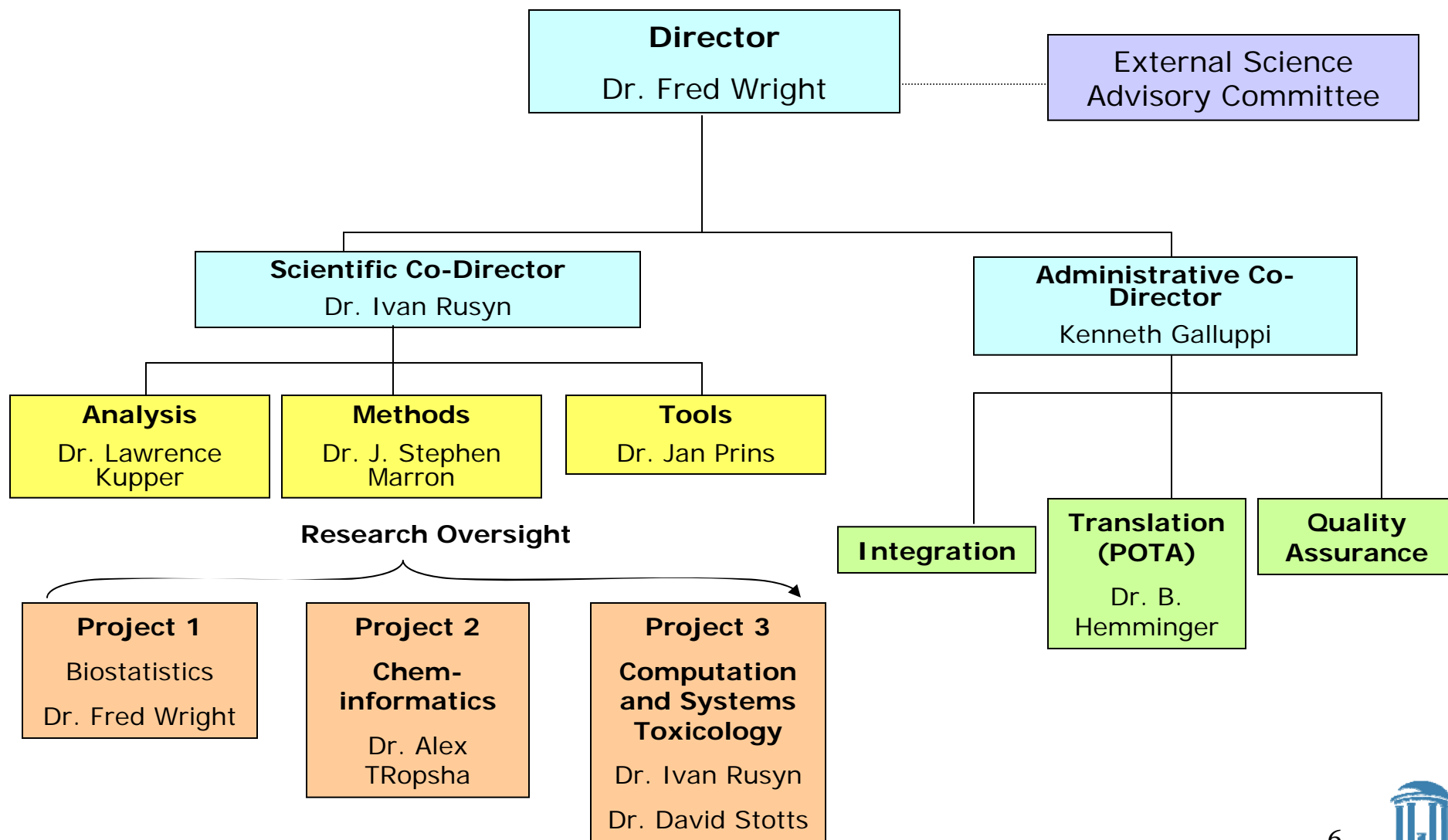


Figure 5. Workflow Logic of Model Generation Process





# CEBRC Center Organization



# Project 1

## Biostatistics in Computational Toxicology



- Fred Wright, Ph.D. (P.I.) – statistical genetics, genomic analysis
- Mayetri Gupta, Ph.D. – sequence analysis, motif detection
- Young Troung, Ph.D. – Bayesian network genetic analysis, SVM methods for metabolomic data
- Joseph Ibrahim, Ph.D. – Bayesian analysis of microarray data
- Danyu Lin, Ph.D. – haplotype-phenotype analysis, microarray analysis
- Fei Zou, Ph.D. – statistical genetics, genomic analysis
- Andrew Nobel, Ph.D. – clustering, data dimensional reduction, genetic pathway analysis
- Master's trained personnel



# Project 1 objectives

- to provide **analysis** capability to the environmental sciences community
- to develop appropriate new **methods** to apply to public data from the EPA and the broader community
- to develop computational **tools** to further the objectives.
- to **disseminate** research findings to the computational toxicology community, **train** students, and to **coordinate** additional statistical research in computational toxicology.





## Methods (to name a few)

- Sample size estimation for high-throughput data
- P-value computation, significance testing
- Multiple-testing issues, false discovery rates
- Dose-response modeling
- New measures of differential expression
- Transcriptional regulation and motif discovery
- Network analysis, discrimination methods
- Pathway analysis

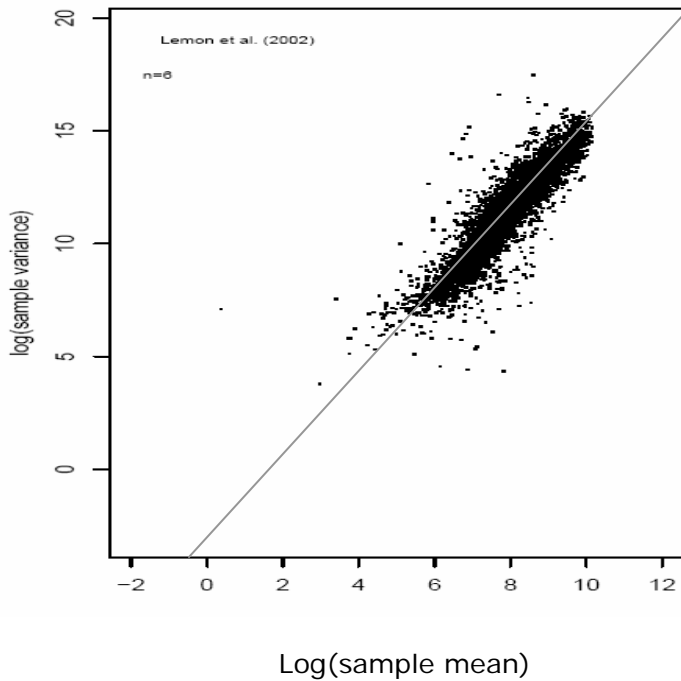


# Tools

- Much of initial code has been implemented in R/Bioconductor. This is directly useful to other statistical investigators.
- Work with project 3 investigators and students to produce user-friendly web-based and/or standalone applications
- Work to increase utility of methods by integration with informatics and biological annotation
- We view the SAM software as a model for independent successful dissemination. Project 3 personnel are training to implement appropriate procedures in ArrayTrack

