Supplementary Materials

Genomic indicators in the blood predict drug-induced liver injury

Jianping Huang[1,8,*], Weiwei Shi[2,*], John Zhang[3,*], Jeff W. Chou[4,12], Richard S. Paules[5], Kevin Gerrish[5], Jianying Li[4,13] , Jun Luo[3], Russell D. Wolfinger[6], Wenjun Bao[6], Tzu-Ming Chu[6], Yuri Nikolsky[2], Tatiana Nikolskaya[2,11], Damir Dosymbekov[7], Marina O. Tsyganova[7], Leming Shi[8], Xiaohui Fan[1,8], J. Christopher Corton[9], Minjun Chen[8], Yiyu Cheng[1], Weida Tong[8], Hong Fang[10], & Pierre R. Bushel[4]

[1]Zhejiang University, Hangzhou, China. [2]GeneGO, Inc., St. Joseph, Michigan, USA. [3]Systems Analytics, Inc., Waltham, Massachusetts, USA. [4]Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA. [5]Microarray Group, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA. [6]Genomics Division, SAS, Cary, North Carolina, USA. [7]Vavilov Institute for General Genetics, Russian Academy of Sciences, Moscow B133, Russia. [8]Center for Toxicoinformatics, National Center for Toxicological Research, Food and Drug Administration, Jefferson Arkansas, USA [9]National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA, [10]Z-Tech Corporation, an ICF International Company at NCTR/Food and Drug Administration, Jefferson, Arkansas, USA, [11]Systems Biology Lab, Institute for General Genetics, Moscow, Russia

[12]Present address:
Jeff Chou
Wake Forest University School of Medicine
Department of Biostatistical Sciences
Medical Center Blvd
Winston-Salem, North Carolina 27157
Email: jchou@wfubmc.edu

[13]Present address:
Jianying Li
University of North Carolina – Chapel Hill
Lineberger Comprehensive Cancer Center
130 Mason Farm Road
Chapel Hill, North Carolina 27599
Email: jyli@med.unc.edu

*The individuals contributed equally to lead authorship of the paper

Correspondence should be addressed to: P.R.B (bushel@niehs.nih.gov)

Summary of incorrect predictions from the external validation of the gene-based and pathway-based classifiers

For the gene-based classifiers, three treatment time “dark” samples and nine treatment time “light” samples were predicted incorrectly, all at the higher doses of acetaminophen. Of the former, two were dosed with 2000 mg/kg acetaminophen (one [animal ID 526] for 6 hrs and one [animal ID 545] 24 hrs) and one [animal ID 427] dosed with 1500 mg/kg for 6 hrs. The latter had four dosed with 2000 mg/kg acetaminophen (three [animal IDs 501, 503 and 514] for 6 hrs and one [animal ID 522] for 48 hrs) and five dosed with 1500 mg/kg (two [animal IDs 413 and 414] for 6 hrs, one [animal ID 404] for 18 hrs and two [animal IDs 423 and 424] for 48 hrs).

For the pathway-based classifiers, all of the acetaminophen samples predicted incorrectly by the NC gene-based classifier were predicted incorrectly by the RF-b pathway-based classifier except animal IDs 423, 424 and 503 but included five “dark” samples; two (animal IDs 232 and 233) with 50 mg/kg acetaminophen for 24 hrs, two with 1500 mg/kg (one [animal ID 438] for 6 hrs the other [animal ID 440] for 18 hrs), one (animal ID 527) with 2000 mg/kg for 6 hrs and three “light” samples; one (animal ID 416) with 1500 mg/kg acetaminophen for 18 hrs and two with 2000 mg/kg (one [animal ID 513] for 6 hrs and the other [animal ID 521] for 24 hrs).

Supplementary Table 1. Signature Transferability of gene-based classifiers.

