Supplementary Information

A comprehensive study design reveals treatment- and transcript abundance-dependent concordance between RNA-seq and microarray data

Charles Wang[1,*], Binsheng Gong[2,*], Pierre R. Bushel[3,4,*,†], Jean Thierry-Mieg[5], Danielle Thierry-Mieg[5], Joshua Xu[2], Hong Fang[6], Huixiao Hong[2], Jie Shen[2], Zhenqiang Su[2], Joe Meehan[2], Xiaojin Li[7], Lu Yang[7], Haiqing Li[7], Paweł P. Łabaj[8], David P. Kreil[8,9], Dalila Megherbi[10], Caiment Florian[11], Stan Gaj[11], Joost van Delft[11], Jos Kleinjans[11], Andreas Scherer[12], Devanarayan Viswanath[13], Jian Wang[14], Yong Yang[14], Hui-Rong Qian[14], Lee J. Lancashire[15], Marina Bessarabova[15], Yuri Nikolsky[15], Cesare Furlanello[16], Marco Chierici[16], Davide Albanese[16], Giuseppe Jurman[16], Samantha Riccadonna[16], Michele Filosi[16], Roberto Visintainer[16], Ke K. Zhang[17], Jianying Li[3,18], Jui-Hua Hsieh[19], Daniel L. Svoboda[20], James C. Fuscoe[21], Youping Deng[22], Leming Shi[2,23], Richard S. Paules[24], Scott S. Auerbach[19,†] and Weida Tong[2,†]

[1]Center for Genomics and Division of Microbiology & Molecular Genetics, School of Medicine, Loma Linda University, Loma Linda, California, USA

[2]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA

[3]Microarray and Genome Informatics Group, National Institute of Environmental Health Sciences, RTP, North Carolina, USA

[4]Biostatistics Branch, National Institute of Environmental Health Sciences, RTP, North Carolina, USA

[5]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

[6]The Office of Scientific Coordination, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA

[7]Functional Genomics Core, Department of Molecular Medicine, Beckman Research Institute, City of Hope, Duarte, California, USA

[8]Boku University Vienna, Austria

[9]Life Sciences, University of Warwick, Coventry, U.K.

[10]CMINDS Research Center, Department of Electrical and Computer Engineering, Francis College of Engineering, University of Massachusetts, Lowell, Massachusetts, USA

[11]Department of Toxicogenomics, Maastricht University, Maastricht, The Netherlands

[12]Australian Genome Research Facility Ltd, The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia

[13]Abbott Laboratories, Abbott Park, Illinois, USA

[14]Research Informatics and Statistics, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana, USA

[15] Thomson Reuters, IP & Science, Carlsbad, California, USA

[16]Fondazione Bruno Kessler, Trento, Italy

[17]Bioinformatics core, Department of Pathology, University of North Dakota, Grand Forks, North Dakota, USA

[18]Kelly Government Solutions, Inc., Durham, North Carolina, USA

[19]Biomolecular Screening Branch, Division of the National Toxicology Program, National Institute of Environmental Health Sciences, RTP, North Carolina, USA

[20]SRA International, Durham, North Carolina, USA

[21]Division of Systems Biology, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA

[22]Department of Internal Medicine and Biochemistry, Rush University Medical Center, Chicago, Illinois, USA

[23]Center for Pharmacogenomics, Schools of Life Sciences and Pharmacy, Fudan University, Shanghai, China

[24]Laboratory of Toxicology and Pharmacology, National Institute of Environmental Health Sciences, RTP, North Carolina, USA

*These authors contributed equally to this work.

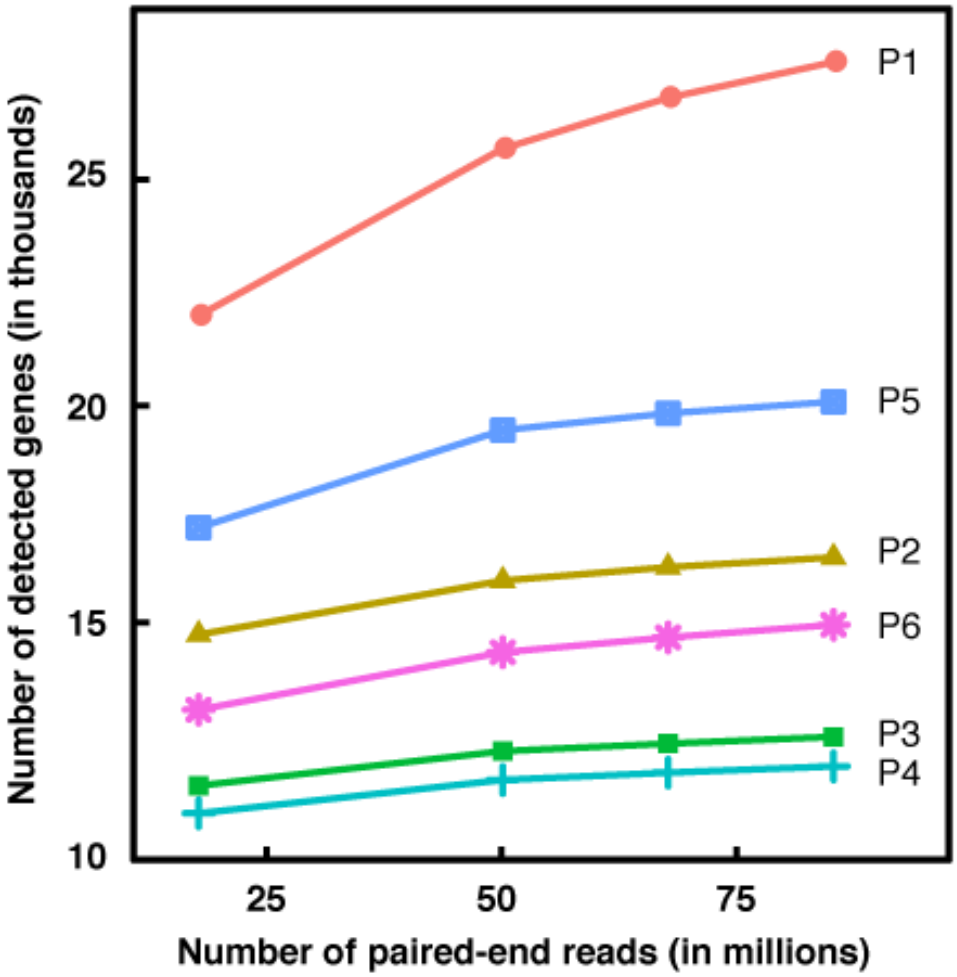
†Corresponding authors: Weida Tong (weida.tong@fda.hhs.gov), Scott S. Auerbach (AuerbachS@niehs.nih.gov), and Pierre R. Bushel (bushel@niehs.nih.gov)

Disclaimer: The views presented in this article do not necessarily reflect current or future opinion or policy of the US Food and Drug Administration. Any mention of commercial products is for clarification and not intended as an endorsement. This article may be the work product of an employee or group of employees of the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), however, the statements, opinions or conclusions contained therein do not necessarily represent the statements, opinions or conclusions of NIEHS, NIH or the United States government.

Table of Contents

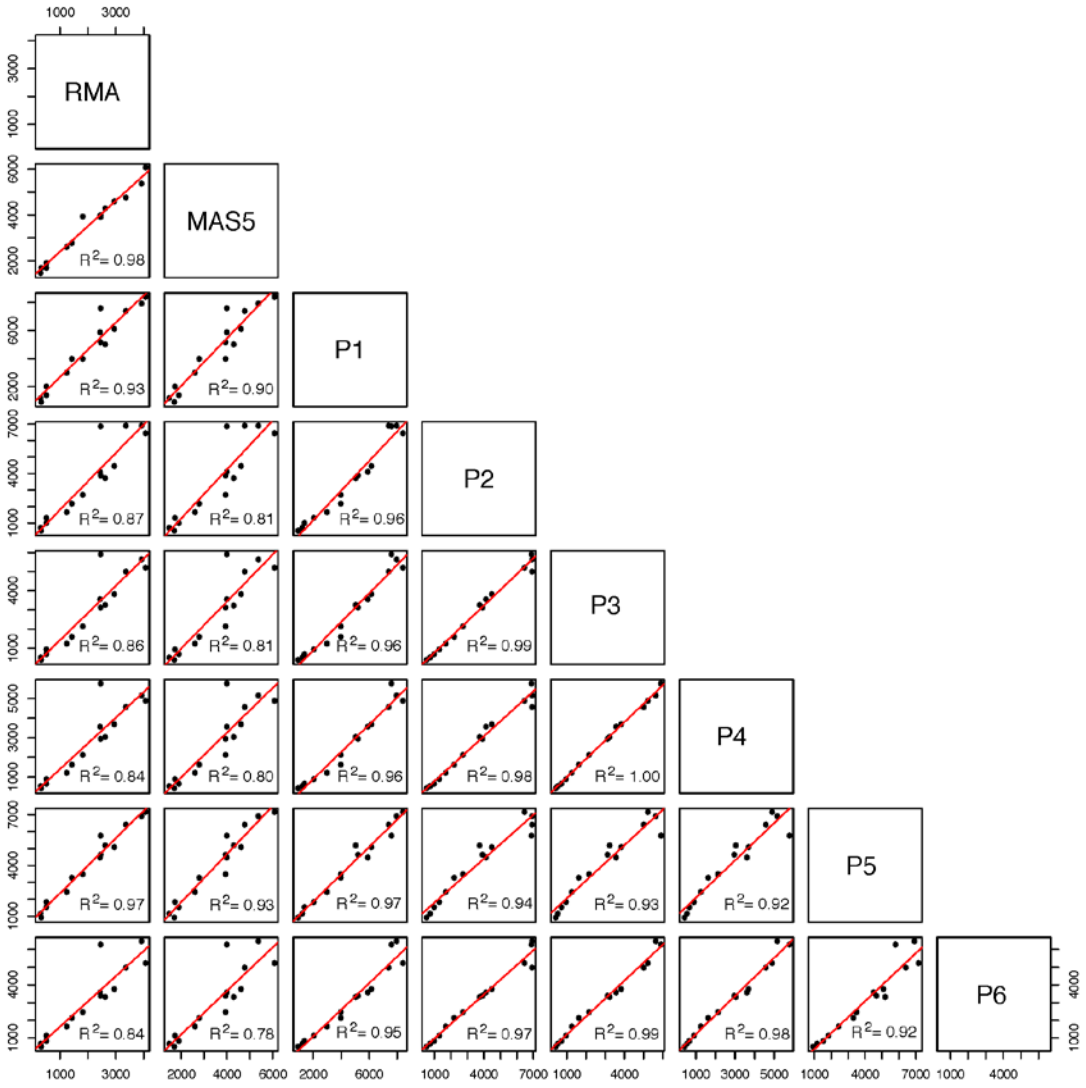
Supplementary Figures

Supplementary Figure 1. Gene detection at different sequencing depths.