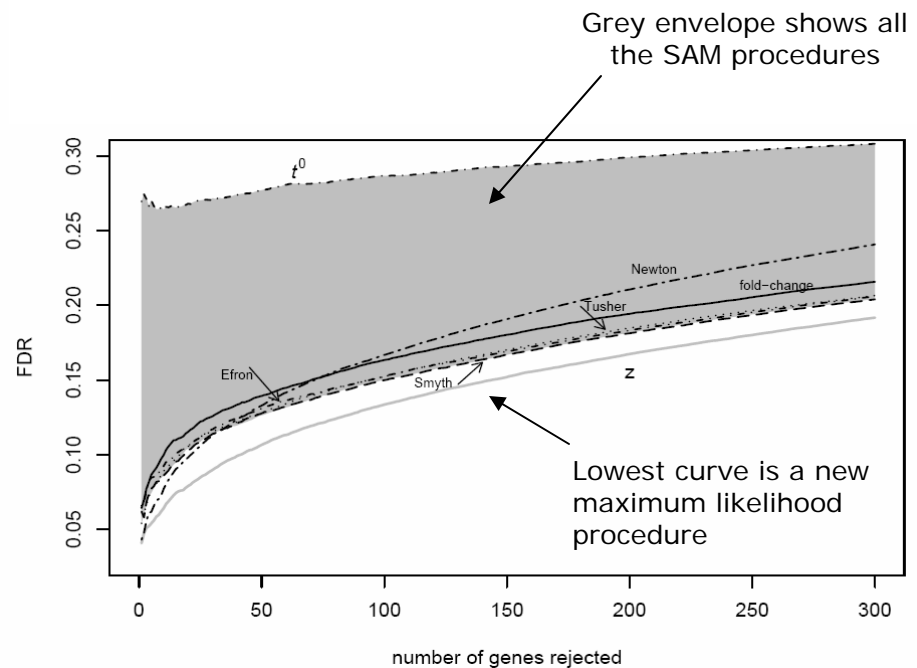


# Example 1. New ways of detecting differential expression

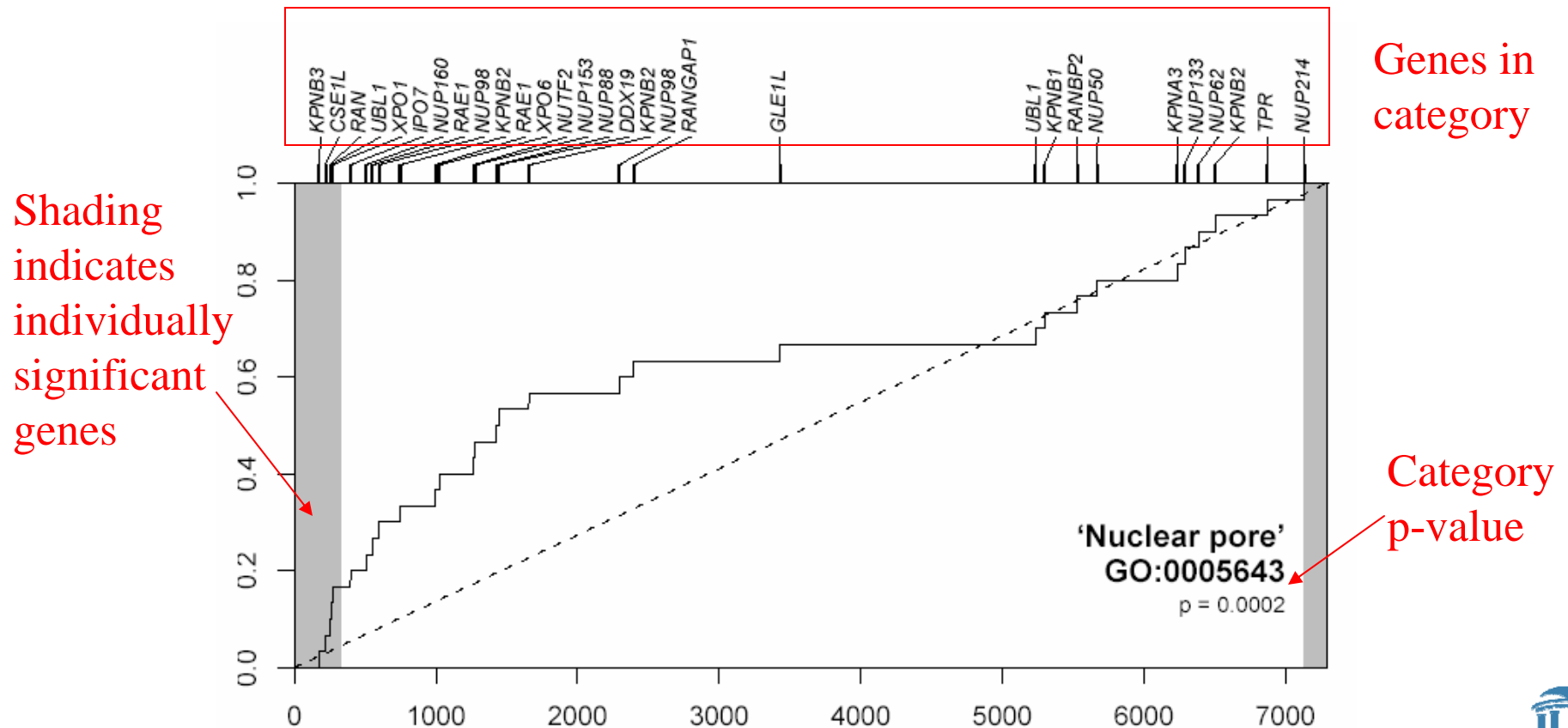


Expression measurements show a mean-variance relationship...

Which we can exploit to reduce the false discovery rate...

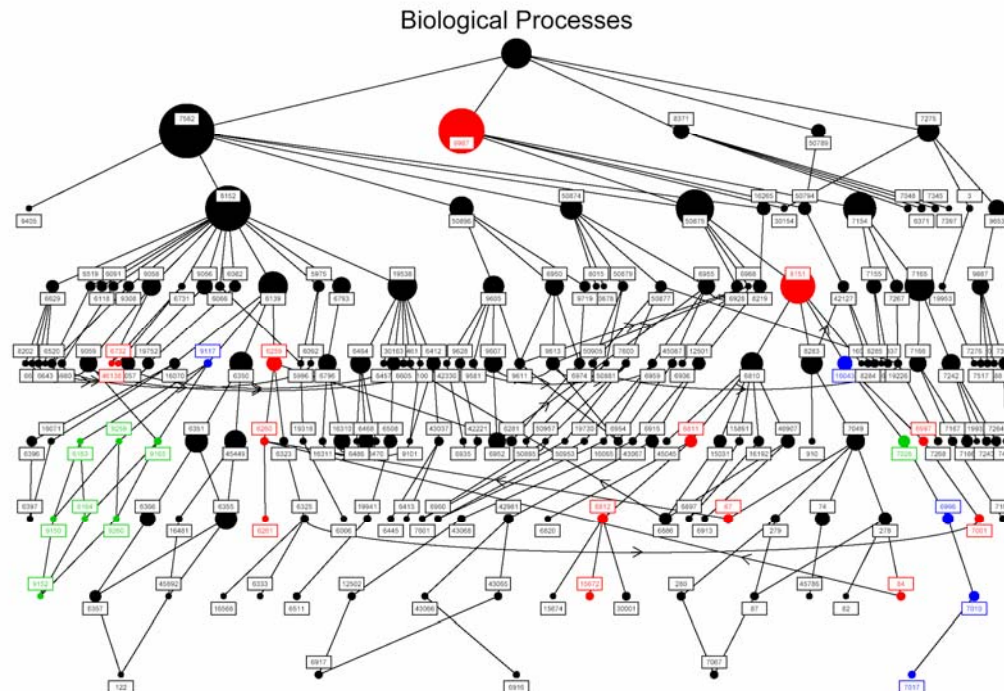


## Example 2. Significant genes/pathways/categories: the Significance Analysis of Function and Expression procedure (honest pathway significance testing)

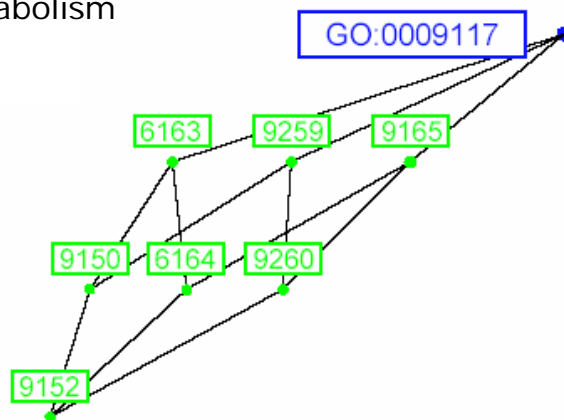


# GO Tree with significant nodes

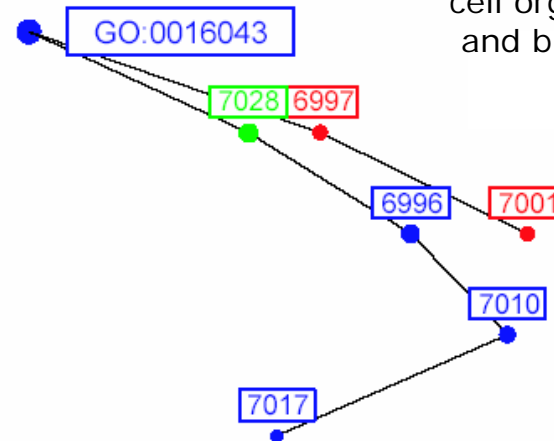
Key: blue ( $p < 0.001$ ) green ( $0.001 \leq p < 0.01$ ), red ( $0.01 \leq p < 0.1$ ).



nucleotide metabolism



cell organization and biogenesis



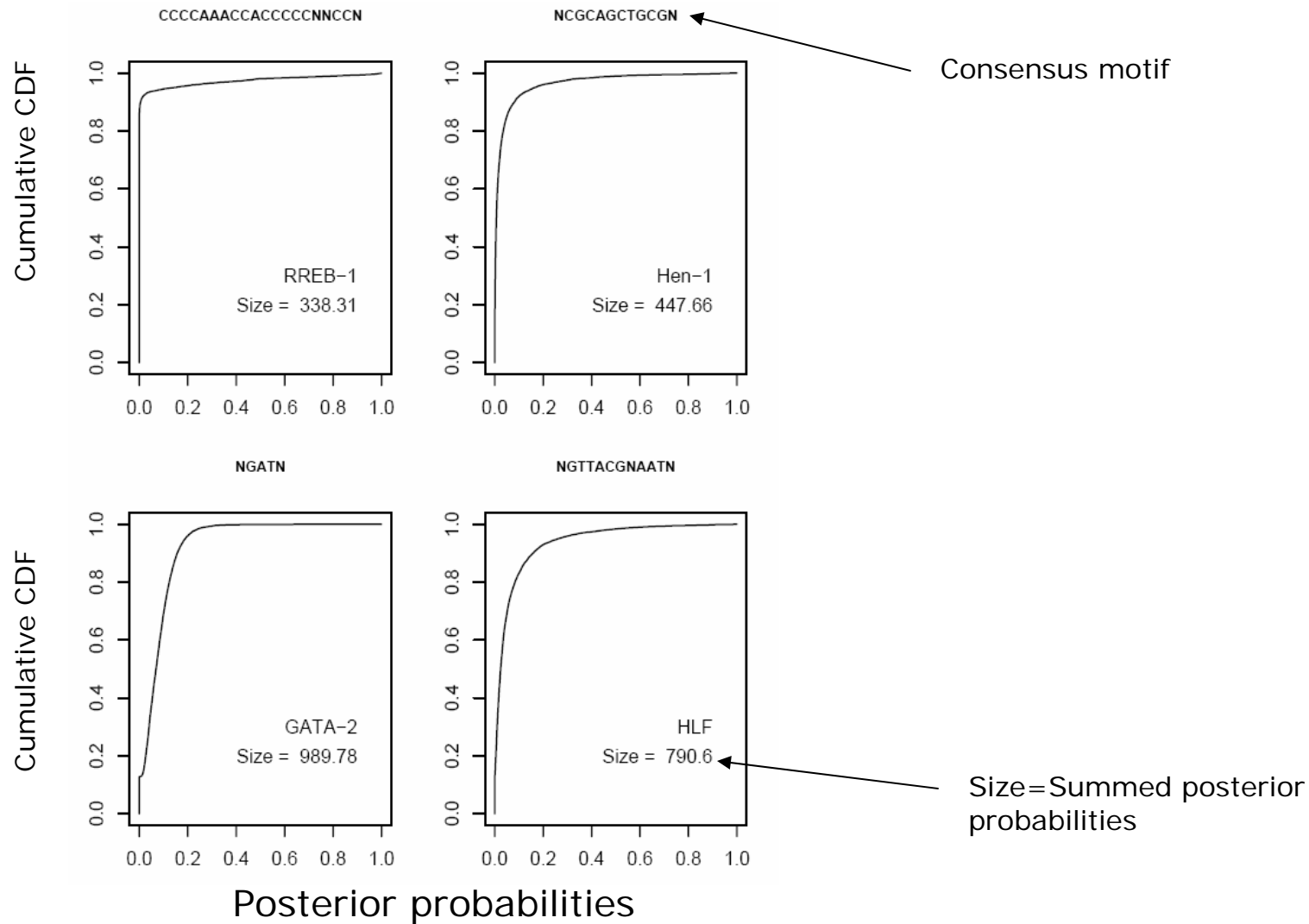
### Example 3. A Bayesian approach for finding probabilities of transcription factor binding sites

The complicated stuff (M. Gupta and colleagues)...