Option	ID	Accuracy	Accuracy
		(P < 0.05, FC > 2)	(P < 0.05, FC > 1.5)
Blood → Liver	RF	0.783 → 0.888	0.734 → 0.874
	SVM	0.786 → 0.817	0.797 → 0.829
	KNN	0.759 → 0.845	0.763 → 0.804
	NC	0.774 → 0.817	0.770 → 0.798
Liver → Blood	RF	0.888 → 0.720	0.895 → 0.727
	SVM	0.858 → 0.639	0.888 → 0.720
	KNN	0.879 → 0.586	0.853 → 0.660
	NC	0.834 → 0.651	0.878 → 0.597

* The top section gave the accuracy results based on Supplementary Figure 2. Accuracy on the left side of the " arrow” was yielded from the blood test set that was predicted using classifiers built on the blood training set, and accuracy on the right side was yielded from the liver test set that was predicted using the liver training set and the signature genes from the corresponding blood classifier. The bottom section was vice versa.

Supplementary Table 2. Prediction accuracies of gene-based classifiers acquired from Agilent and Affymetrix data.

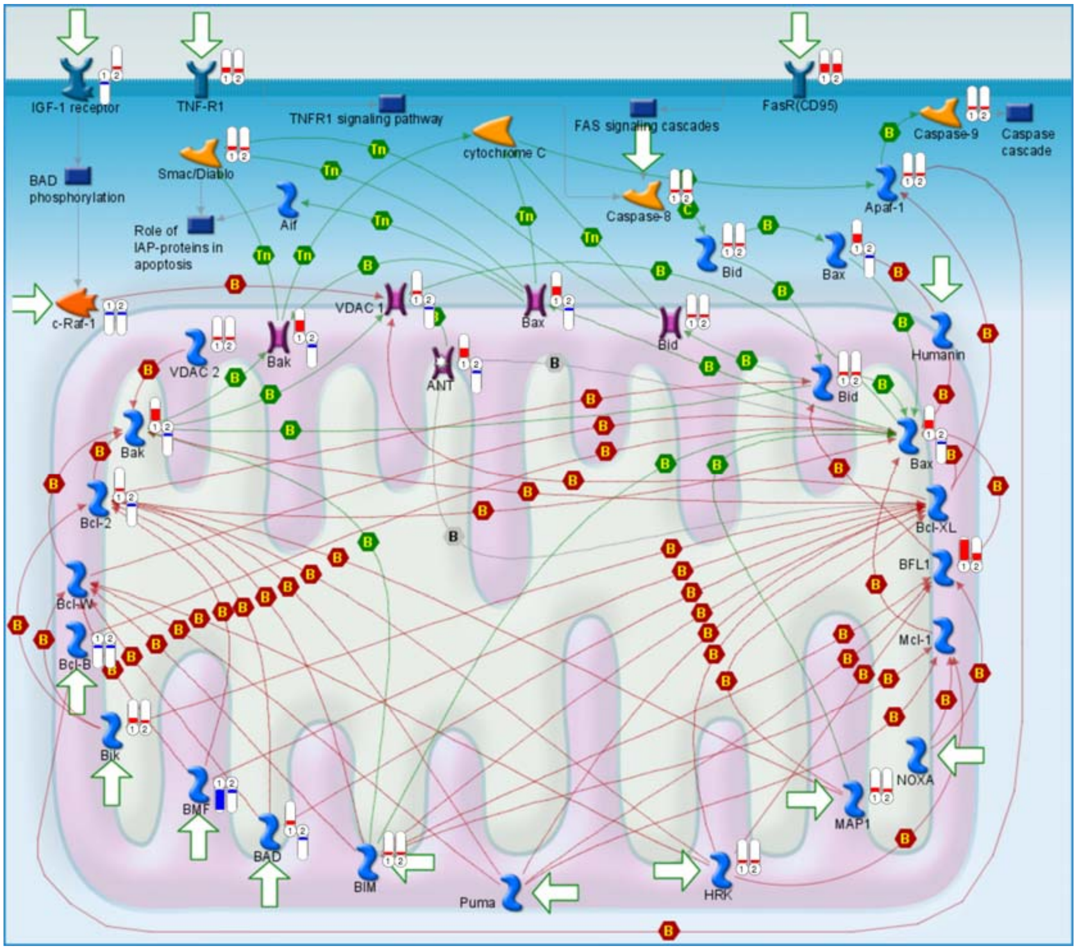
Classifiers	Mapping	Line 2	Line 2'	Line 1	Line 3	Line 3'	Line 5	Line 5'	Line 4	Line 6	Line 6'
SVM (mean)	SeqMap	0.616	0.728	0.837	0.615	0.724	0.516	0.526	0.905	0.516	0.530
	RefSeq	0.636	0.741	0.825	0.636	0.742	0.518	0.569	0.901	0.517	0.574
	Unigene	0.631	0.749	0.839	0.630	0.748	0.518	0.563	0.899	0.517	0.559
KNN euc (mean)	SeqMap	0.499	0.771	0.825	0.499	0.772	0.516	0.508	0.909	0.515	0.506
	RefSeq	0.484	0.775	0.815	0.485	0.775	0.516	0.521	0.912	0.515	0.513
	Unigene	0.484	0.790	0.813	0.485	0.790	0.516	0.515	0.909	0.515	0.511
DLDA (mean)	SeqMap	0.490	0.810	0.754	0.490	0.810	0.516	0.579	0.859	0.515	0.580
	RefSeq	0.487	0.807	0.732	0.487	0.807	0.516	0.569	0.861	0.515	0.566
	Unigene	0.487	0.808	0.732	0.489	0.806	0.516	0.584	0.857	0.515	0.573