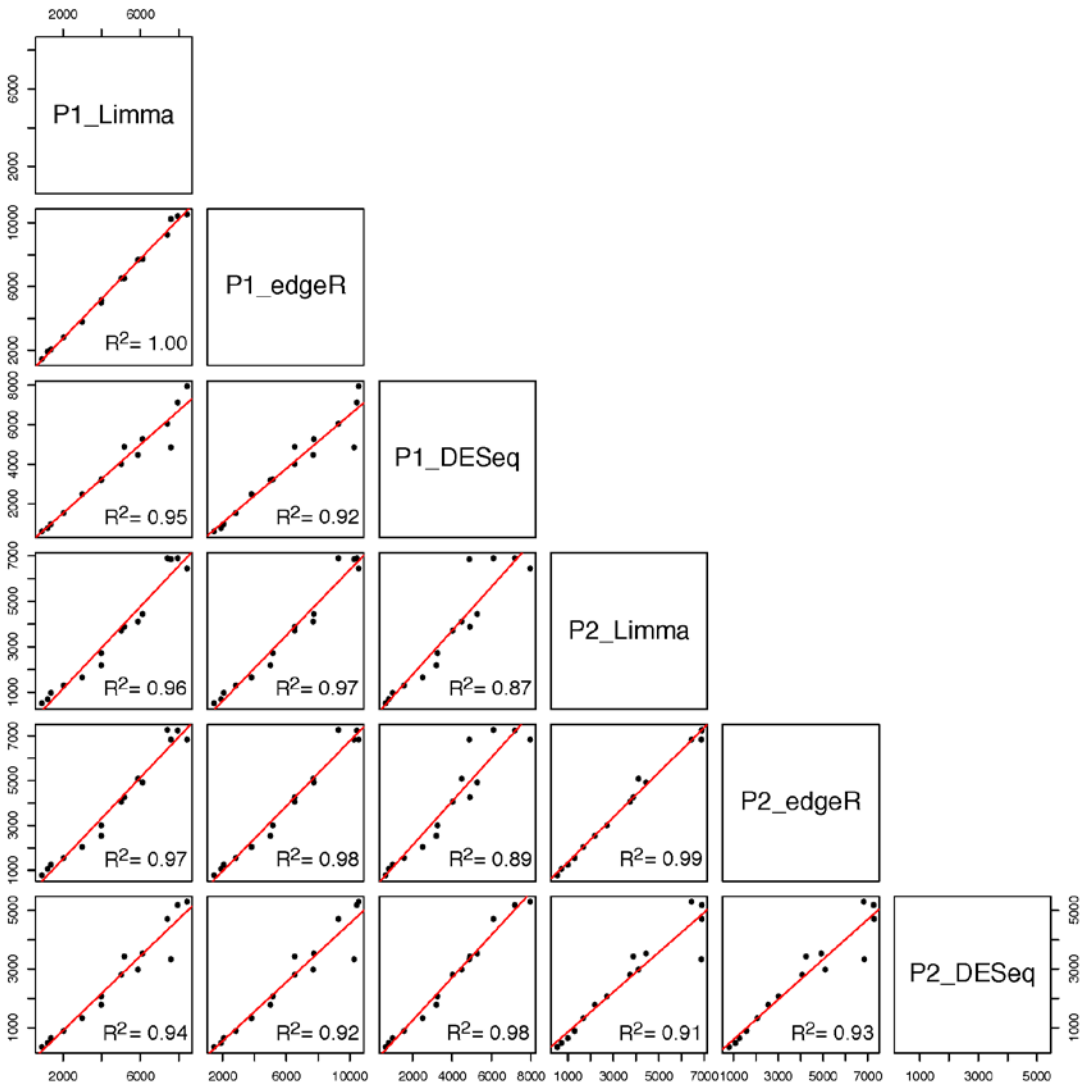
The x-axis is the number of pair-end reads for a sample treated by chemical AFL. The y-axis is the number of genes detected for each RNA-seq data analysis pipeline (P1-P6). A gene is detected if no less than 4 read pairs or 8 single-end reads mapped depending on whether reads are mapped in pair by the RNA-seq data analysis pipeline.

Supplementary Figure 2. Pair-wise correlation of the number of DEGs for 15 chemicals among two microarray data processing methods (RMA and MAS5) and six RNA-seq data analysis pipelines..



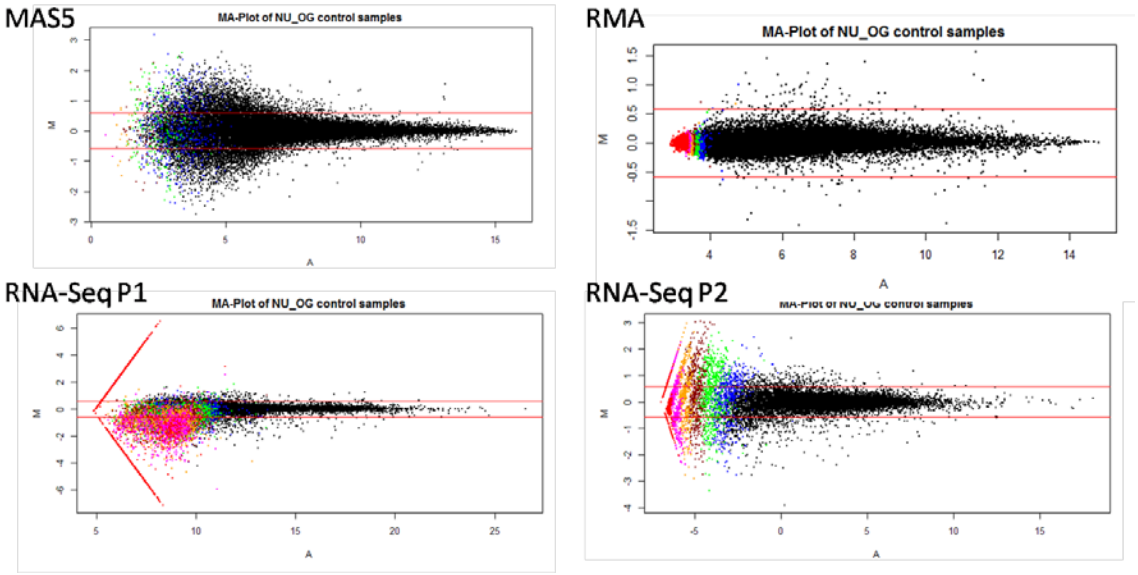
For each subplot, the x-axis is the number of DEG for the data processing method shown in the column head while the y-axis is the number of DEG for the method shown in the row end. Genes or probesets are determined by limma to be differentially expressed if the absolute fold change greater than 1.5 and P-value less than 0.05.

Supplementary Figure 3. Pair-wise correlation of the number of DEGs for 15 chemicals among three commonly used DEG methods (limma, edgeR, and DESeq) for two chosen RNA-seq data analysis pipelines (P1 and P2).



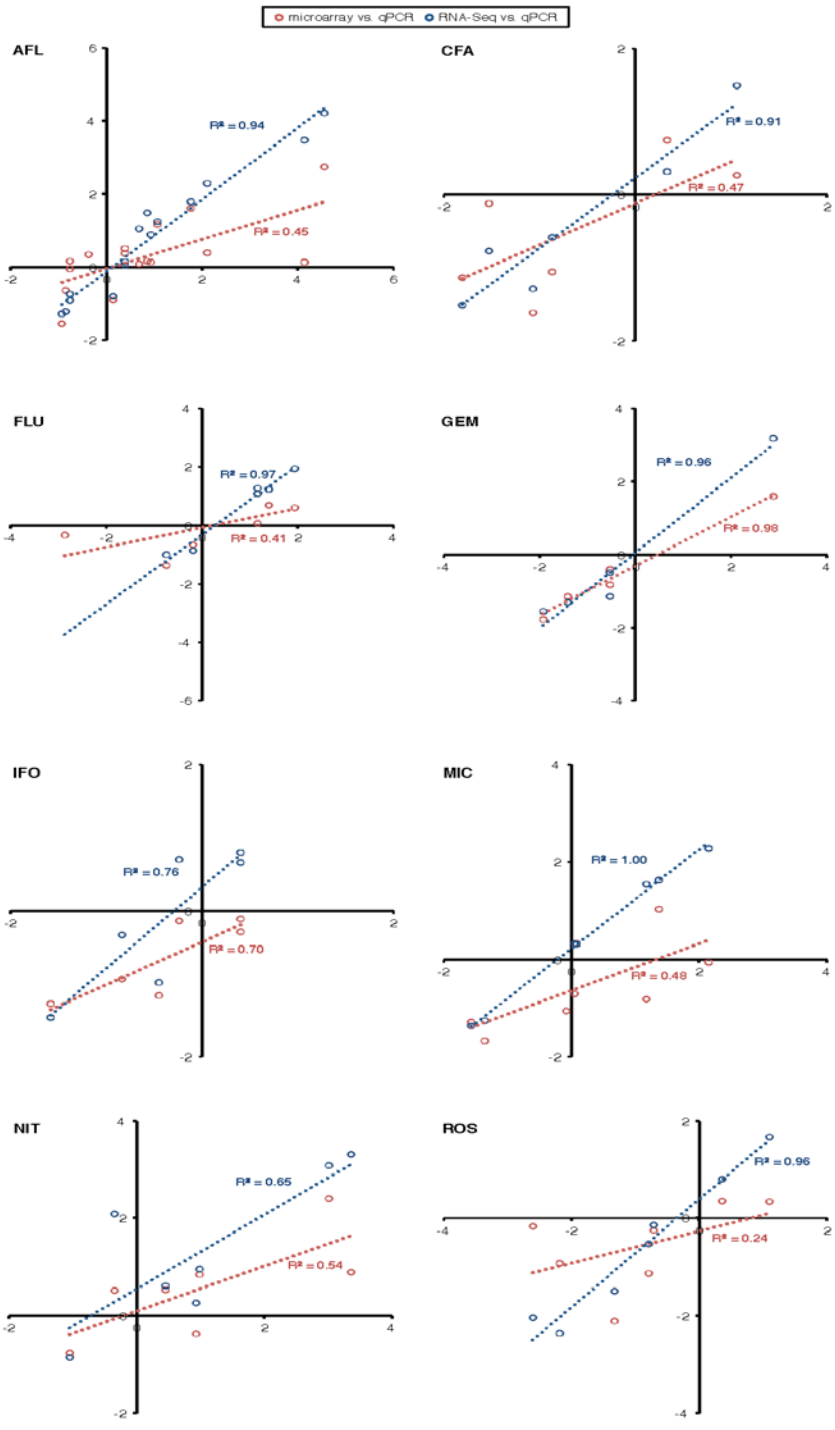
For each subplot, the x-axis is the number of DEG for the data analysis combination shown in the column head while the y-axis is the number of DEG for the data analysis combination shown in the row end. Genes or probesets are determined to be differentially expressed if the absolute fold change greater than 1.5 and P-value less than 0.05.