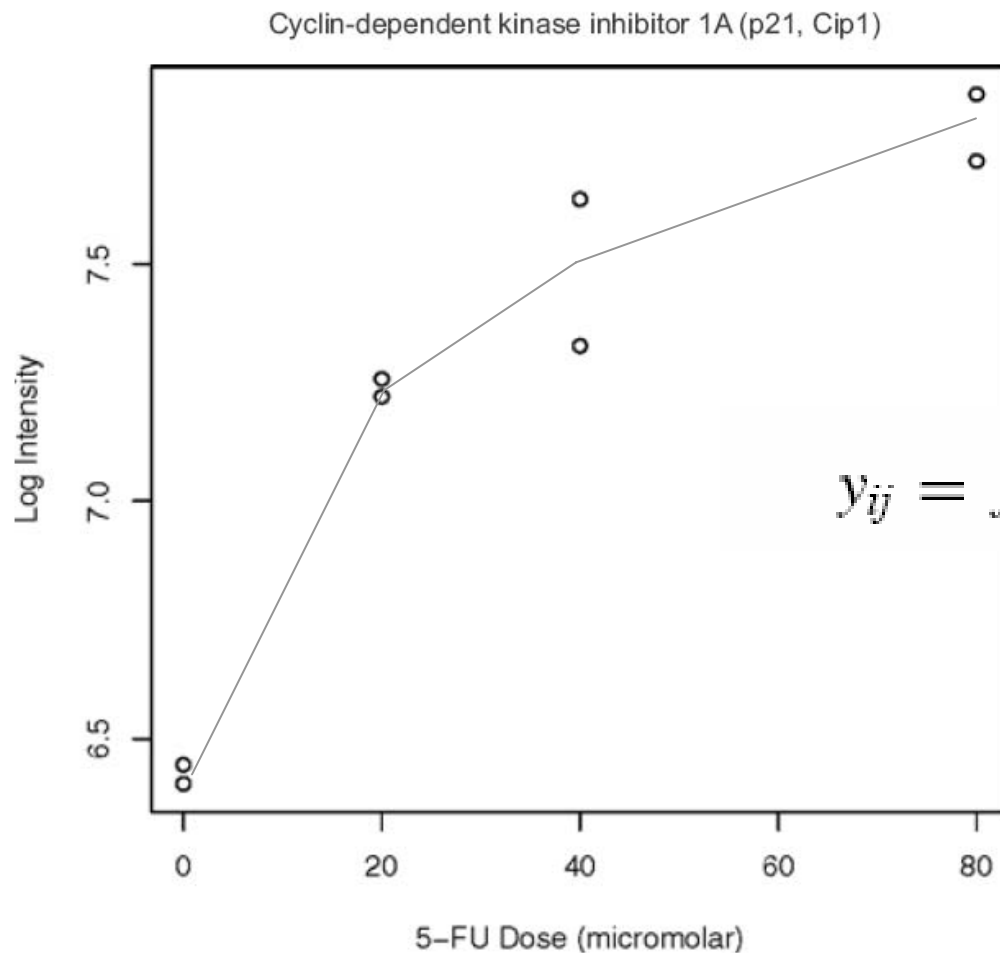
$$\begin{aligned} LR(\mathbf{x}) &= \frac{Pr(\mathbf{x}|H_{A,\text{single}})}{Pr(\mathbf{x}|H_0)} = \frac{\sum_{j=1}^{L-w+1} Pr(\mathbf{x} \cap \text{the motif starts at position } j)}{Pr(\mathbf{x}|H_0)} \\ &= \frac{\sum_{j=1}^{L-w+1} Pr(\mathbf{x}_{[1:j-1]}|\theta_0) \cdot Pr(\mathbf{x}_{[j:j+w-1]}|\Theta) \cdot Pr(\mathbf{x}_{[j+w:L]}|\theta_0) \cdot \beta \cdot (1-\beta)^{L-w}}{Pr(\mathbf{x}|\theta_0) \cdot (1-\beta)^L} \\ &= \frac{\beta}{(1-\beta)^w} \cdot \sum_{j=1}^{L-w+1} \frac{Pr(\mathbf{x}_{[j:j+w-1]}|\Theta)}{Pr(\mathbf{x}_{[j:j+w-1]}|\theta_0)} \end{aligned}$$



The simple results – each gene/transcript has a posterior probability of containing the motif. Here are results for the Affy U95A array, 5000bp upstream sequences, using the JASPAR database



## Example 4. Isotonic regression: gene expression dose-response data



$$y_{ij} = f(d_i) + \epsilon_{ij},$$

Model -  
 $f$  should be  
strictly  
increasing or  
decreasing

Hu et al., 2005, *Bioinformatics*  
21: 3524-3529).





## **Pyrethroid Biomarker Project (J. Harrill, K. Crofton and colleagues, U.S. E.P.A)**

- **Problem:** Lack of a cost efficient biomarker of effect hampers assessments of the cumulative risk of pyrethroid insecticides.
- **Aim:** Develop a biochemical biomarker of effect for pyrethroids that reflects changes in neuronal firing rates.
- **Methods:** Use gene arrays and RT-PCR to identify dose-responsive transcripts in rat CNS. Permethrin and deltamethrin each examined at four doses, Affymetrix arrays.

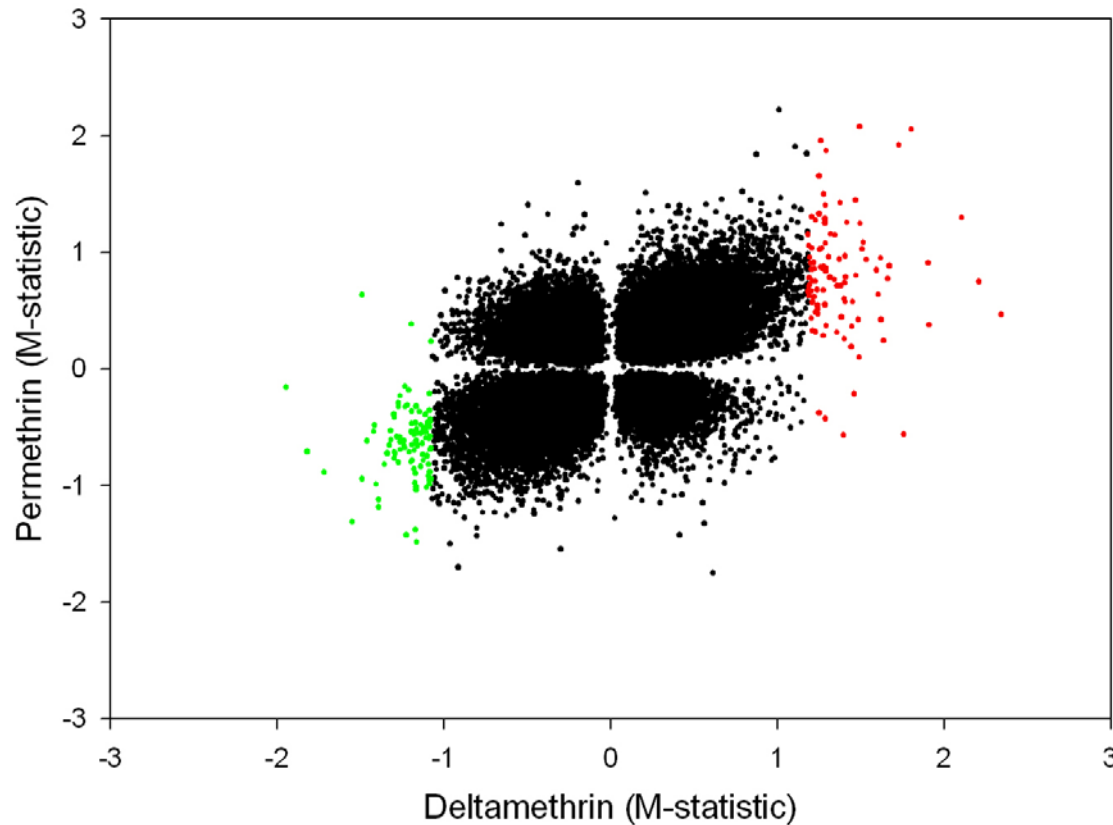


## Dose-response, cont.: a statistic to rank genes...

$$M = \frac{\hat{f}(dose_{highest}) - \hat{f}(dose_{lowest})}{\sqrt{v}}$$

Standard error estimate.  
Could be improved.

Comparison of M-statistics



Dose-response data  
on pyrethroid in rat  
brains, courtesy of  
J. Harrill and K.  
Crofton, U.S. E.P.A.



# Project 2

## Chem-informatics



- Alex Tropsha, Ph.D. (P.I.) –computational chemistry, QSAR
- Weifan Zheng, Ph.D. – computational methods in drug discovery, QSAR
- Alexander Golbraikh, Ph.D. – mathematical approaches in QSAR development
- Yufeng Liu, Ph.D. – Support vector machines, semi-supervised machine learning
- additional personnel