Log base 2 intensity data for the Affymetrix Liver arrays (AFX Liver) and log base 2 average ratio data for the Agilent Blood arrays (AG2 Blood) were used with the probe-sets in three mappings (array technology features to each other): Sequence, Refseq and Unigene. Classification performed 100 times with stratified splits of the training and test data at 70% and 30 % respectively. SVM: support vector machines (linear kernel with C=1), KNN: k-nearest neighbors euc (k=5, Euclidian distance), DLDA: diagonal linear discriminant analysis. Line numbers with a prime (‘) denote that batch correction was performed on the data. Line number is donated in Figure 2a.

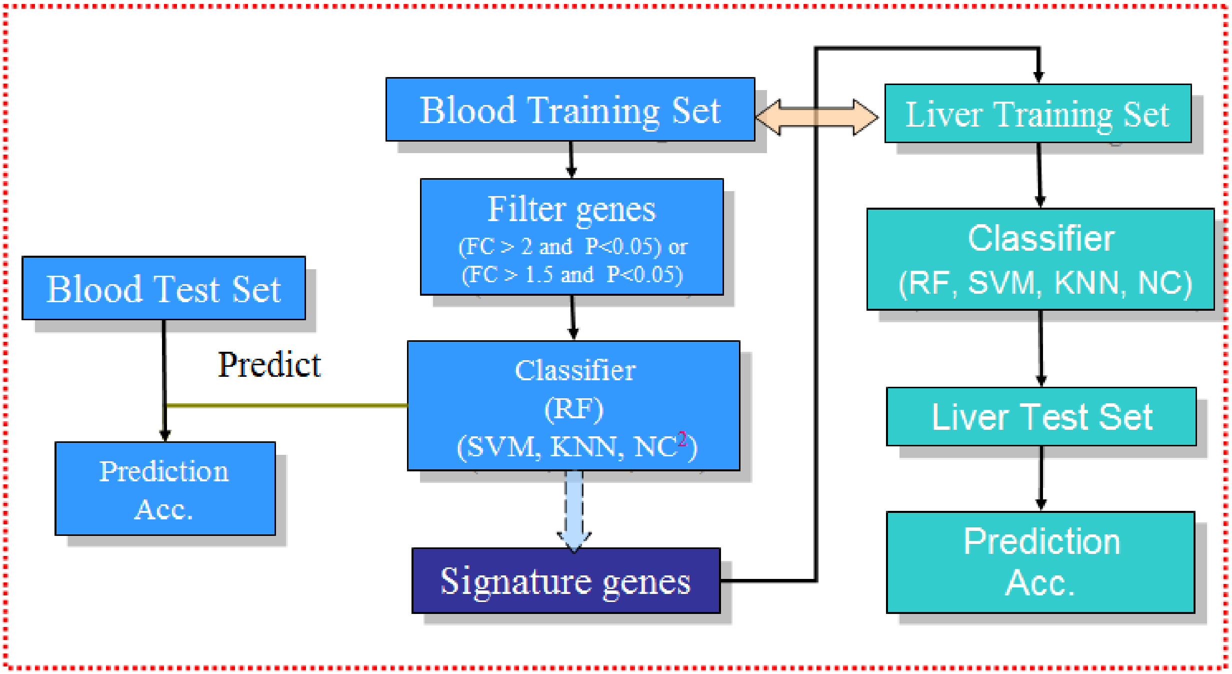
Supplementary Table 3. Biological processes over-represented by genes* in the biclusters.