Supplementary Figure 4. The MA plot between two “virtual” samples.



The MA plot between two “virtual” samples for two microarray data processing methods (RMA and MAS5) and two chosen RNA-seq data analysis pipelines (P1 and P2). Specifically, one virtual sample is the pooling of the six controls in the test set by taking their average expression for each gene, the other is the pooling of the six controls in the training set with the same vehicle (NU(non-nutritive)) and route (OG(oral gavage)). Different colors represent different numbers of samples (>= 10, 8, 6, 4, 2, or 0) whose expression level are in the low 5% of the data range. Horizon parallel red lines are the 1.5 fold change.

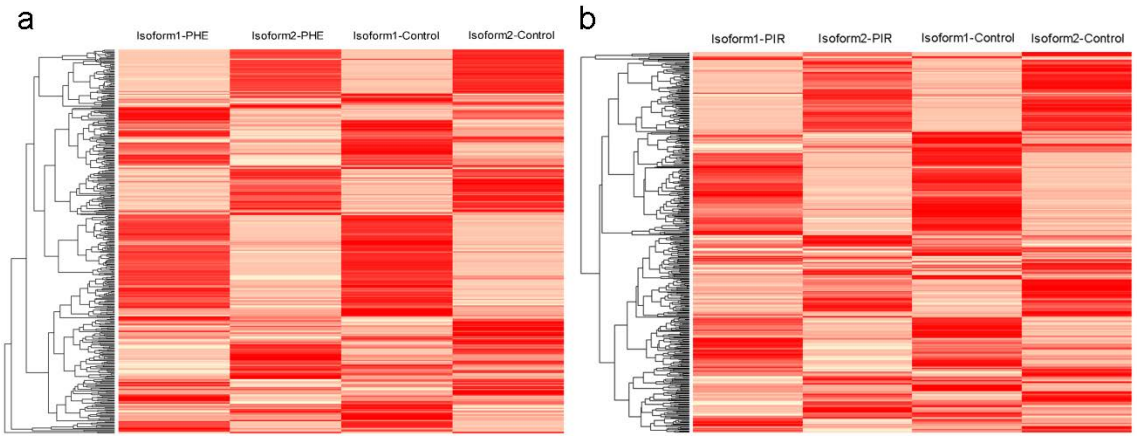
Supplementary Figure 5. Correlation of log2 fold change between qPCR and the two high-throughput platforms.



Correlation of log2 fold change between qPCR and the two high-throughput platforms (RNA-seq and microarray) of assayed genes for each of eight selected chemicals. No fold change or P-value cut-off was applied. The x-axis is the log2

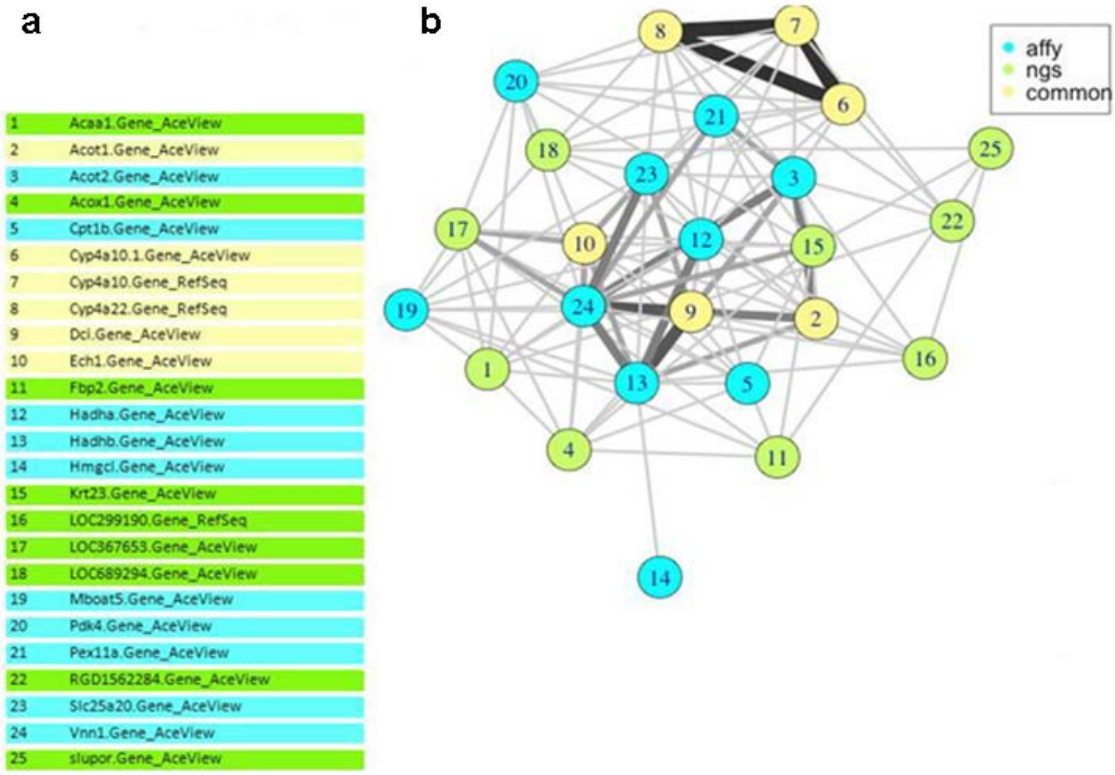
fold change for qPCR and the y-axis for the high throughput platforms. Dots and regression lines are colored blue for RNA-seq and red for microarray.

Supplementary Figure 6. Clustering and heatmap based on raw counts of reads mapped to an isoform and transformed to proportions within a sample.



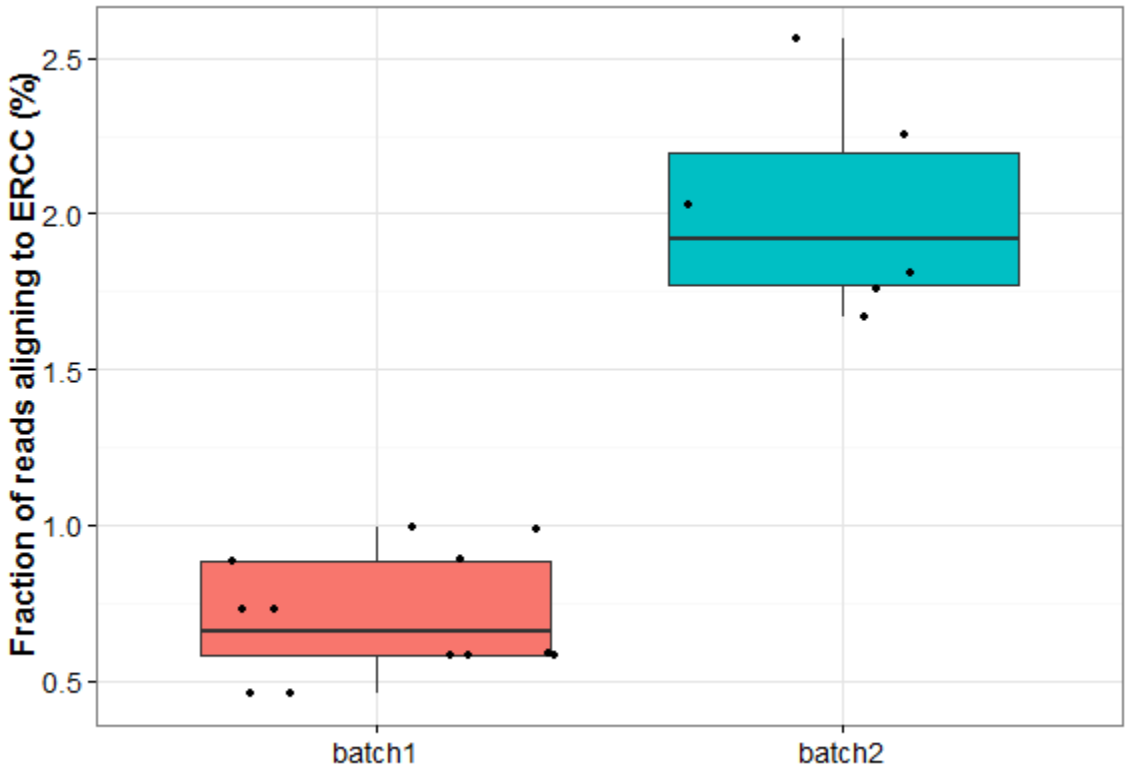
a) 408 significant transcripts (y-axis) in PHE. b) 449 significant transcripts (y-axis) in PIR. Statistical significance of a differentially expressed isoform between the treated and control was determined using a Fisher's exact test based on MISO inferred assignment of reads to each isoform (see Online Methods). Hierarchical clustering of the transcripts based on the Bray-Curtis dissimilarity matrix and performed by average linkage grouping. Red denotes highly abundant isoforms, yellow denotes lowly abundant isoforms. The longer the cluster branch, the more dissimilar transcripts are in terms of isoform usage between treated and control.

Supplementary Figure 7. Consistency of gene co-expression networks for the RNA-seq and microarray data.



Several genes and interactions are consistently found to be in common in co-expression networks reconstructed for the five MOAs on the training data (WGCNA networks, co-expression evaluated as TOM similarity) for the top 25 features (A) of an unified signature built on the AceView expression index table. The signature was derived from a GGSSL (semi-supervised) classifier, with ReliefF and KNN used for feature ranking, on 400 features (200 most discriminant for RNA-seq pipeline P1 and 200 for microarray), achieving 69% accuracy on validation data. Only top 10% interactions are shown in the graph (b), with link width proportional to weight, ranging from 1 (thinner: link in only one of the MOA specific networks) to 5 (thicker: link common to all the 5 MOA nets).

Supplementary Figure 8. Fraction of reads aligning to ERCC spike-ins for samples with duplicate libraries.



Fraction of reads aligning to ERCC spike-ins for samples with duplicate libraries shows a considerable variation between library preparation batches. Libraries in batch1 were sequenced twice, once on an Illumina HiScanSQ system and the other on an Illumina HiSeq 2000 system.