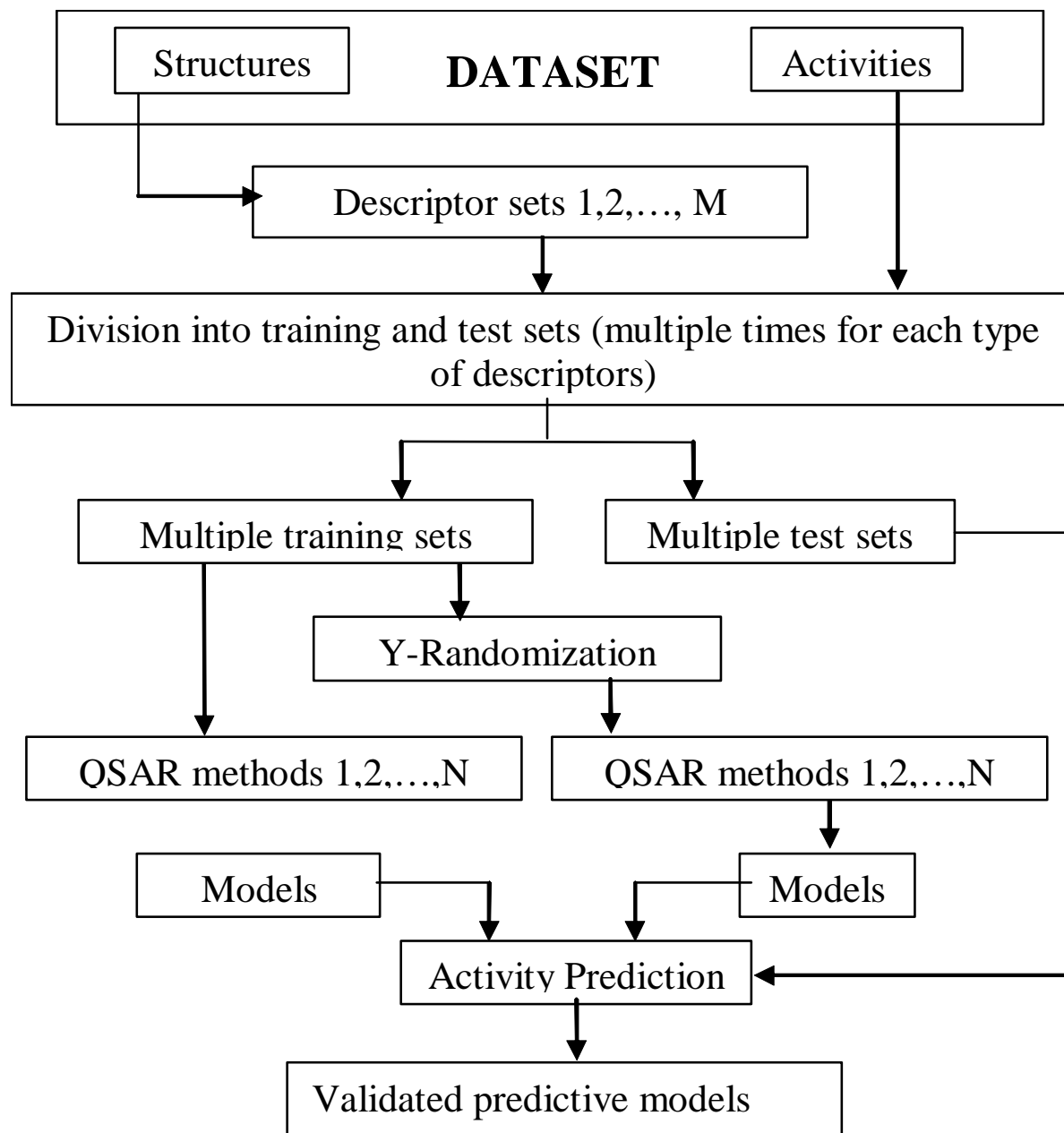
## Project 2 objectives

- to develop an innovative QSPR modeling workflow based on the principles of combinatorial QSPR modeling, model validation and consensus prediction
- to develop toxicity predictors using the workflow
- to integrate modeling tools and endpoint predictors using workflow design middleware and workflow deployment in a predictive toxicology web portal
- Applied to toxicology datasets



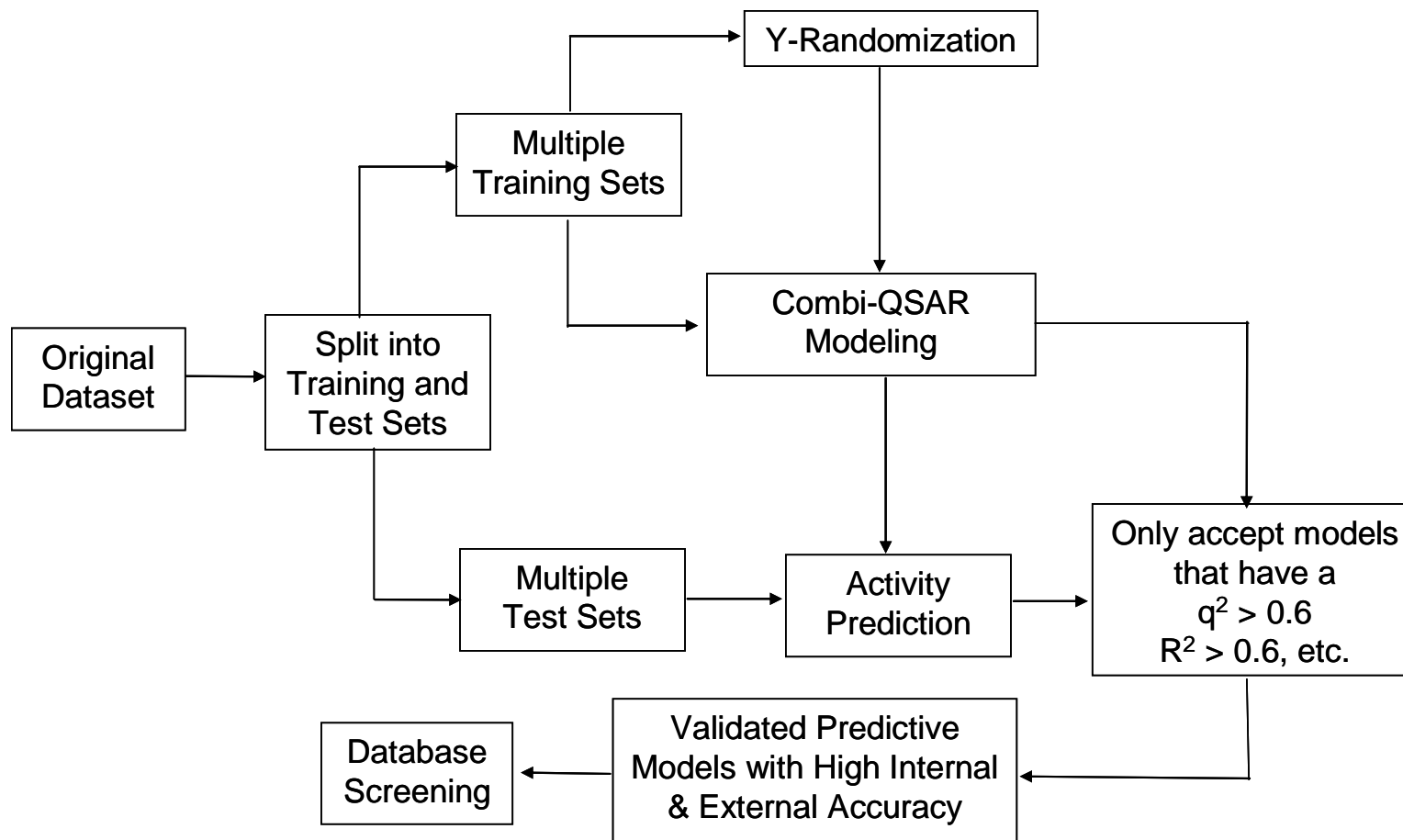
- Project 2 builds on years of research in the Tropsha lab on QSAR/QSPR modeling and developing robust predictors
- Many of the machine learning and cross-validation ideas are used in statistical genomics
- Descriptors – topological molecular indices, size and shape, hydrophilic/phobic indices, physical properties, etc.
- Try to predict biological activity
- Analysis of the Carcinogenic Potency Database (collaboration with Dr. A. Richard, EPA) was performed, applied to 693 compounds, with classification kNN QSAR prediction accuracies estimated at 85%-90%.





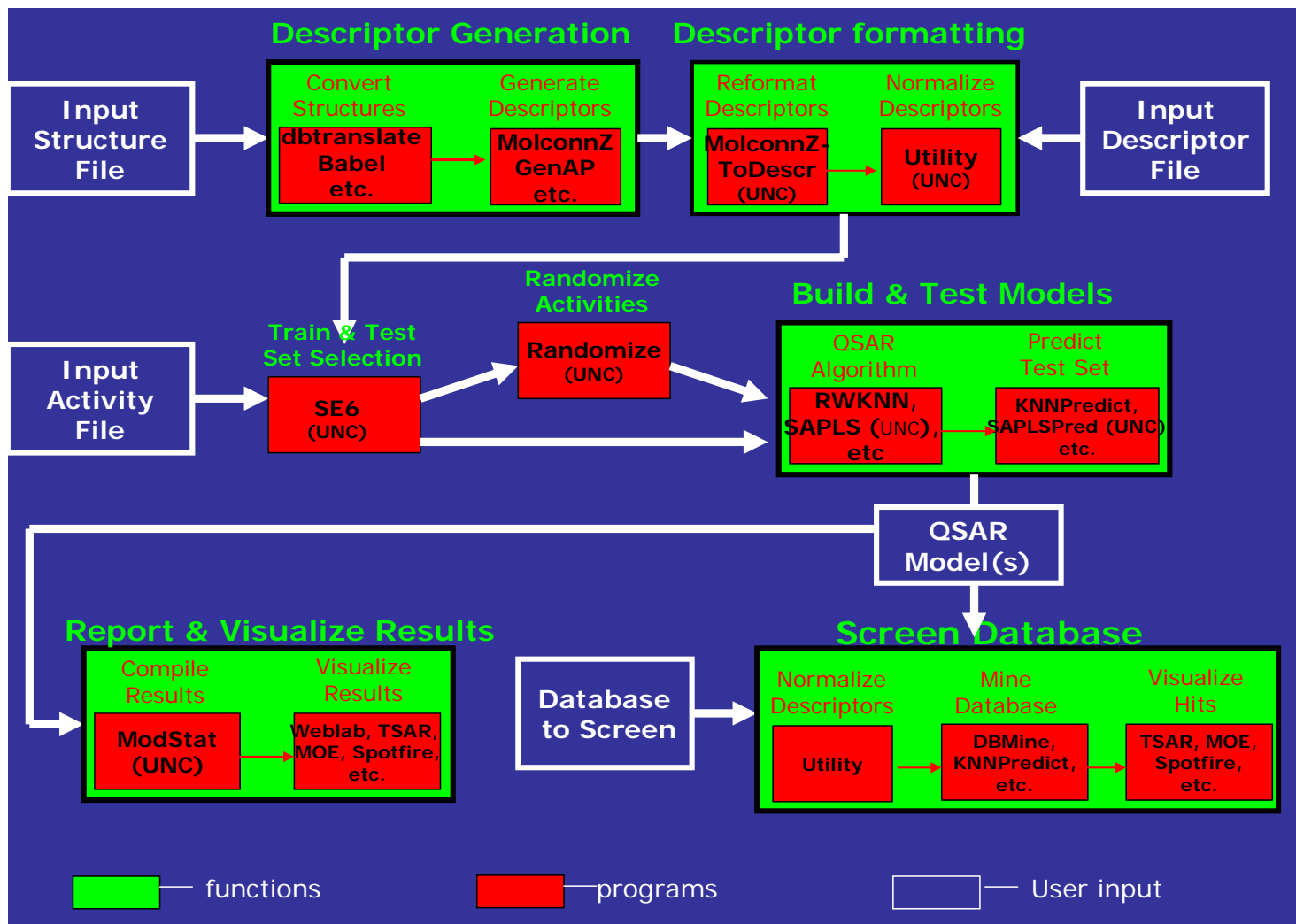
Flowchart of the combinatorial QSAR methodology





Flowchart of predictive toxicology framework based on validated combi-QSAR models. Numerous public datasets proposed.