Identifier	Term	p-value
Up-regulated		
GO:0002274	myeloid leukocyte activation	3.04E-05
GO:0002444	myeloid leukocyte mediated immunity	7.08E-05
GO:0045321	leukocyte activation	1.95E-04
GO:0002349	histamine production during acute inflammatory response	1.21E-03
GO:0001821	histamine secretion	1.21E-03
GO:0012502	induction of programmed cell death	1.54E-03
GO:0002532	production of molecular mediator of acute inflammatory response	2.39E-03
GO:0043067	regulation of programmed cell death	2.40E-03
GO:0051052	regulation of DNA metabolic process	2.95E-03
GO:0032496	response to lipopolysaccharide	3.11E-03
GO:0006626	protein targeting to mitochondrion	4.51E-03
GO:0019221	cytokine and chemokine mediated signaling pathway	5.78E-03
GO:0002443	leukocyte mediated immunity	6.19E-03
GO:0006935	chemotaxis	6.58E-03
GO:0006952	defense response	6.91E-03
GO:0043068	positive regulation of programmed cell death	9.83E-03
rno04670	Leukocyte transendothelial migration	2.50E-04
rno00450	Selenoamino acid metabolism	3.71E-02
rno04650	Natural killer cell mediated cytotoxicity	4.56E-02
rno00010	Glycolysis / Gluconeogenesis	7.89E-02
Down-regulated		
GO:0006091	generation of precursor metabolites and energy	2.16E-10
GO:0008610	lipid biosynthetic process	2.28E-09
GO:0006695	cholesterol biosynthetic process	6.87E-07
GO:0008203	cholesterol metabolic process	3.62E-06
GO:0006631	fatty acid metabolic process	6.22E-06
GO:0044262	cellular carbohydrate metabolic process	4.21E-04
GO:0042221	response to chemical stimulus	7.85E-04
GO:0005975	carbohydrate metabolic process	8.15E-04
GO:0009896	positive regulation of catabolic process	1.05E-03
GO:0050818	regulation of coagulation	1.40E-03
GO:0007596	blood coagulation	2.68E-03
GO:0045819	positive regulation of glycogen catabolic process	3.03E-03
GO:0007599	hemostasis	3.71E-03
GO:0042493	response to drug	5.78E-03
GO:0002526	acute inflammatory response	8.75E-03
GO:0050819	negative regulation of coagulation	9.20E-03
rno01040	Polyunsaturated fatty acid biosynthesis	2.17E-05
rno04610	Complement and coagulation cascades	3.70E-05
rno00960	Alkaloid biosynthesis II	1.28E-02
rno03320	PPAR signaling pathway	2.31E-02
rno00071	Fatty acid metabolism	2.52E-02