Supplementary Tables

Supplementary Table 1. List of prior research work with comparison of RNA-seq and microarray

Study	*Journal*	*Year*	*Experimental design*	*Conclusions*
Marioni et al.	Genome Research	2008	Liver vs kidney	A single lane of Illumina RNA-seq data appears comparable to that in a single array in enabling identification of differentially expressed genes.
Sultan et al.	Science	2008	Human HEK 293T cells and B cells, Illumina HumanRef8 V2.0 BeadChips vs RNA-seq	RNA-seq was more sensitive than microarrays, with genes detected only by RNA-seq being in the lowest range of expression levels.
Fu et al.	BMC Genomics	2009	Using proteomics, microarray, and RNA-seq measure expression of human brain samples	RNA-seq provided more accurate estimation of absolute transcript levels.
Bradford et al.	BMC Genomics	2010	Exon-array vs RNA-seq based on human normal vs cancer tissue	High degree of correspondence between two platforms; RNA-seq more sensitive to detecting differentially expressed exons than the Exon array.
Xiong et al.	Nature Genetics	2010	Comparing RNA-seq and microarray in detecting dosage compensation of the active X-chromosome in human and mouse	RNA-seq is more sensitive than microarray.
Xu et al.	PNAS	2011	Comparing GGH array to RNA-seq based on liver and muscle samples	Both platforms detect similar expression changes at the gene level, the GG-H array is more sensitive at the exon level.
Liu et al.	Nucleic Acids Research	2011	HJAY array vs RNA-seq based on human/chimpanzee/rhesus cerebellum RNA samples	RNA-seq has significantly improved gene coverage and increased sensitivity for differentially expressed genes.
Łabaj et al.	Bioinformatics	2011	human mammary epithelial cell (HMEC) line 184A1, GeneChip Human Gene 1.0 ST Array (Affymetrix) vs RNA-seq, 3 technical replicates	The number of transcripts measured reliably (i.e., with good precision) was higher for microarrays than for RNA-seq, with the strength of the effect depending on data processing.
Su et al.	Chemical Research in Toxicology	2011	Comparing microarray and RNA-seq based on aristolochic acid-treated rat kidneys	RNA-seq was more sensitive in detecting genes with low expression levels, while similar gene expression patterns were observed for both platforms. Moreover, although the overlap of the DEGs was only 40-50%, the biological interpretation was largely consistent between the RNA-seq and microarray data.

Raghavachari et al.	BMC Medical Genomics	2012	Comparing exon-array and RNA-seq on whole blood clinical specimens	Microarrays remain useful and accurate for transcriptomic analysis of clinical samples with low input requirements, while RNA-seq technology complements and extends microarray measurements for novel discoveries.
Nookaew et al.	Nucleic Acids Research	2012	Saccharomyces cerevisiae strain CEN.PK 113-7D, grown under two different conditions (batch and chemostat)	The results obtained based on RNA-seq data were in good agreement with microarray data. The inconsistencies found in DEG identification between RNA-seq and microarrays were shown to be mainly due to genetic variation found on the ORF and on the microarray probes.
Kogenaru et al.	BMC Genomics	2012	Comparing microarray and RNA-seq using HrpX regulome from pathogenic bacteria	RNA-seq and microarray together provide a more comprehensive picture of HrpX regulome. RNA-seq and microarray complement each other in transcriptome profiling.
Van Delft et al.	Toxicological Sciences	2012	HepG2 cells upon exposure to benzo[a]pyrene (BaP)	RNA-seq detects about 20% more genes than microarray-based technology but almost threefold more significantly differentially expressed genes. Functional enrichment analyses show that RNA-seq yields more insight into the biology and mechanisms related to the toxic effects caused by BaP, i.e., two- to five-fold more affected pathways and biological processes.
Sirbu et al.	PLOS ONE	2012	RNA-seq vs dual- and single-channel microarray for time series datasets for the Drosophila melanogaster embryo development	RNA-seq displayed highest sensitivity to differential expression and single-channel arrays performed similarly. Both are more sensitive to the dual-channel array; RNA-seq has an advantage in quantifying extreme transcription levels.
Mooney et al.	PLOS ONE	2013	B-Cell Lymphomas of Canis familiari, GeneChip Canine Genome V2.0 Array (Affymetrix) vs RNA-seq	While RNA-seq gave more sensitive detection of transcripts, a higher percentage of differentially expressed genes was identified by microarrays.
Sekhon et al.	PLOS ONE	2013	Maize gene atlas developed from 11 maize organs by microarray and RNA-seq using the same RNA set	Both technologies produced similar transcriptome profiles (r ranges from 0.70 to 0.83). RNA-seq provided enhanced coverage of the transcriptome (82.1% > 56.5%). RNA-seq provided higher resolution for identifying tissue-specific expression as well as for distinguishing the expression profiles of closely related paralogs as compared to microarray-derived profiles.
Xu et al.	BMC Bioinformatics	2013	Comparing microarray and RNA-seq using 5-azadeoxy-cytidine (5-Aza) treated HT-29 colon cancer cells	RNA-seq has advantages over microarray in identification of DEGs with the most consistent results generated from DESeq and SAM methods.
Merrick et al.	PLOS ONE	2013	Comparing microarray and RNA-seq using 1 ppm aflatoxin B1 treated male rats	RNA-seq detects more DEGs than microarray, and provides additional information on novel transcripts, isoforms, and exons.

Supplementary Table 2. Modes of action and exposure of the chemicals used in the study

SET	MIE/MOA	Chemical Name	Abbreviation	Dose (mg/kg/day)	Duration (days)	Structure Activity Group; Therapeutic indication
TRAINING	AHR	3-METHYLCHOLANTHRENE	3ME	300	5	Toxicant, Ah receptor agonist, DNA alkylator
TRAINING	AHR	BETA-NAPHTHOFLAVONE	NAP	1500	5	Toxicant, Ah receptor agonist
TRAINING SET	AHR	LEFLUNOMIDE	LEF	60	5	Inhibits pyrimidine /purine metabolism, dihydroorotase inhibitor; Antirheumatic Disease Modifying Agent
TRAINING	CAR/PXR	ECONAZOLE	ECO	334	5	Sterol 14-demethylase inhibitor, fluconazole like; Antifungal azole
TRAINING	CAR/PXR	METHIMAZOLE	MET	100	3	Thyroperoxidase inhibitor; Thyroid and Antithyroid Agent
TRAINING	CAR/PXR	PHENOBARBITAL	PHE	54	5	GABA agonist; Antiepileptics / Anticonvulsants
TRAINING	CYTOTOX	CARBON TETRACHLORIDE	CAR	1175	7	Toxicant, free radical generator
TRAINING	CYTOTOX	CHLOROFORM	CHL	600	5	Toxicant, free radical generator
TRAINING	CYTOTOX	THIOACETAMIDE	THI	200	5	Toxicant, free radical generator
TRAINING	DNA DAMAGE	AFLATOXIN B1	AFL	0.3	5	Toxicant, DNA alkylator
TRAINING	DNA DAMAGE	IFOSFAMIDE	IFO	143	3	DNA-alkylator, nitrogen mustard; Antineoplastic
TRAINING	DNA DAMAGE	N-NITROSODIMETHYLAMINE	NIT	10	5	Toxicant, DNA alkylator
TRAINING	PPARA	BEZAFIBRATE	BEZ	617	7	Peroxisome proliferator; Hypolipidemic Agent
TRAINING	PPARA	NAFENOPIN	NAF	338	5	Peroxisome proliferator; Hypolipidemic Agent
TRAINING	PPARA	PIRINIXIC ACID	PIR	364	5	Peroxisome proliferator; Hypolipidemic Agent
TEST SET	CAR/PXR	CLOTRIMAZOLE	CLO	89	5	Sterol 14-demethylase inhibitor; Antifungal azole
TEST SET	CAR/PXR	FLUCONAZOLE	FLU	394	5	Sterol 14-demethylase inhibitor, fluconazole like; Antifungal azole
TEST SET	CAR/PXR	MICONAZOLE	MIC	920	5	Sterol 14-demethylase inhibitor, fluconazole like; Antifungal azole
TEST SET	ER	BETA-ESTRADIOL	BES	150	5	Estrogen receptor agonist, steroidal; Bone Mineral Homeostasis
TEST SET	ER	ETHINYLESTRADIOL	EES	10	5	Estrogen receptor agonist, steroidal; Hormone replacement
TEST SET	ER	NORETHINDRONE	NOR	375	5	Progesterone receptor agonist; Ovulation inhibitor
TEST SET	HMGCOA	CERIVASTATIN	CER	7	5	HMG-CoA reductase inhibitor; Hypolipidemic Agent
TEST SET	HMGCOA	LOVASTATIN	LOV	450	5	HMG-CoA reductase inhibitor, non-aromatic; Hypolipidemic Agent
TEST SET	HMGCOA	SIMVASTATIN	SIM	1200	3	HMG-CoA reductase inhibitor, non-aromatic; Hypolipidemic Agent
TEST SET	PPARA	CLOFIBRIC ACID	CFA	448	5	PPAR alpha agonist, fibric acid; Hypolipidemic Agent
TEST SET	PPARA	GEMFIBROZIL	GEM	700	7	PPAR alpha agonist, fibric acid; Hypolipidemic Agent
TEST SET	PPARA	ROSIGLITAZONE	ROS	1800	5	PPAR gamma agonist, thiazolidinedione, antidiabetic

MOA denotes the mode of action. MIE denotes the molecular initiating event.

Supplementary Table 3. RNA-seq data and mapping status summary based on data analysis pipeline P1

See attached Excel file.

Supplementary Table 4. List of common pathways enriched for each of the MOA chemical groups that are shared by both RNA-seq and microarray platforms

1. PPARA (46)
Acetone d+B3egradation I (to methylglyoxal)
Acute phase response signaling
Acyl-CoA hydrolysis
Androgen biosynthesis
Antigen presentation pathway
Aryl hydrocarbon receptor signaling
Bupropion degradation
Complement system
Dopamine degradation
Estrogen biosynthesis
Ethanol degradation II
Ethanol degradation IV
FXR/RXR activation
Fatty acid activation
Fatty acid α-oxidation
Fatty acid β-oxidation I
Fatty acid β-oxidation III (unsaturated, odd Number)
Glutaryl-CoA degradation
Glycerol-3-phosphate shuttle
Hepatic cholestasis
Histamine degradation
Isoleucine degradation I
LPS/IL-1 mediated inhibition of RXR function
LXR/RXR activation
Melatonin degradation I
Mitochondrial L-carnitine shuttle pathway
NRF2-mediated oxidative stress response
Nicotine degradation II
Nicotine degradation III
Noradrenaline and adrenaline degradation
Oxidative ethanol degradation III
PXR/RXR activation
Phenylalanine degradation IV (mammalian, via side chain)
Putrescine degradation III
Retinoate biosynthesis I
Role of pattern recognition receptors in recognition of bacteria and viruses
Serotonin degradation
Stearate biosynthesis I (animals)
Superpathway of melatonin degradation
Tryptophan degradation III (eukaryotic)