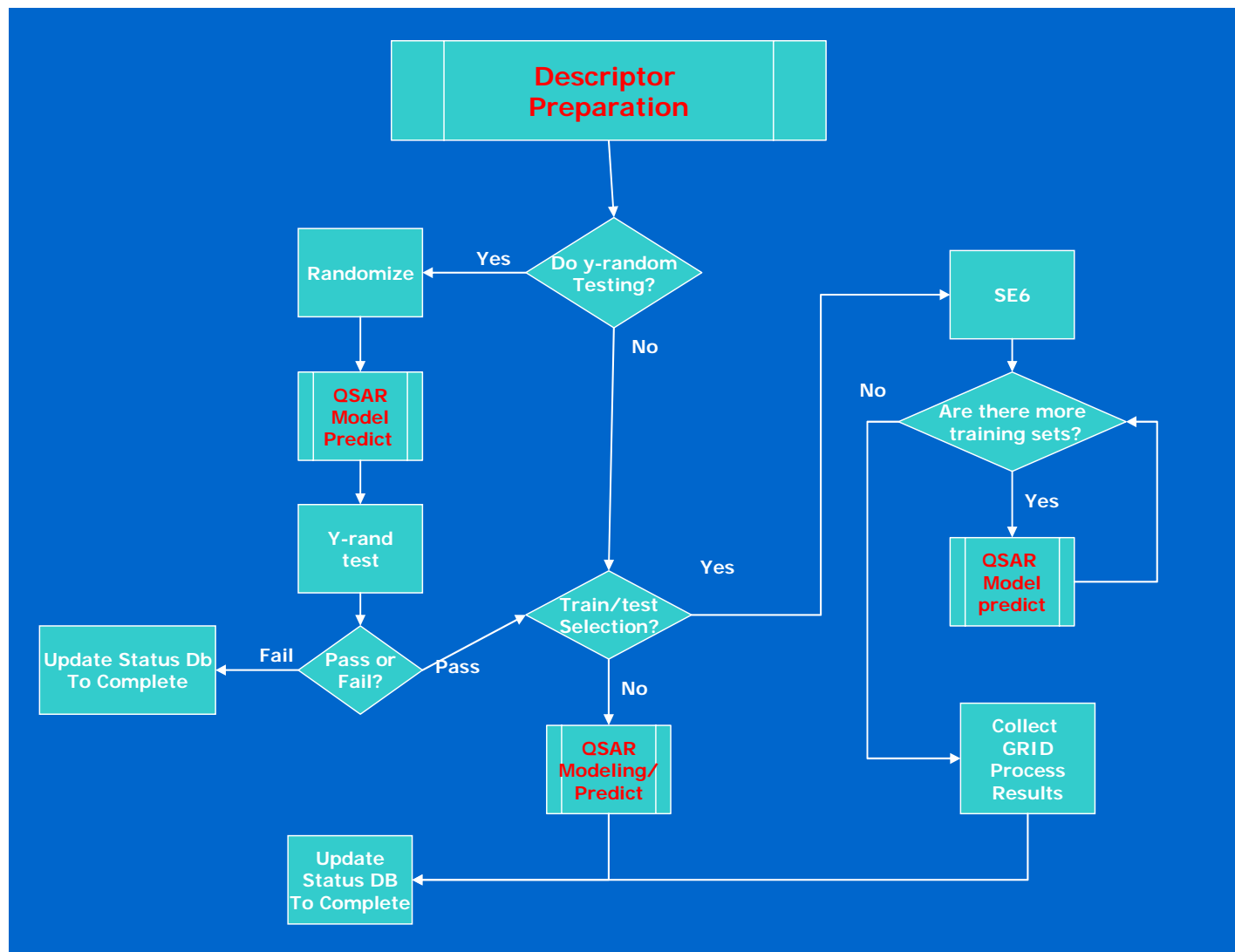




## Predictive QSPR modeling workflow







Workflow logic of model generation process





## Project 3



# Computational Infrastructure for Systems Toxicology

- David Stotts, Ph.D. (co-P.I.) – computer science, software engineering
- Ivan Rusyn, Ph.D. (co-P.I.) – toxicology, genomics
- Wei Wang, Ph.D. – computer science, data mining
- Brad Hemminger, Ph.D. – informatics and metadata issues
- David Threadgill, Ph.D. – mammalian genetics, genomics
- Additional programmers and students



## Project 3 objectives

- Develop and implement algorithms that streamline the analysis of multi-dimensional data streams in dose-response assessment and cross-species extrapolation.
- Facilitate the development of an industry-standard workflow for (i) analysis of the -omics data, (ii) linkages to classical indicators of adverse health effects, and (iii) integration with other types of biological information such as genome sequences and genetic differences between species.



## Project 3 objectives, cont.

- Build web-based, open-source and user-friendly graphical interfaces associated with interoperable computational tools for data analysis that facilitate incorporation of new data streams into basic research and decision-making pipelines (methods from Projects 1 and 2).
- Provide an interdisciplinary computer science resource to the environmental sciences and toxicology community
- Longer-term objectives include new software engineering methods for better execution and maintenance of above, and sharing and disseminating results



## A driving biological problem:

- Toxicogenetic analysis of susceptibility to toxicant-induced organ injury
- The model is being used by Drs. Threadgill and Rusyn involves extensive profiling of numerous mouse strains (over 40) for relevant organs
- Early data on acetaminophen and alcohol on liver
- Proposals for trichloroethylene and other toxicants on liver, kidney, and other organs



# The Mouse as a Model for Studying Genotype-Phenotype Interactions

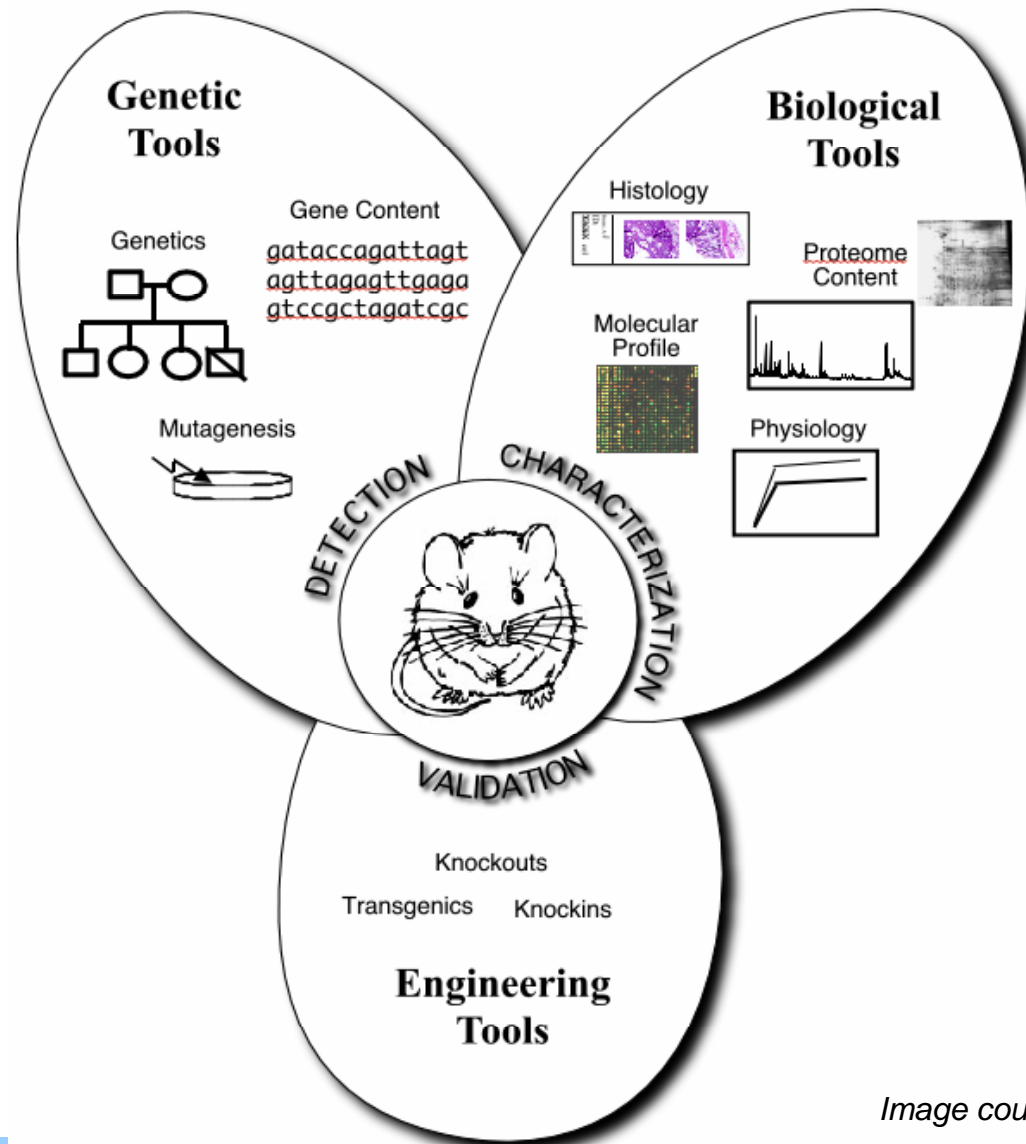
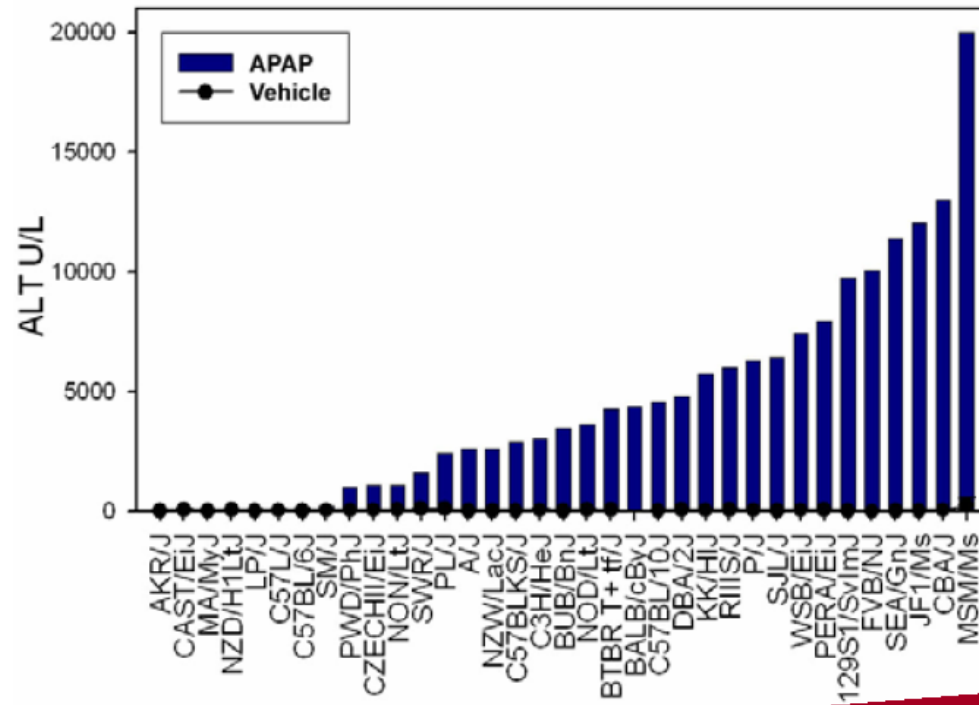