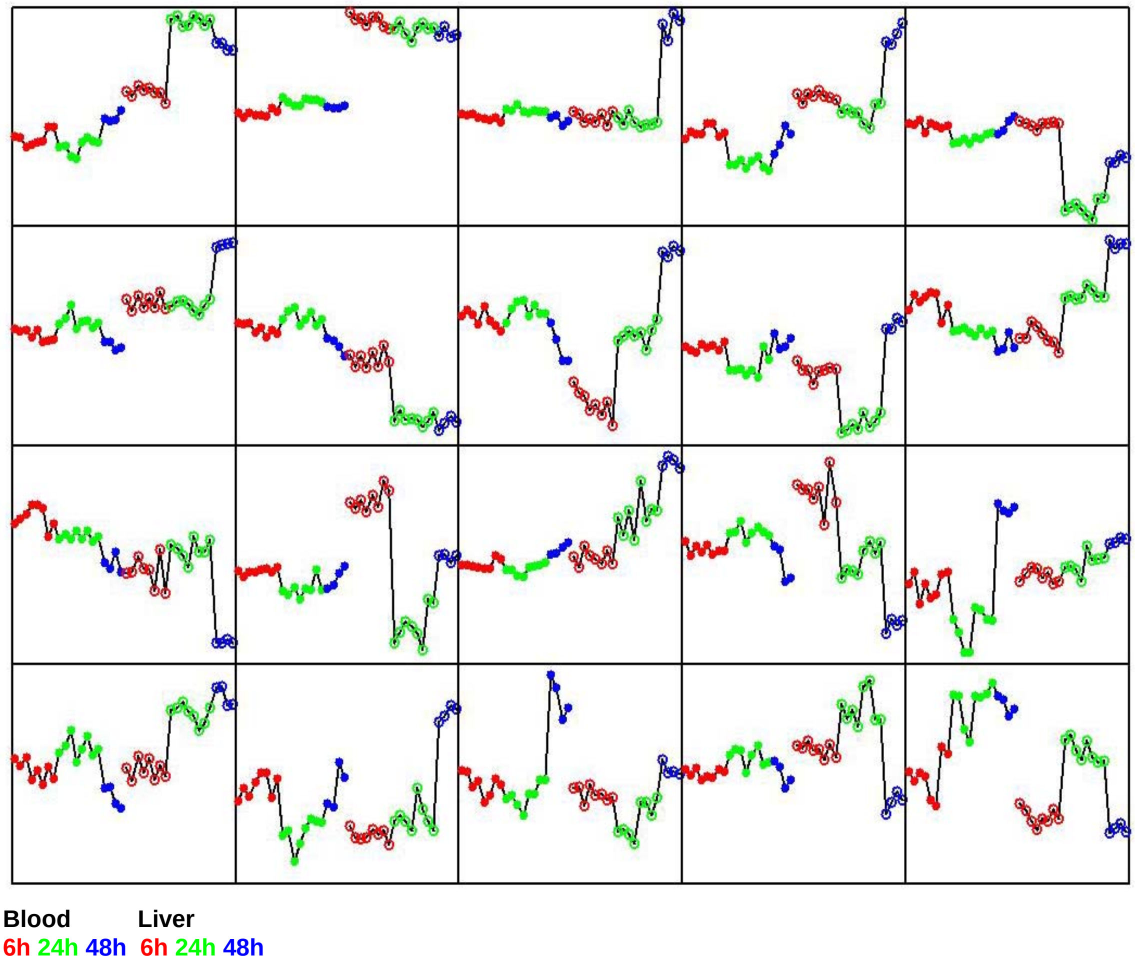
*The co-expressed genes are located in Supplementary file C.