	Tryptophan degradation X (mammalian, via tryptamine)
	Type II diabetes mellitus signaling
	Valine degradation I
	Xenobiotic metabolism signaling
	α-tocopherol degradation
	γ-linolenate biosynthesis II (animals)
2. CAR/PXR (7)	
	Aryl hydrocarbon receptor signaling
	Glutathione-mediated detoxification
	LPS/IL-1 mediated inhibition of RXR function
	NRF2-mediated oxidative stress response
	Nicotine degradation II
	PXR/RXR activation
	Xenobiotic metabolism signaling
3. AhR (10)	
	Acetone degradation I (to methylglyoxal)
	Aryl hydrocarbon receptor signaling
	Bupropion degradation
	LPS/IL-1 mediated inhibition of RXR function
	Melatonin degradation I
	Nicotine degradation II
	Nicotine degradation III
	Retinoate biosynthesis I
	Superpathway of melatonin degradation
	Xenobiotic metabolism signaling
4. Cytotoxic (15)	
	Acetone Degradation I (to Methylglyoxal)
	Bupropion degradation
	Cell cycle: G2/M DNA damage checkpoint regulation
	Citrulline biosynthesis
	Estrogen biosynthesis
	LPS/IL-1 mediated inhibition of RXR function
	Melatonin degradation I
	Methylglyoxal degradation III
	NRF2-mediated oxidative stress response
	Nicotine degradation II
	Pyrimidine ribonucleotides de novo biosynthesis
	Regulation of eIF4 and p70S6K signaling
	Superpathway of melatonin degradation
	Superpathway of methionine degradation
	Xenobiotic metabolism signaling
5. DNA damage (2)	
	Cell cycle: G2/M DNA damage checkpoint regulation
	Xenobiotic metabolism signaling

Supplementary Table 5. Selected 18 Genes for quantitative PCR (qPCR) Validation

Probeset	AceView	Entrez_ID	RefSeq	Gene Name	Function
1369122_at	Bax	24887	Bax	similar to BAX protein, cytoplasmic isoform delta; Bcl2-associated X protein	activates apoptosis
1369180_at	Bcl2l1	24888	Bcl2l1	similar to Bcl2-like 1 isoform 3; Bcl2-like 1	regulates cell death by blocking the voltage-dependent anion channnel
1371027_at	Cblb	171136	Cblb	Cas-Br-M (murine) ecotropic retroviral transforming sequence b	promotes degradation of its substrates by the proteasome
1378550_at	Cblc	292699	Cblc	Cas-Br-M (murine) ecotropic retroviral transforming sequence c	regulates of EGFR mediated signal transduction
1387391_at	Cdkn1a	114851	Cdkn1a	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	regulates cell cycle progression at G1
1387813_at	Erbb2	24337	Erbb2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	involves in transcriptional regulation
1369600_at	Fgf11	170632	Fgf11	fibroblast growth factor 11	involves in embryonic development, cell growth, morphogenesis, tissue repair, tumor growth and invasion
1387640_at	Fgf15	170582	Fgf15	fibroblast growth factor 15	involve in the suppression of bile acid biosynthesis
1369616_at	Fgf22	170579	Fgf22	fibroblast growth factor 22	involves in embryonic development, cell growth, morphogenesis, tissue repair, tumor growth and invasion
1387709_at	Figf	360457	Figf	c-fos induced growth factor	stimulates angiogenesis, lymphangiogenesis and endothelial cell growth
1369468_at	Fzd4	64558	Fzd4	frizzled homolog 4 (Drosophila)	regulates cell growth and differentiation
1370458_at	Hdgfrp3	252941	Hdgfrp3	hepatoma-derived growth factor, related protein 3	enhances DNA synthesis
1377428_at	Lama5	140433	Lama5	laminin, alpha 5	mediates the attachment, migration and organization of cells into tissues during embryonic development
1369218_at	Met	24553	Met	met proto-oncogene	regulates many physiological processes including proliferation, scattering, morphogenesis and survival
1368308_at	Myc	24577	Myc	myelocytomatosis oncogene	regulates transcription of specific target genes
1391559_at	Tlcd1	287472	Tlcd1	TLC domain containing 1	regulates many aspects of growth, development and cellular signaling
1388174_at	Wnt2b	116466	Wnt2b	wingless-type MMTV integration site family, member 2B	regulates cell growth and differentiation
1369248_a_at	Xiap	63879	Xiap	X-linked inhibitor of apoptosis	inhibits apoptosis induced by menadione

Supplementary Table 6. Fold Change and p-value of DEG test in qPCR, microarray and RNA-seq

Chemical	AffyID	AceViewID	Category	isDEG_qPCR*	FC_qPCR†	PValue_qPCR	FC_microarray	PValue_microarray	lowExpr_microarray#	FC_RNA-seq	PValue_RNA-seq	lowExpr_RNA-seq#
NIT	1369122_at	Bax	Both	yes	1.98±0.07	8.90E-06	1.80	2.21E-04		1.93	3.68E-03	
MIC	1369122_at	Bax	Both	yes	2.58±0.22	7.50E-08	2.05	4.81E-05		3.10	8.36E-05	
FLU	1369122_at	Bax	Both	yes	2.62±0.24	3.50E-08	1.62	9.84E-03		2.36	1.14E-03	
AFL	1369122_at	Bax	Both	yes	3.4±0.01	6.70E-07	3.06	3.72E-06		3.45	1.87E-04	
MIC	1378550_at	Cblc	Both	yes	2.25±0.14	4.10E-09	-1.76§	4.67E-04	yes	2.92	8.55E-04	
AFL	1387391_at	Cdkn1a	Both	yes	23.5±0.32	7.70E-09	6.70	1.52E-06		18.42	2.71E-06	
NIT	1387391_at	Cdkn1a	Both	yes	8.07±0.23	3.80E-05	5.31	5.45E-06		8.42	6.97E-04	
NIT	1387709_at	Figf	Both	yes	10.27±0.98	5.30E-06	1.85	9.98E-03	yes	9.93	1.69E-03	
NIT	1369468_at	Fzd4	Both	yes	-2.07±0.25	3.00E-05	-1.69	1.12E-03		-1.81	7.75E-03	
GEM	1369468_at	Fzd4	Both	yes	-2.65±0.45	1.50E-04	-2.21	7.36E-05		-2.47	9.15E-04	
MIC	1369468_at	Fzd4	Both	yes	-2.97±0.16	4.00E-06	-2.43	2.05E-04		-2.56	6.80E-04	
IFO	1369468_at	Fzd4	Both	yes	-2.99±0.01	0	-2.41	1.21E-04		-2.74	6.46E-03	
CFA	1369468_at	Fzd4	Both	yes	-3.5±0.32	6.06E-06	-2.20	2.88E-05		-2.86	6.22E-04	
CFA	1369218_at	Met	Both	yes	-1.83±0.18	3.94E-06	-2.08	1.87E-03		-1.50	2.54E-02	
AFL	1369218_at	Met	Both	yes	-1.9±0.13	2.30E-04	-2.93	1.76E-05		-2.44	4.25E-03	
MIC	1369218_at	Met	Both	yes	-2.56±0.49	2.20E-05	-3.18	1.31E-04		-2.37	7.44E-04	
AFL	1368308_at	Myc	Both	yes	2.1±0.94	3.20E-02	2.23	1.17E-02		2.38	3.06E-02	
FLU	1368308_at	Myc	Both	yes	2.24±0.12	5.80E-05	2.12	8.50E-03		2.41	3.63E-03	
GEM	1368308_at	Myc	Both	yes	-3.81±0.08	7.60E-08	-3.40	1.14E-04		-2.93	9.85E-04	
FLU	1391559_at	Tlcd1	Both	yes	3.81±0.12	1.70E-07	1.53	1.55E-02		3.84	5.53E-05	
GEM	1391559_at	Tlcd1	Both	yes	7.37±0.68	2.00E-06	3.02	2.81E-04		9.17	1.26E-05	
IFO	1369248_a_at	Xiap	Both		-1.37±0.07	7.30E-02	-2.24	4.41E-04		-1.97	4.80E-02	

FLU	1369248_a_at	Xiap	Both	yes	-1.68±0.14	5.60E-06	-2.57	2.62E-03		-2.00	7.79E-04	
CFA	1369248_a_at	Xiap	Both	yes	-2.1±0.25	4.14E-07	-3.07	4.83E-06		-2.45	6.46E-03	
ROS	1369248_a_at	Xiap	Both	yes	-2.51±0.13	1.50E-10	-4.31	3.60E-07		-2.82	1.30E-04	
MIC	1369180_at	Bcl2l1	AFX		1.03±0.06	3.90E-01	-1.62	8.75E-04	yes	1.25	1.26E-01	
IFO	1387391_at	Cdkn1a	AFX		-1.79±0.14	1.38E-01	-1.91	4.69E-03	yes	-1.26	5.78E-01	
FLU	1387813_at	Erbb2	AFX		0.88±0.17	3.00E-02	-1.59	4.00E-03	yes	-1.81	8.06E-02	yes
ROS	1369218_at	Met	AFX	yes	-1.74±0.08	1.20E-06	-2.19	5.63E-03		-1.45	2.12E-02	
MIC	1388174_at	Wnt2b	AFX		-1.06±1.24	8.70E-01	-2.07	1.35E-03	yes	NA	NA	yes
AFL	1369248_a_at	Xiap	AFX		1.1±0.07	5.96E-01	-1.85	4.77E-03		-1.75	1.03E-01	
GEM	1369248_a_at	Xiap	AFX		-1.44±0.41	5.60E-03	-1.77	3.64E-02		-1.40	3.39E-02	
AFL	1369180_at	Bcl2l1	NGS	yes	1.9±0.03	2.00E-04	1.10	5.40E-01	yes	1.86	2.52E-02	
NIT	1371027_at	Cblb	NGS		1.37±0.2	7.60E-04	1.45	6.65E-02		1.53	9.04E-03	
ROS	1387391_at	Cdkn1a	NGS	yes	-4.54±1.33	5.10E-03	-1.89	9.65E-02		-5.17	3.64E-02	
AFL	1387813_at	Erbb2	NGS	yes	1.6±0.2	3.40E-05	1.07	6.99E-01	yes	2.08	1.09E-02	yes
AFL	1369600_at	Fgf11	NGS	yes	-1.7±0.1	5.30E-04	1.12	4.84E-01	yes	-1.89	2.45E-02	yes
CFA	1369600_at	Fgf11	NGS	yes	-2.89±0.33	4.87E-05	-1.09	4.54E-01	yes	-1.71	4.09E-02	yes
ROS	1369600_at	Fgf11	NGS	yes	-6.1±0.57	3.20E-06	-1.12	3.10E-01	yes	-4.10	6.05E-05	yes
AFL	1387709_at	Figf	NGS	yes	1.8±0.3	1.90E-04	1.12	4.62E-01	yes	2.81	1.91E-03	yes
MIC	1387709_at	Figf	NGS	yes	4.42±0.49	3.50E-07	-1.04	7.84E-01	yes	4.86	5.27E-05	yes
AFL	1369468_at	Fzd4	NGS	yes	-1.8±0.08	1.50E-02	-1.56	1.36E-01		-2.34	3.76E-02	
NIT	1370458_at	Hdgfrp3	NGS		-1.27±0.34	8.60E-09	1.43	1.62E-02	yes	4.24	1.18E-04	yes
AFL	1370458_at	Hdgfrp3	NGS	yes	4.3±0.3	6.90E-08	1.32	9.44E-02	yes	4.92	6.85E-05	yes
AFL	1377428_at	Lama5	NGS	yes	17.5±0.6	6.30E-07	1.09	5.76E-01	yes	11.05	1.42E-05	yes
CFA	1391559_at	Tlcd1	NGS	yes	2.08±0.49	1.31E-03	1.21	2.27E-01	yes	2.82	3.71E-04	
ROS	1391559_at	Tlcd1	NGS	yes	2.15±0.64	2.80E-03	1.27	2.14E-01	yes	3.18	1.69E-03	