Image courtesy of D.W. Threadgill

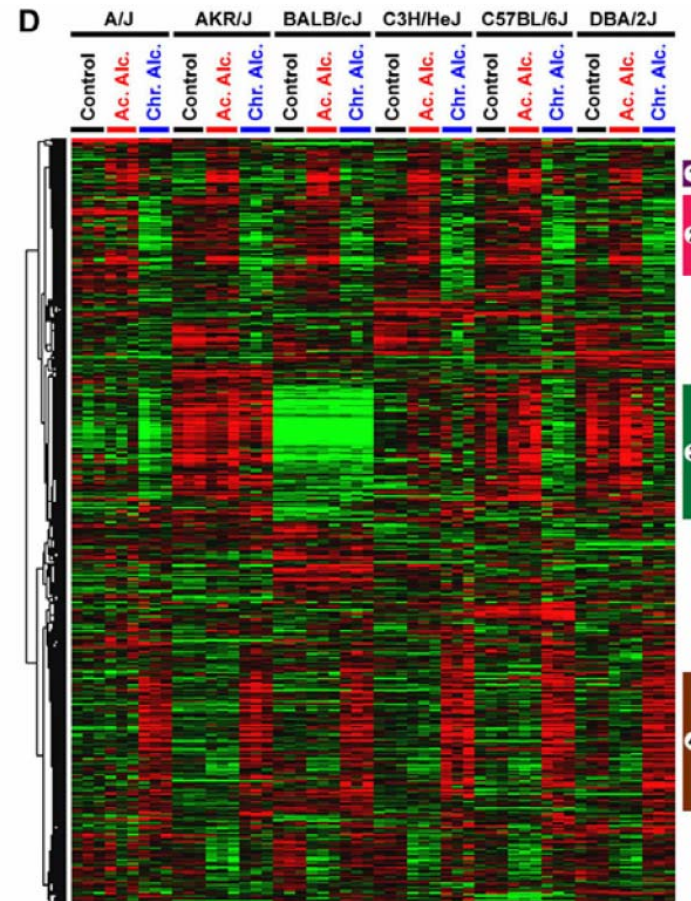
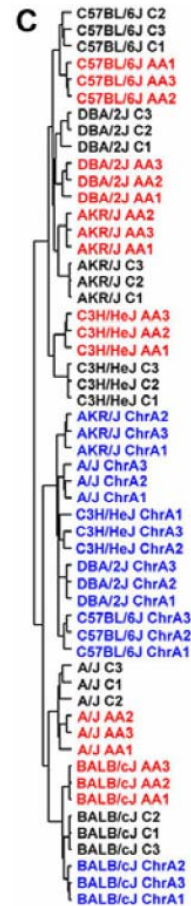
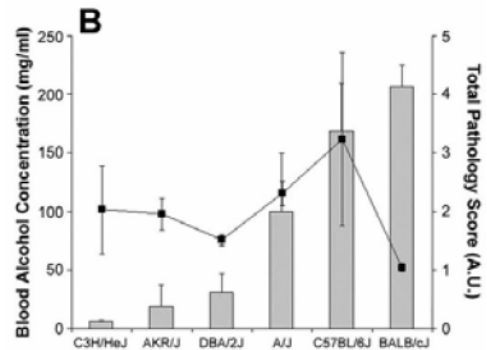
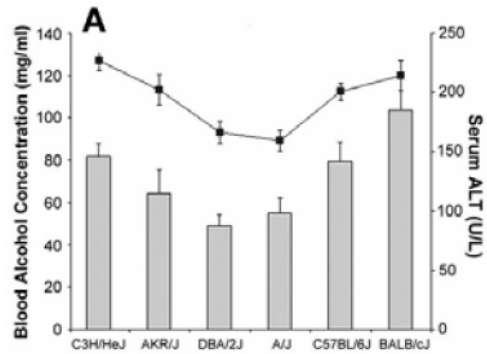


## A large variation in response by genetic background...



Strain-specific susceptibility to acetaminophen (APAP)-induced liver injury. Serum ALT levels (top panel) and tissue histopathological changes (bottom panel) were assessed 24 hrs after a single dose exposure to APAP (300 mg/kg, i.g., 24 hrs).





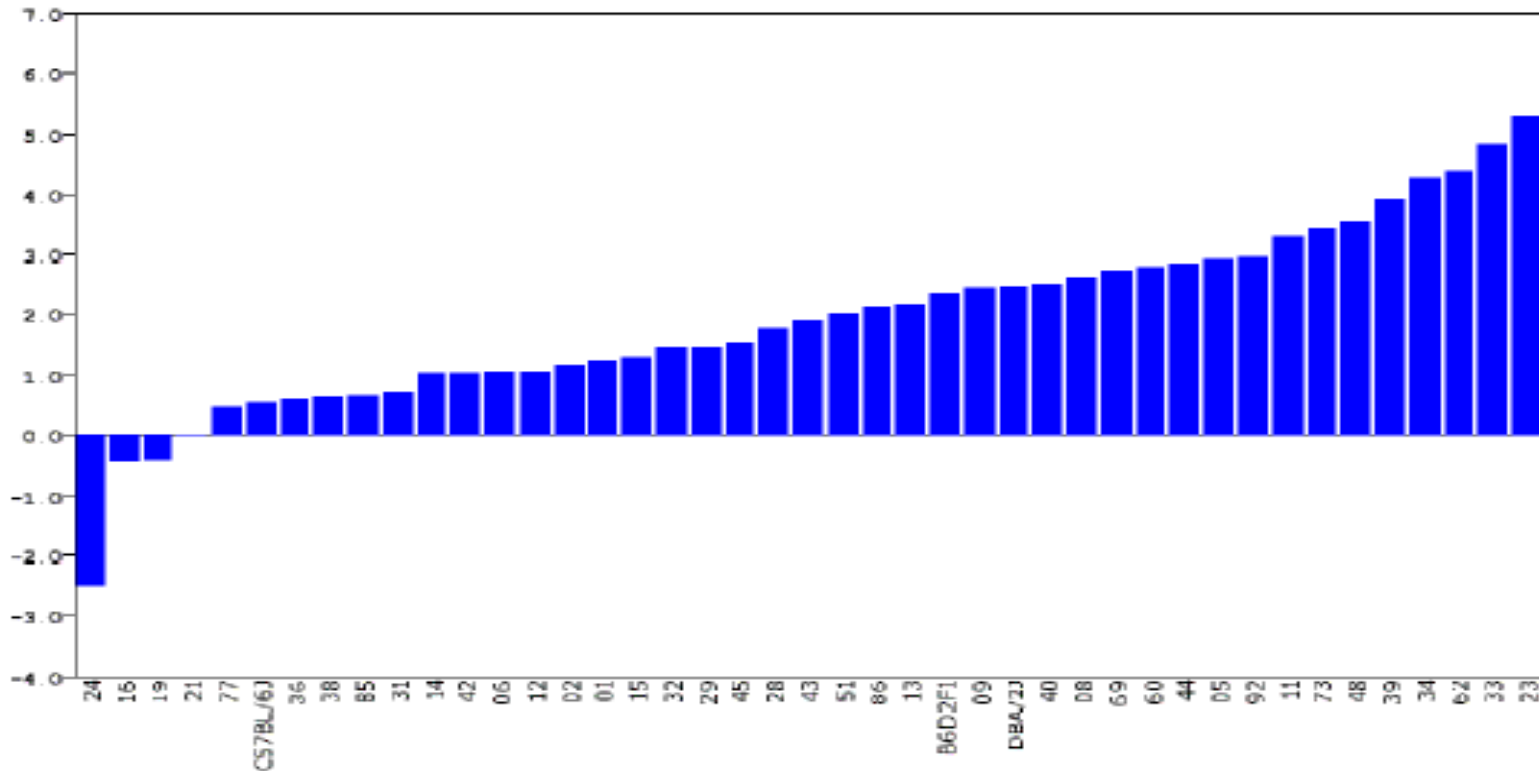
Toxicological and expression analysis of genotype-specific responses to ethanol in liver. Serum and liver tissues were collected from mice of 6 different strains after acute (5 g/kg, 6 hrs; A) or subchronic (4 weeks, B) treatment with ethanol.





# Variation in expression of potentially important genes...

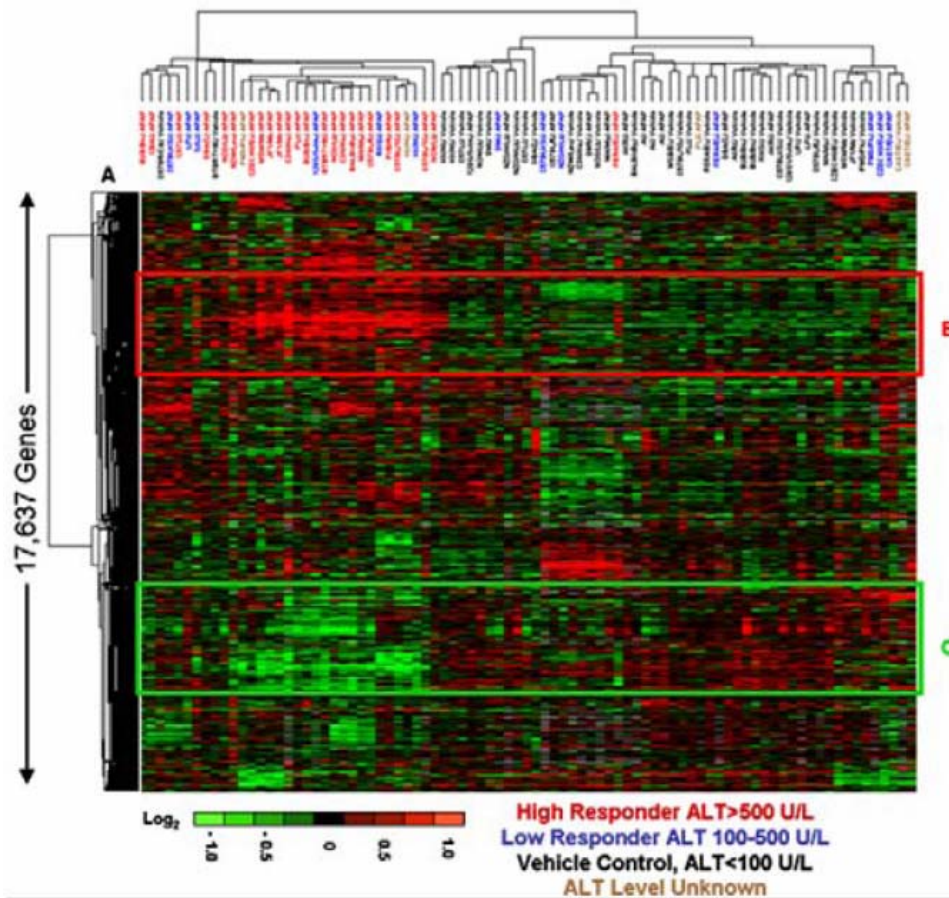
## Fmo3 : flavin containing monooxygenase 3



WebQTL can be used to select BXD strains that model genetic background-dependent variability in metabolism genes across the population. The difference in background expression of Fmo3 across BXD strains is shown.