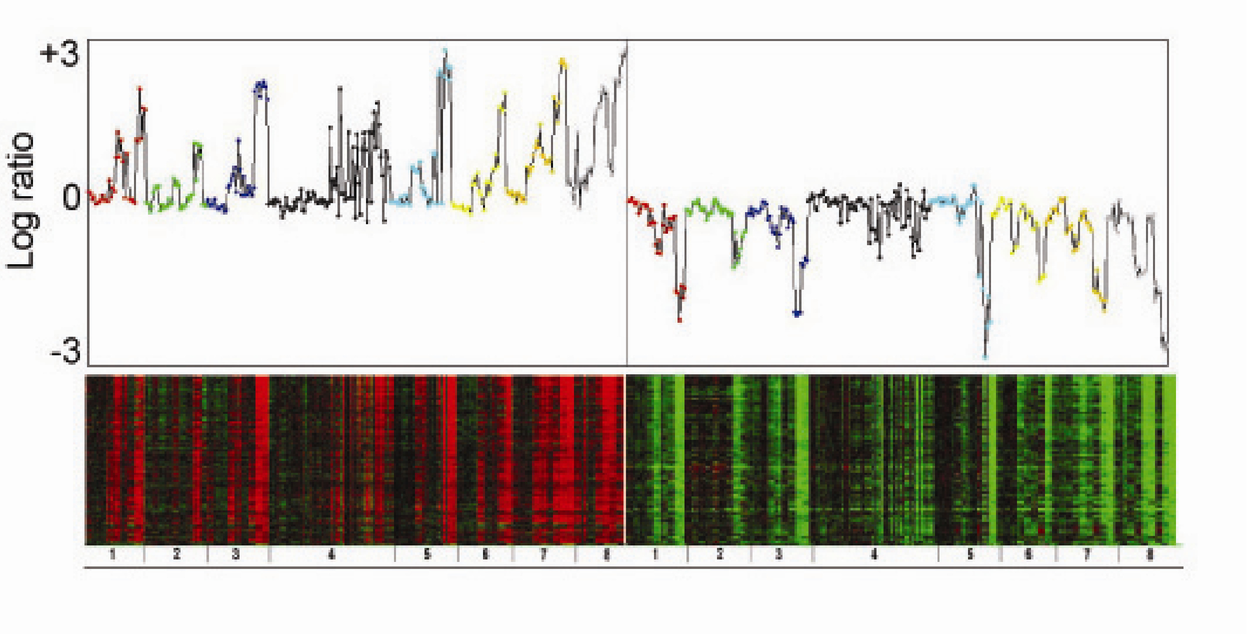
Supplementary Figure 1. The descriptor pathway map for Regulation of Apoptosis by Mitochondrial Proteins (MetaCore) with superimposed co-expressed genes from cc-Biclustering. Up- and down-regulated genes are shown as red and blue “thermometer” histograms, correspondingly. The height of the thermometer” indicates the numerical value of differential gene expression. Thermometer 1 is the fold change of means with sign of necrosis vs. those with no necrosis for liver samples; Thermometer 2 is the fold change of means with sign of necrosis vs. those with no necrosis for blood samples. The co-expressed genes are located in Supplementary file C.



Supplementary Figure 2. Signature Transferability was performed by splitting 318 blood samples into training samples (175) and test samples (143). SVM, KNN, and NC classifiers were built using a forward feature selection with a five-fold internal cross validation on the training data set. The gene signatures from the blood were transferred to the liver training set and predicted liver test set (Supplementary Table 1 top) and Signature Transferability was done vice versa by training on the liver to predict the blood sample (Supplementary Table 1 bottom).



Supplementary Figure 3. Extracting Patterns and Identifying co-Expressed Genes (EPIG) patterns of gene expression from the (Agilent ratio data) blood and the liver samples for 1,2-dichlorobenzene at dose 1500mg. In each pattern, on the x-axis, the left three groups (6h, 24h and 48h) are blood samples, the right three groups (6h, 24h and 48h) are liver samples. The y-axis is the log base 2 ratio representing the average of the top 5 gene expression profiles in each pattern. The patterns are listed by number in increasing sequential order from the left to the right, row by row. At the high dose (1500 mg), the liver showed severe damage. A large number of genes were significantly differentially expressed in the liver only (i.e. patterns 1 and 7).



Supplementary Figure 4. Expression patterns. Top: Representative expression patterns (Agilent data) from the genes in the biclusters of the rat liver samples exposed to any one of eight compounds. The left is the pattern from the up-regulated genes, the right from the down-regulated genes. The x-axis: treatment of the animals with the compounds (each represented by a color and number) and the y-axis: the average of the log base 2 ratio of the expression intensity from the top 5 genes with the highest expression profile correlation to the bicluster pattern. The compounds are: (1) red:1,2-Dichlorobenzene, (2) green: 1,4-Dichlorobenzene, (3) blue: Bromobenzene, (4) black: Diquat, (5) light blue: Galactosamine, (6) yellow: Monocrotaline, (7) gold: N-Nitrosomorpholine and (8) grey: Thioacetamide. Each compound has its dose exposure from low (left) to high (right). The time duration from 6 (left), 24 (middle), to 48 hr (right) within each dose. Bottom: Heat map of the gene expression patterns in the biclusters. There are 330 genes up-regulated (left half) and 409 genes down-regulated (right half). The co-expressed genes are located in Supplementary file C. The red color indicates up-regulation and the green color down-regulation.