AFL	**1371027_at**	**Cblb**	**None**		**1.3±0.13**	**1.48E-01**	**1.42**	**1.59E-01**		**1.10**	**7.64E-01**	
AFL	**1378550_at**	**Cblc**	**None**		**1.3±0.5**	**1.35E-01**	**1.30**	**9.90E-02**	**yes**	**1.04**	**8.83E-01**	**yes**
CFA	**1387391_at**	**Cdkn1a**	**None**		**1.26±0.32**	**4.10E-01**	**1.67**	**1.98E-01**		**1.24**	**4.91E-01**	
GEM	**1387391_at**	**Cdkn1a**	**None**		**-1.44±0.63**	**2.50E-01**	**-1.32**	**4.30E-01**		**-2.21**	**1.14E-01**	
IFO	**1387813_at**	**Erbb2**	**None**		**-1.18±0.21**	**2.12E-01**	**-1.10**	**6.78E-01**	**yes**	**1.62**	**1.23E-01**	**yes**
NIT	**1387813_at**	**Erbb2**	**None**	**yes**	**1.9±0.62**	**4.60E-03**	**-1.30**	**1.03E-01**	**yes**	**1.20**	**4.57E-01**	**yes**
FLU	**1369600_at**	**Fgf11**	**None**	**yes**	**-7.3±1.6**	**8.20E-04**	**-1.26**	**7.31E-02**	**yes**	**NA**	**NA**	**yes**
AFL	**1387640_at**	**Fgf15**	**None**		**Undetermined**		**1.16**	**3.47E-01**	**yes**	**NA**	**NA**	**yes**
IFO	**1387640_at**	**Fgf15**	**None**		**Undetermined**		**-1.20**	**4.16E-01**	**yes**	**NA**	**NA**	**yes**
AFL	**1369616_at**	**Fgf22**	**None**		**Undetermined**		**1.08**	**6.48E-01**	**yes**	**1.10**	**8.08E-01**	**yes**
IFO	**1387709_at**	**Figf**	**None**		**1.32±0.04**	**4.10E-02**	**-1.22**	**3.84E-01**	**yes**	**1.58**	**2.26E-01**	**yes**
ROS	**1369468_at**	**Fzd4**	**None**	**yes**	**-1.64±0.05**	**2.20E-03**	**-1.19**	**1.72E-01**		**-1.10**	**5.19E-01**	
IFO	**1370458_at**	**Hdgfrp3**	**None**		**1.32±0.85**	**8.51E-01**	**-1.08**	**7.40E-01**	**yes**	**1.74**	**1.02E-01**	**yes**
FLU	**1370458_at**	**Hdgfrp3**	**None**	**yes**	**2.23±0.99**	**1.10E-02**	**1.04**	**7.36E-01**	**yes**	**2.13**	**1.10E-01**	**yes**
MIC	**1368308_at**	**Myc**	**None**		**1.06±0.13**	**6.30E-01**	**1.25**	**3.40E-01**		**1.25**	**3.10E-01**	
ROS	**1368308_at**	**Myc**	**None**		**1.28±0.97**	**3.70E-01**	**1.28**	**5.07E-01**		**1.75**	**2.04E-01**	
AFL	**1391559_at**	**Tlcd1**	**None**	**yes**	**-1.7±0.2**	**1.00E-04**	**-1.03**	**8.54E-01**	**yes**	**-1.67**	**7.56E-02**	**yes**
AFL	**1388174_at**	**Wnt2b**	**None**		**-1.3±0.65**	**2.81E-01**	**1.26**	**1.53E-01**	**yes**	**NA**	**NA**	**yes**

*Genes or probesets are determined as DEG if absolute Fold Change >1.5 and P-Value <0.05.

[†]Positive Fold Change (FC) indicates up-regulation; negative FC indicates down-regulation.

[#]Genes or probesets are considered as “lowly expressed” if the average expression levels are below median expression of all genes (or probesets) for both treated samples and control samples.

[§]The fold change direction was opposite to those of qPCR and RNA-seq.

Supplementary Table 7. Summary information for classifier development

Methodology	RNA-Seq data	AFX data	Endpoint	Feature selection	Prediction method
M1	Gene index from pipeline-1; All indexes less than 8 are set to 8; All genes	MAS5 normalized data	(1) agents (2) MOAs	None	Briefly, each individual is represented by a vector with gene index as coordinates, and the cosine distance based on the genes with index over 8 in at least one of two vectors is used to compare individuals. 18 groups of rats in training set are selected and their average vectors are computed. Each individual rat is then compared to these groups and the cosines are plotted for the 63 individuals on the x axis. Each agent is colored according to the MOA. To classify the blinded test set, the 48 new runs were similarly compared to the 18 groups.
M2	RPM normalized data from pipeline-3. The mean value of the control samples were calculated. Then the ratio values of the other samples were calculated. The genes with any control mean value equals “0” were filtered out. The zero value was replaced by a smaller value of the minimum (min/1.2). Afterward, the data set were transformed using log2.	MAS5 normalized data; Similarly, the mean intensities of the matched control samples were calculated first and were then used to calculate the ratios for the treated samples by dividing the intensities of the treated samples using the mean intensities of the matched controls	MOAs	Genes were ranked by Information Gain approaches and candidate genes were then selected from the top 22 genes with a step size of 1	The RBF SVM was used. The parameters C and γ were determined with grid searching. The average accuracy of 10 times 5-fold cross validation was calculated as the evaluation criteria. The grid ranges for C and γ were from -5 to 15 and from -15 to 5 respectively with a step of 2. Because the prediction is not binary, 5 binary SVM models were built and each was associated with one MOA and a possibility ranging from 0 to 1. Final determination was made based on the 5 probabilities as follow: Pred =[p1, p2, p3, p4, p5]; MOApred = MOAi if pi = Max(Pred) and pi > 0.5 or MOApred = new MOA if Max(Pred) ≤ 0.5)
M3	RPM normalized data from pipeline-2; Ratio data were generated by comparing treated samples to matched controls.	MAS5 normalized data were used; Ratio data were generated by comparing treated samples to matched controls.	MOAs	t-test was conducted for each MOA: using 9 treated samples vs. 12 samples of vehicle matched controls and identified signatures for both NGS and Affy, respectively; A)top 100 P value ranking; B)top 100 fold change ranking with P<0.01; C)The common gene signature shared between NGS and Affy.	KNN in ArrayTrack was used for classifiers’ trainings and predictions. The training model was built one MOA at a time for PPARA and Car/PXR, respectively. Each model was based on 45 training samples, divided into two groups: group 1 (36 samples) and group 2 (9 samples of the specific MOA) with K=5 External validations were done using the 36 sample unknown testing set
M4	Gene index from pipeline-1; Ratio data were generated by comparing treated samples to matched controls	MAS5 normalized data were used and ratio data also genereated	MOAs	Using "sequential MOA removal ANOVA based voting" to select features. Briefly, ANOVA test was done first for all five MOAs in the training set to derive a list of MOA differentiating gene signatures. PPARA, Cytotoxicity, and DNA Damage were then sequentially removed and ANOVA tests were done at each step. R/limma was adopted for per-agent DEG test with p-value threshold of 0.05 and fold change cut-off of 1.5	The test samples were classified by a visual reading of the Hierarchical Clustering Analysis (HCA) and Principal Component Analysis (PCA) of both training and test samples based on those genes selected by all 4 ANOVA tests were used to classify the test samples.
M5	RPKM data from pipeline-6	MAS5 normalized data	MOAs	Genes were ranked in terms of importance (Robnik-Sikonja and Kononenko, 2003) and sequentially selected to built kNN models which were assessed with Matthews Correlation Coefficient (MCC).	A general graph-based semi-supervised learning (GGSSL) algorithm with novel class discovery (Nie et al., 2010) was adopted for prediction and novelty detection. GGSSL training was based on 20 random subsampling (Monte Carlo) runs on the training data set, stratified over the five classes, splitting the data into internal training and internal test sets (75-25% training-test proportion)

Supplementary Table 8. MOA prediction results

Group	Predictor	OVERALL*	MOA1*	MOA2*	UNKNOWN*	Platform
M6 - Round 1	A-Model-1	50	44	22	67	microarray
M6 - Round 1	A-Model-2	22	44	22	11	microarray
M6 - Round 1	N-Model-1	44	22	22	67	RNA-seq
M6 - Round 1	N-Model-2	58	67	56	56	RNA-seq
M6 - Round 2	A-Model-1	50	44	22	67	microarray
M6 - Round 2	A-Model-2	22	44	22	11	microarray
M6 - Round 2	A-Model-3	50	11	22	83	microarray
M6 - Round 2	N-Model-1	44	22	22	67	RNA-seq
M6 - Round 2	N-Model-2	58	67	56	56	RNA-seq
M6 - Round 2	N-Model-3	42	22	33	56	RNA-seq
M5 - Round 1	N-FPKM1	75	67	33	100	RNA-seq
M5 - Round 1	N-FPKM2	61	44	0	100	RNA-seq
M5 - Round 1	N-RLE	69	100	0	89	RNA-seq
M5 - Round 2	N-genes1	69	67	11	100	RNA-seq
M5 - Round 2	N-genes2	67	67	0	100	RNA-seq
M5 - Round 2	N-isoforms	61	44	0	100	RNA-seq
M5 - Round 2	A-AFFY	67	78	33	78	microarray
M5 - Round 2	A-AFFY2	69	78	44	78	microarray
M5 - Round 2	N-Aceview	53	0	33	89	RNA-seq
M5 - Round 2	N-FPKM	56	0	56	83	RNA-seq
M5 - Round 2	N-FPKM2	58	0	56	89	RNA-seq
M5 - Round 2	N-FPKM3	56	0	56	83	RNA-seq
M5 - Round 2	N-FPKM4	58	0	56	89	RNA-seq
M5 - Round 2	N-RLE	69	44	33	100	RNA-seq
M5 - Round 2	N-RLE2	69	44	33	100	RNA-seq
M5 - Round 2	N-RLE3	69	44	33	100	RNA-seq
M5 - Round 2	N-RLE4	69	44	33	100	RNA-seq
M5 - Round 3	A-GGSSL_100probes_Affy	75	67	44	94	microarray
M5 - Round 3	A-GGSSL_400probes_Affy	75	67	33	100	microarray
M5 - Round 3	A-GGSSL_70probes_Affy-hook-quantile-farms	72	67	33	94	microarray
M5 - Round 3	N-GGSSL_200probes_200genes_Aceview	72	67	33	94	RNA-seq
M5 - Round 3	N-GGSSL_400genes_Aceview	56	0	22	100	RNA-seq
M5 - Round 3	N-GGSSL_5000probes_Aceview	72	56	33	100	RNA-seq