Source: Ivan Rusyn and colleagues





Unsupervised hierarchical clustering of liver gene expression from APAP- (300 mg/kg, 24 hrs) or vehicle-treated mice shows distinct grouping of samples that correlates with the degree of liver injury.

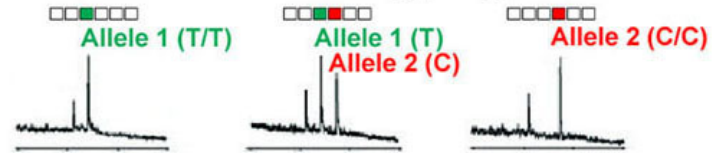


# Systems Biology Approach

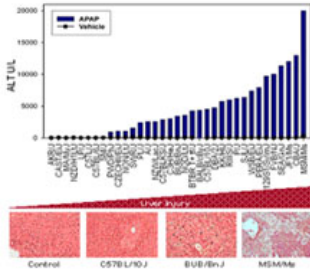
## Mouse Models



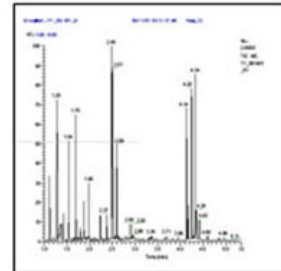
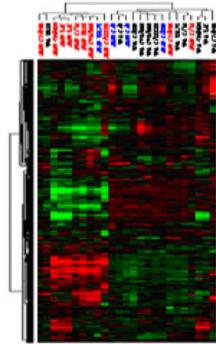
## SNP Genotyping



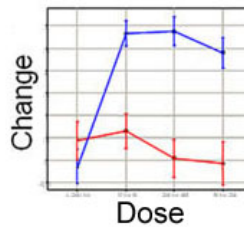
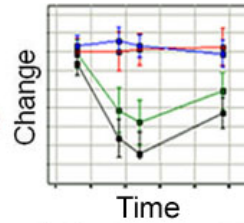
## Toxicity



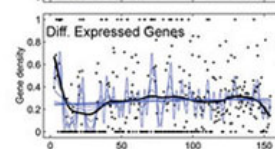
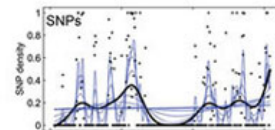
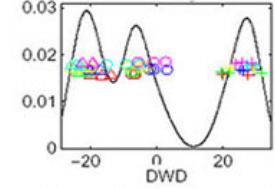
## Gene Expression



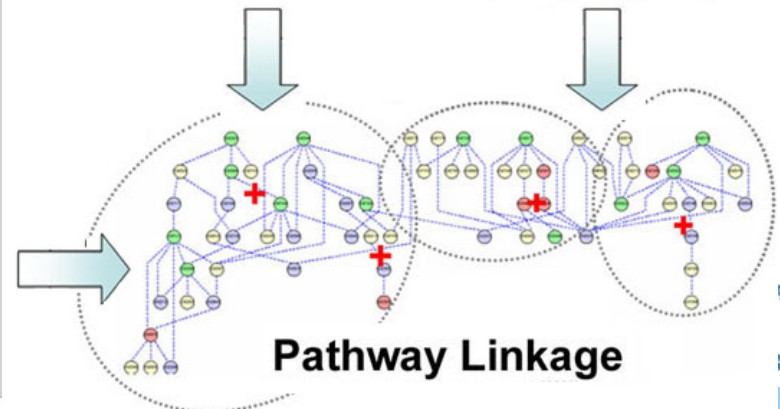
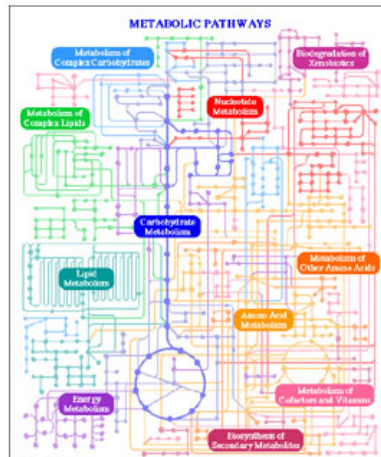
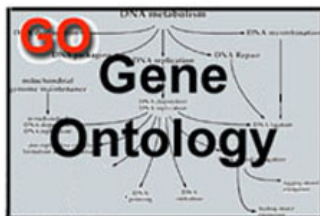
## Point Analysis



## Composite Analysis



BIOCARTA



Transcriptome map of forebrain. The physical location of each gene on the microarrays (y-axis is plotted with the genetic location of QTLs that regulate the steady-state level of its transcripts. The three major patterns of regulation are marked.

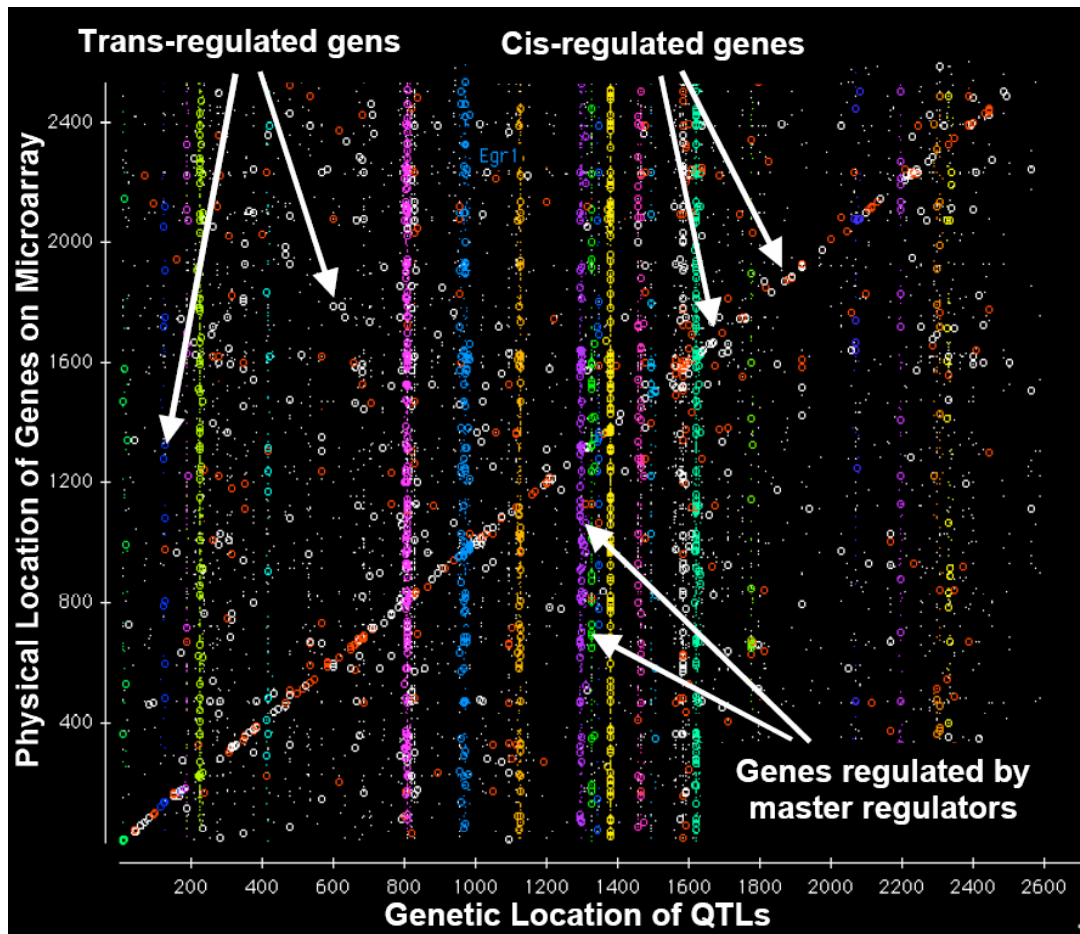
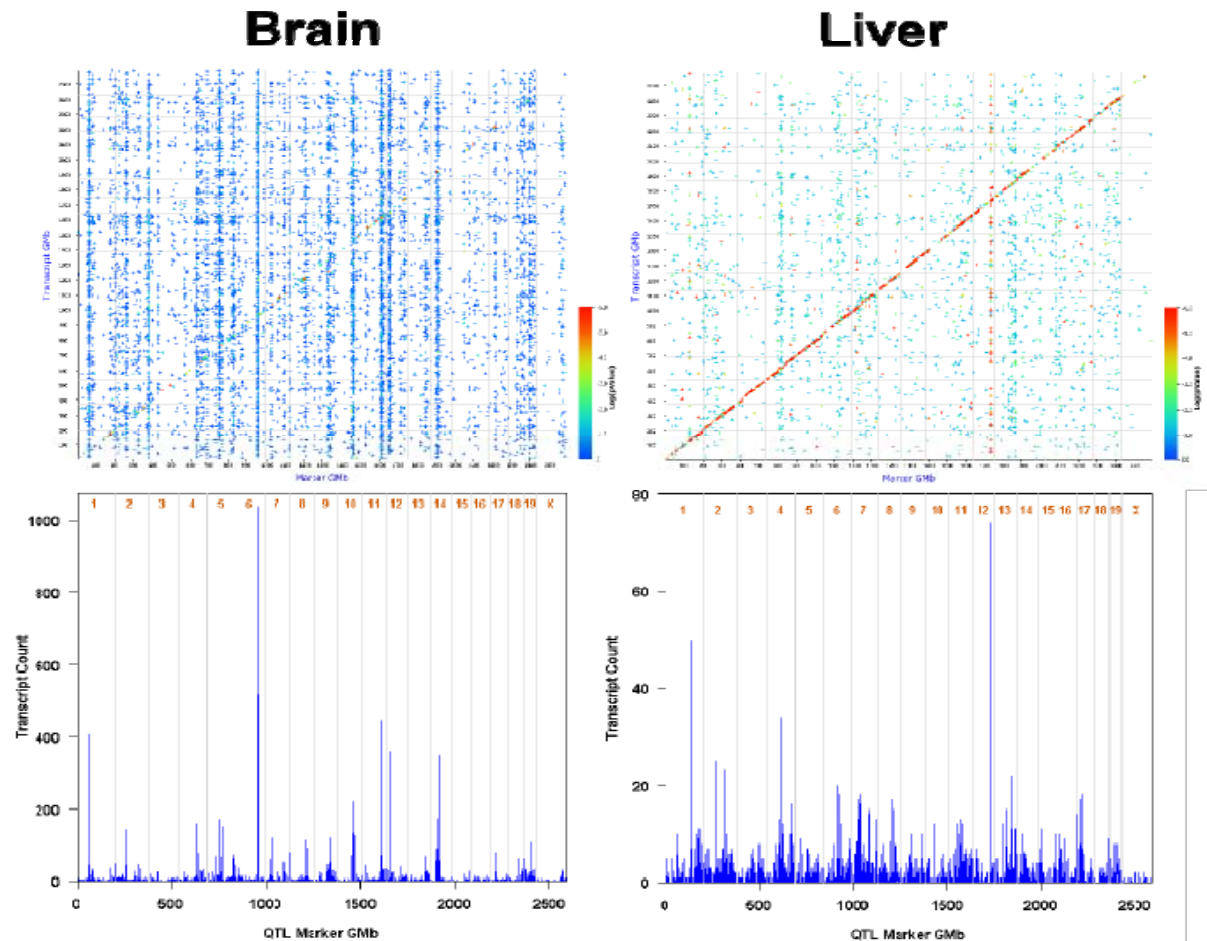