M5 - Round 3	N-GGSSL_ALL_Aceview	67	67	0	100	RNA-seq
M1	Oct19	50	100	0	50	RNA-seq
M1	Oct23	100	100	100	100	RNA-seq
M3	A-EPIG_221_probes_KNN_5_neighbors_prediction	67	67	44	78	microarray
M3	A-EPIG_n221_KNN_5_neighbors_Model	58	67	44	61	microarray
M4	A-Affy_pred_Model1	86	100	67	89	microarray
M4	A-Affy_pred_Model2	78	100	44	83	microarray
M4	N-NGS_pred_Model-1	58	67	0	83	RNA-seq
M4	N-NGS_pred_Model-2	69	67	33	89	RNA-seq
M2 - Round 1	A-Model1	39	33	22	50	microarray
M2 - Round 1	A-Model2	44	0	33	72	microarray
M2 - Round 1	N-Model1	44	0	11	83	RNA-seq
M2 - Round 1	N-Model2	44	0	22	78	RNA-seq
M2 - Round 2	A-Model1	53	33	56	61	microarray
M2 - Round 2	A-Model2	39	33	44	39	microarray
M2 - Round 2	N-Model-1	64	56	56	72	RNA-seq
M2 - Round 2	N-Model-2	64	67	56	67	RNA-seq
M3 - Round 2	A-Best_Model_Affy_Car topF100_PPARA_comm8F	72	67	56	83	microarray
M3 - Round 2	A-Comm_topF_Affy_Ratio_Car131REF_PPARA8	69	67	44	83	microarray
M3 - Round 2	A-Comm_topF_Affy_Row_Car13REF_PPARA8	53	67	56	44	microarray
M3 - Round 2	A-Top100F_Affy	67	33	56	89	microarray
M3 - Round 2	A-Top100P_Affy	64	44	33	89	microarray
M3 - Round 2	N-Best_Model_NGS_Car_topF100_PPARA_comm8F	58	56	56	61	RNA-seq
M3 - Round 2	N-Comm_top100F_NGS_Ratio_Car131REF_PPARA8	53	78	67	33	RNA-seq
M3 - Round 2	N-Comm_top100F_NGS_Raw_Car131REF_PPARA8	64	67	67	61	RNA-seq
M3 - Round 2	N-Top100F_NGS	56	44	56	61	RNA-seq
M3 - Round 2	N-Top100P_NGS	67	22	100	72	RNA-seq
M7	N-prediction	36	56	11	39	RNA-seq

*Prediction accuracy is in percentage (%).

Supplementary Table 9. Rank-Base scoring of the RNA-seq signatures using NextBio

RNA-seq DEG List (Chemical)[1]	# of DEGs Employed in Analysis[2]	Top Scoring Microarray Signature From DrugMatrix Affymetrix Liver Experiments[3]	Correlation/Direction (Top Scoring Signature)[4]	P-Value (Top Scoring Signature)[5]	Rank of Corresponding Microarray Study (rank in 657 Studies)[6]	Correlation/Direction (Corresponding Microarray Study)[7]	P-Value (Corresponding Miccroarray Study)[8]
3ME	578	Liver of rats + 3-METHYLCHOLANTHRENE at 300mg-kg-d in CMC by oral gavage 5d_vs_vehicle	+	2.50E-53	1	+	2.50E-53
AFL	1939	Liver of rats + AFLATOXIN B1 at .3mg-kg-d in CMC by oral gavage 5d_vs_vehicle	+	7.00E-162	1	+	7.00E-162
BEZ	3942	Liver of rats + NAFENOPIN at 338mg-kg-d in corn oil by oral gavage 5d_vs_vehicle	+	4.9E-324	6	+	1.80E-279
CAR	5128	Liver of rats + MICONAZOLE at 920mg-kg-d in corn oil by oral gavage 5d_vs_vehicle	+	6.70E-256	6	+	6.00E-234
CHL	3850	Liver of rats + MICONAZOLE at 920mg-kg-d in corn oil by oral gavage 3d_vs_vehicle	+	6.30E-245	3	+	2.60E-223
ECO	2342	Liver of rats + ECONAZOLE at 334mg-kg-d in corn oil by oral gavage 5d_vs_vehicle	+	7.40E-157	1	+	7.40E-157
IFO	1130	Liver of rats + DOXORUBICIN at 3mg-kg-d in saline by intraveneous 3d_vs_vehicle	+	4.30E-64	4	+	1.10E-40
LEF	3128	Liver of rats + LEFLUNOMIDE at 60mg-kg-d in corn oil by oral gavage 5d_vs_vehicle	+	7.70E-210	1	+	7.70E-210
MET	1567	Liver of rats + METHIMAZOLE at 100mg-kg-d in water by oral gavage 3d_vs_vehicle	+	5.90E-89	1	+	5.90E-89
NAF	5340	Liver of rats + BEZAFIBRATE at 617mg-kg-d in corn oil by oral gavage 3d_vs_vehicle	+	4.9E-324	3	+	4.9E-324
NAP	422	Liver of rats + BETA-NAPHTHOFLAVONE at 1500mg-kg-d in CMC by oral gavage 5d_vs_vehicle	+	4.60E-26	1	+	4.60E-26
NIT	5532	Liver of rats + N-NITROSODIMETHYLAMINE at 10mg-kg-d in saline by intraperitoneal 5d_vs_vehicle	+	6.90E-269	1	+	6.90E-269
PHE	787	Liver of rats + PHENOBARBITAL at 54mg-kg-d in water by oral gavage 5d_vs_vehicle	+	9.20E-75	1	+	9.20E-75
PIR	3382	Liver of rats + NAFENOPIN at 338mg-kg-d in corn oil by oral gavage 5d_vs_vehicle	+	4.9E-324	2	+	4.9E-324
THI	5006	Liver of rats + THIOACETAMIDE at 200mg-kg-d in saline by intraperitoneal 5d_vs_vehicle	+	4.40E-288	1	+	4.40E-288

[1]RNA-seq DEG List (Chemical) - Gene list identifier. List used for query was provided by RNA-seq data analysis pipeline P1.
[2]# of DEGs Employed in Analysis - Number of DEG features recognized by NextBio that were employed in the analysis. Note not all identified DEGs from the RNA-seq analysis were using the rank-based scoring. Only those genes recognized by NextBio were considered.

[3]Top Scoring Microarray Signature From DrugMatrix Affymetrix Liver Experiments - RNA-seq signatures were queried against the DrugMatrix Affymetrix Liver microarray signatures found in NextBio. Scoring in done by a running fisher test. The gene list found in this column is the top scoring gene list (of the 657 found in the DrugMatrix Affymetrix liver study set). Note: Distinct sets of control samples were used to generate the DEGs lists from NextBio and the SEQC RNA-seq study presented in this manuscript. While this is not ideal, we believe this further strengthens the rank-based concordance findings presented in the table.
[4]Correlation/Direction (Top Scoring Signature) - Directionality of the correlation between RNA-seq and top scoring Affymetrix gene list
[5]P-Value (Top Scoring Signature) - Running fisher P-value top scoring Affymetrix gene list
[6]Rank of Corresponding microarray study (rank in 657 Studies) - The rank of the corresponding (to samples used to generate RNA-seq data) microarray study gene list. Rank is determined by running fisher P-value. In nearly half of the cases the top scoring gene list was also the corresponding microarray study gene list
[7]Correlation/Direction (Corresponding microarray study) - Directionality of the correlation between RNA-seq and corresponding Affymetrix gene list
[8]P-Value (Corresponding microarray study) - Running fisher P-value for corresponding Affymetrix gene list

Supplementary Table 10. List of transcripts with shortened 3’ UTRs detected from the samples treated by chemicals PHE and PIR

See attached Excel file.

Supplementary Table 11. List of differentially spliced isoforms detected in samples treated by chemicals PHE and PIR

See attached Excel file.

Supplementary Table 12. Transcription regulator signaling

TRID	TR	Microarray # of DSTs	Microarray GCS	Microarray p-value	RNA-Seq	RNA-Seq	RNA-Seq
T08991	PPARalpha	2	0.49	4.60E-03	2	0.76	1.00E-04
T14640	RelA-p65:NF-	2	0.29	3.81E-02	2	0.68	4.00E-04
T02453	Sp3	31	55.36	3.00E-04	23	28.86	8.00E-04
T15084	C/EBPdelta	2	0.54	2.70E-03	2	0.59	1.30E-03
T04747	E2F	2	0.62	1.10E-03	2	0.61	1.50E-03
T00459	C/EBPbeta(LAP)	6	2.97	3.90E-03	6	3.09	2.30E-03
T00991	PPARalpha	5	1.32	6.33E-02	4	1.55	3.10E-03
T05017	HNF-4	8	3.42	3.51E-02	5	2.06	4.80E-03
T15086	rno-miR-1	3	0.65	2.80E-02	3	0.93	5.30E-03
T00754	Sp1	38	58.69	8.42E-02	22	24.19	6.00E-03
T14594	ChREBP	4	1.76	1.60E-03	3	0.84	6.40E-03
T08475	GR-alpha	10	3.35	3.71E-01	7	3.33	6.50E-03
T00371	HNF3A	11	7.07	9.90E-03	11	6.23	2.05E-02
T00369	HNF-1alpha	6	1.29	2.55E-01	5	1.50	2.80E-02
T08153	C/EBPbeta	17	12.31	9.19E-02	15	10.33	3.22E-02
T13803	NF-1	6	1.53	1.44E-01	5	1.42	3.40E-02
T15509	c-Fos:c-Jun	3	0.50	6.58E-02	2	0.28	3.43E-02
T00084	CBF(2)	3	0.31	2.19E-01	3	0.54	3.87E-02
T16461	CPBP	3	1.44	1.00E-04	2	0.25	4.56E-02
T02200	Fra-2	5	0.65	4.85E-01	3	0.52	4.71E-02
T00108	C/EBPalpha	7	2.88	2.39E-02	6	1.82	4.73E-02
T00372	HNF-4alpha1	6	1.74	8.06E-02	5	1.30	4.77E-02
T00875	USF1	7	1.66	3.23E-01	4	0.86	4.93E-02
T02115	USF2	11	6.15	3.39E-02	9	3.62	5.97E-02
T08191	NF-1C	3	0.60	3.86E-02	3	0.44	7.41E-02
T02331	NF-1A	4	1.18	1.73E-02	4	0.63	1.32E-01
T03882	Smad4	2	0.78	1.00E-04	1	0	1
T02304	MAZ	2	0.29	3.37E-02	0	0	1

TR signifies transcript regulator, DST is downstream target genes and GCS is the group correlation score.