Image courtesy of D.W. Threadgill



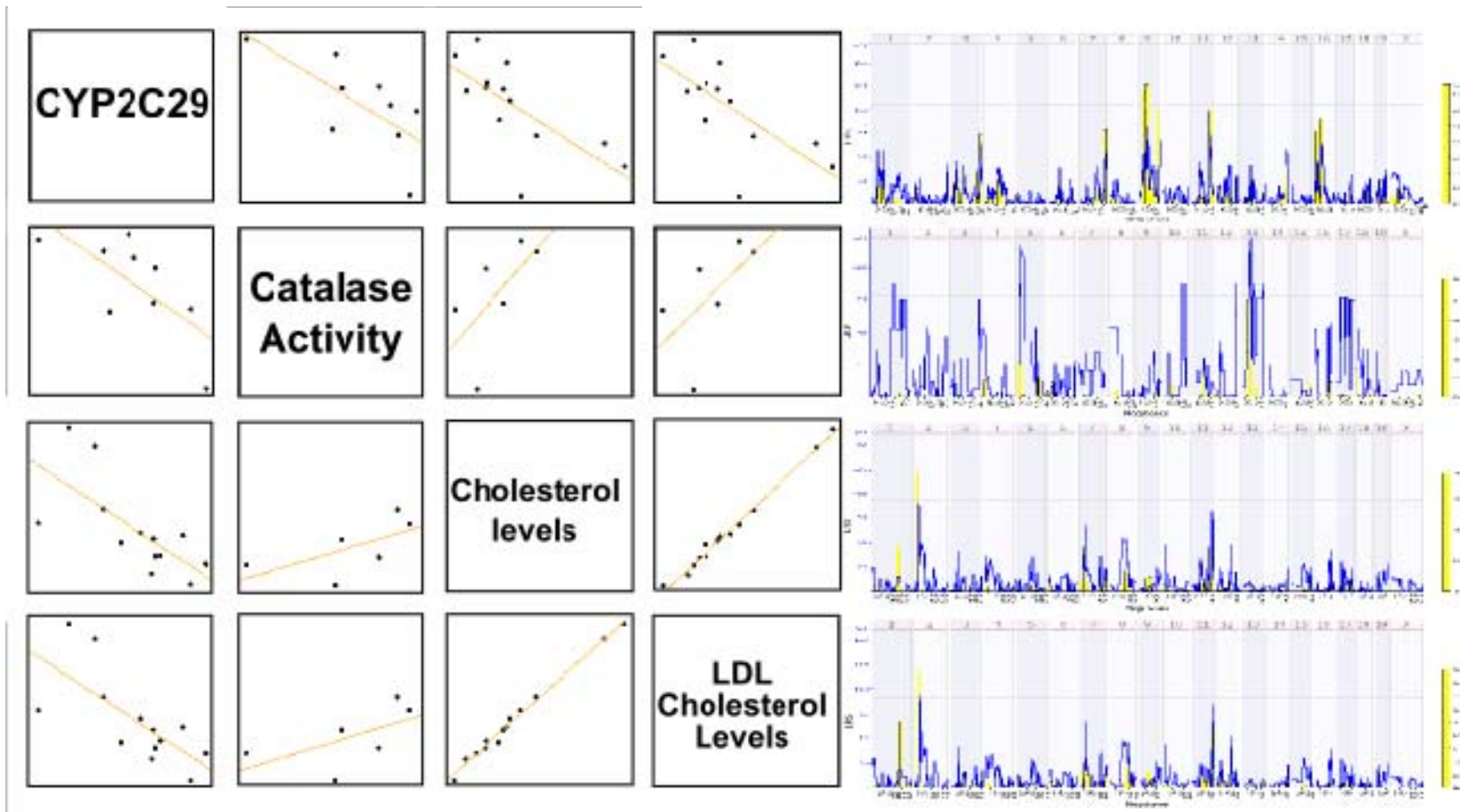
# Examination of genetic networks that regulate gene expression in liver (webQTL and beyond)



Transcriptome map for the murine brain and liver.

Source: Ivan Rusyn and colleagues



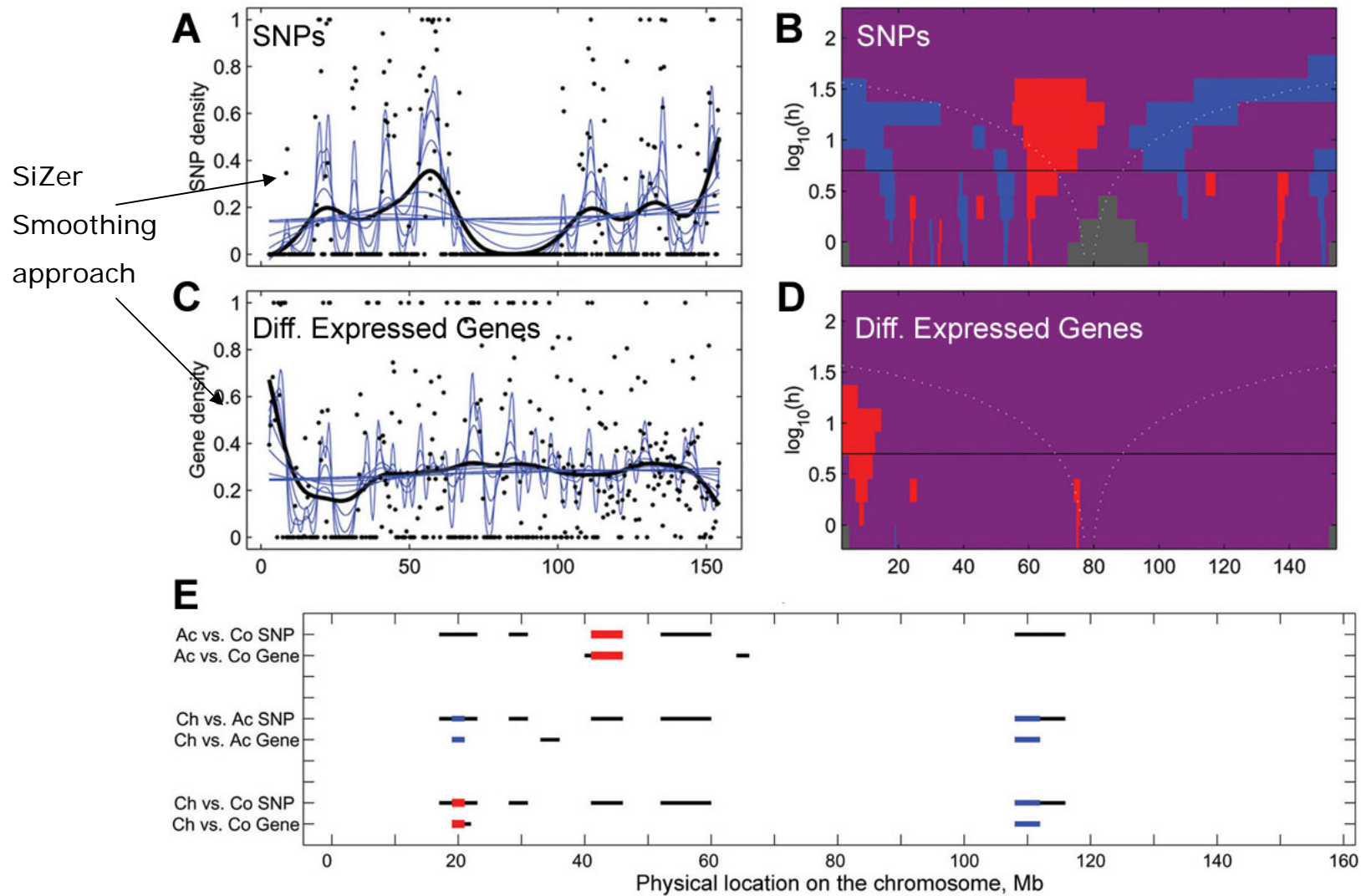


Source: Ivan Rusyn and colleagues

Correlation between gene expression of CYP2C29 and several liver-specific phenotypes recorded for BXD strains.



# Development of new methods for -omics data analysis: Finding associations between gene expression profiles and strain-specific genotyping data



- Data analysis procedures in concert with project 1, including principal component analyses, distance-weighted discrimination, SAFE, etc.
- Specific data mining approaches also proposed, such as subspace clustering (SNPs vs. phenotypes, gene expression), that fall outside of typical statistical framework
- The computational challenges are immense when we compare different –omics platforms (e.g., 100,000 SNPs X 30,000 transcripts)
- This requires serious computer science (activities of UNC SNP group).





## Solutions to a computational infrastructure

- Software technology – federated systems and architectures
- Execution platforms – workstations, grid computing, supercomputing
- UNC has access to resources of the Renaissance Computing Institute
- Data access and management – data mining, formats and data interchange, common abstractions/metadata issues