Supplementary Table 13. Cross platform concordance of DEGs in the common set of genes tested at different sequencing depths

Sequence Depth (M)	number of DEGs in whole set*	number of DEGs in common set*	number of overlapped DEGs†	Concordance#
28.7±4.9	4973	1773	410	33.6%
65.3±16.4	5170	1865	418	33.0%
94.0±16.9	5456	1929	417	32.0%
119.4±17.0	5628	1928	419	32.2%

* The DEG analysis was performed for the chemical AFL using edgeR in the whole set of AceView genes for pipeline P1 with the thresholds (P-value < 0.05 and absolute fold change > 1.5). The DEGs among the common set (see Online Methods) were then extracted and counted. In comparison, the DEG analysis for microarray data (RMA normalization) was conducted using limma and the same thresholds, which resulted in 559 DEGs among the common set of 13079 genes.

† The number of overlapped DEGs was counted by comparing DEGs from both platforms among the common set of genes with the same directionality of fold change.

The concordance was adjusted against random chance. The computation of concordance and its adjustment were explained in Online Methods.

Supplementary Table 14. Master table for mapping Affymetrix probesets to RNA-seq gene annotations

See attached Excel file.

Supplementary Notes

Supplementary Note 1. RNA-seq data analysis pipelines

RNA-seq data analysis – pipeline 1: Alignment and quantification of RNA-seq data using the Magic pipeline

The bulk of analyses were done with this pipeline.

The Magic-pipeline, developed at NCBI, includes QC, alignment, annotation, quantification, and normalization. Read pairs are treated as single objects, using the Fastc modification (ATGNC><CCTATTT) of the fasta format, and sorted alphabetically, merging and counting identical sequences. The mapping strategy is to hash a dense set of seeds of length 16 bases and scan the targets using a fast seed-extension automaton allowing for insertions deletions and auto-adaptable to the mismatch profile of the platform. The program maps RNA discontinuously on the genome, discovering introns/exon junctions, structural rearrangements or, simply, polyA addition sites and adaptors, which are recognized in the unaligned overhanging ends of reads, clipped dynamically, and annotated.

Magic maps the reads in parallel on all targets: the genome (RGSC v3.4, identical to Rn4 for the main chromosomes), the RefSeq and AceView 2008 gene models, the mitochondrial genes, the rRNA genes (manually constructed from multiple GenBank accessions, in the absence of a RefSeq), the spike-in sequences, and finally the imaginary genome as a mapping specificity control. The imaginary whole genome is constructed by complementing the bases (exchange A:T and G:C), but not reversing the order. It has exactly the same statistics as the genome but is completely alien: hits to this target are false positives and used to tune the quality thresholds.

For RNA-seq, each alignment is scored by adding one point per base aligned and removing eight points per mismatch (base substitution, insertion or deletion). The length of the aligned region is then recursively clipped to optimize this score. As a result, no mismatch is reported closer than 8 bases to the edge of the aligned segment,

allowing reliable recognition of substitutions or indels and precise cooperative discovery of exon-intron boundaries. For each read pair, alignments on all targets, genes or genome, are scored, and only the best scores are kept, giving preference to spliced over unspliced targets in order to disfavor pseudogenes. Global quality filters are then applied, tolerant to mismatches in full length, unique alignments. In the recommended quasi-unique mode, read pairs mapping over their entire length and equally well to several genes or genomic loci, up to 9 sites, are assigned in Bayesian proportion seeded on the number of reads uniquely mapping to these loci. Reads mapping to several alternative transcripts of the same gene are kept in each, but contribute only once to the gene count.

Magic exports a richer and more explicit alignment format than Bam/Sam, so that the downstream tracking of mismatches, SNPs and structural rearrangements is simplified. Analysis is faster than data generation time: typically, RNAs are aligned on all targets using around 3 to 6 hours of CPU, 5 to 8 GB of RAM and 2 gigabytes of disk space per gigabase of sequence. It is available on the AceView NCBI website www.aceview.org in the Downloads, Software tab, Magic code (ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic)

The Magic normalized RNA-seq index of expression is directly comparable to a normalized microarray logarithmic luminosity. It is independent of the size of the experiment and factors in the quality of the libraries, the insert length and the noise level. The index is computed as

Index = $\log_2 (z + \sqrt[2]{4 + z^2}) - 1$, with $z = 10^{12}\, n/N'l'$

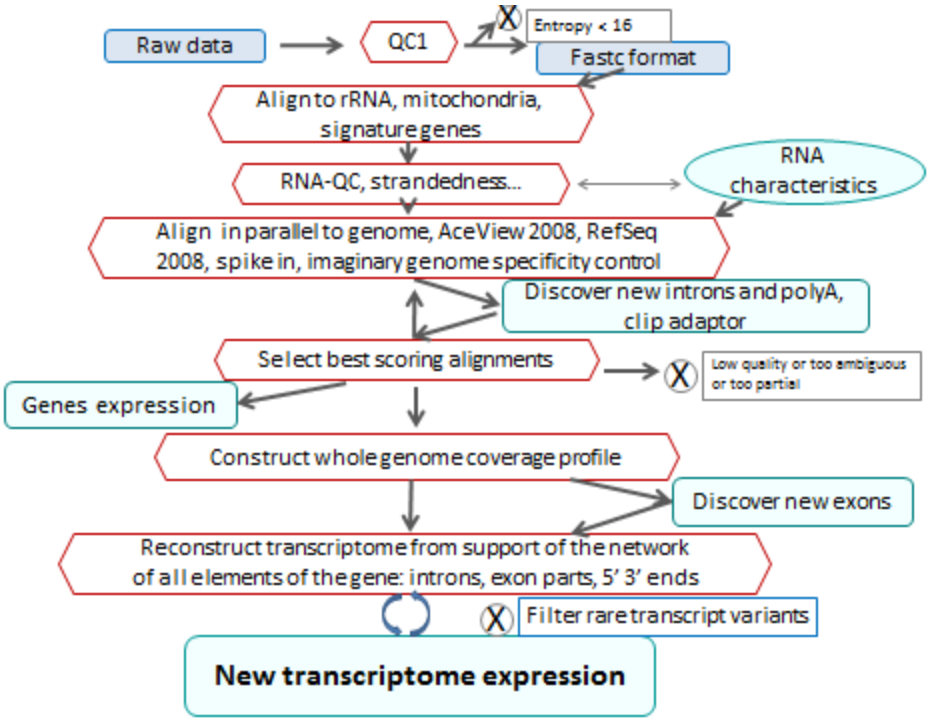
- z estimates the coverage of the gene per terabases aligned
- n is the number of bases aligned in the gene
- $N' /10^{12}$ is the corrected number of terabases aligned to the known transcriptome, after excluding the ribosomal and mitochondrial genes and the genes gathering more than 2% of the total number of bases aligned, to

compensate an apparent variation of the index of all other genes in the presence of highly-variable extremely-expressed genes, such as albumin in liver or hemoglobins in blood.

- l' is the length of the gene effectively sequenceable in the particular experiment. First the coverage of very long ubiquitous genes (such as EEF2 (3.5kb), FLNA (8.5 kb), MALAT1 (7kb), AHNAK (18 kb) or NEAT (22 kb)) is measured and used as a maximal gene length. This length is often around 3 kb, sometimes much longer, sometimes below 1 kb; for a polyA selection protocol, it corresponds mainly to the 3' bias. Second, the average length λ of the insert is measured from the aligned read pairs, and subtracted from the length of the main RNA l. Indeed (l – λ) is the number of possible positions of the insert in the gene; it is set to at least λ.The latter correction matters for short genes, under-represented in libraries with long insert size.

For genes that are not 'significantly expressed', i.e. with less than 4 reads above the experimental noise, the index is given as NA/value, indicating low precision due to high sampling error, and the value is interpolated between the lowest callable index (average 7.8+-1.3 in rat TGx, with 40.4 million reads aligned per run on average) and 0. The noise is evaluated as the fraction of truly intergenic reads after transcriptome reconstruction, it is mainly genomic contamination. At Charles Wang City of Hope laboratory, the noise is low: it varies between 1% and 2%.

The expression index and raw read count tables (dated 20120831) for the AceView/RefSeq 2008 genes are sorted by descending maximal index of expression across all rats. The first columns contain metadata connecting AceView gene to RefSeq and GeneID, as well as the Affymetrix expression array probesets and the rat Rn4 genomic coordinates. The detailed pipeline is illustrated in the figure below. The code is available at ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic.

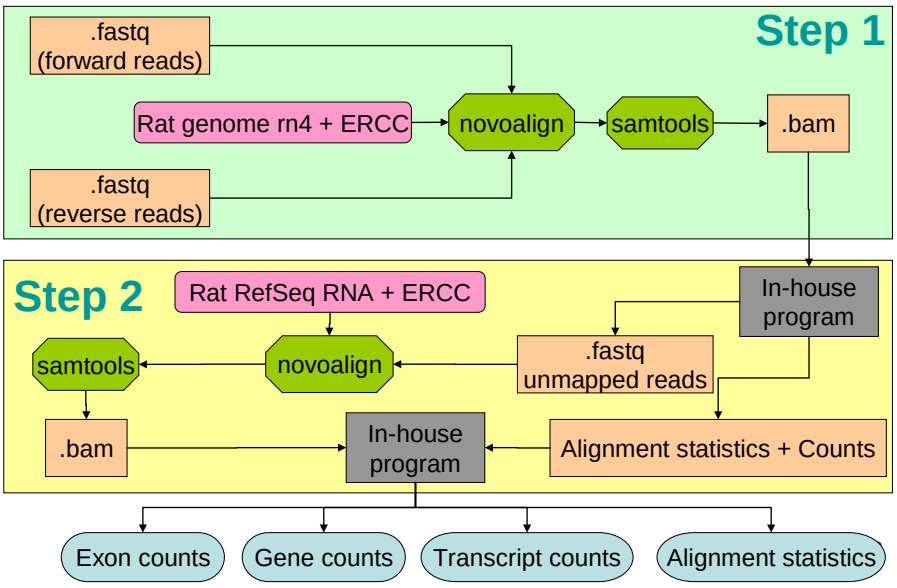


RNA-seq data analysis 2: pipeline-2

This is a two-step alignment pipeline illustrated in the figure below for the quantification and normalization of the RNA-seq data set. In the first step, raw reads are aligned with Novoalign v2.08.01 (www.novocraft.com) against rat genome rn4 downloaded from the UCSC ftp server (ftp://hgdownload.cse.ucsc.edu/goldenPath/rn4) plus 92 ERCC sequences. The intermediate bam files generated with Novoalign are parsed through an in-house program to summarize mapping results. The unmapped reads are then passed to Novoalign again and are mapped to rat RefSeq RNAs (release version 52, March 5, 2012) downloaded from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/refseq) and 92 ERCC sequences. The alignment results are parsed through an in-house program and are merged with the results generated in step one. Finally, four types of outputs including counts at exon, gene, and transcript levels and some statistic measurements are generated. The raw read count for genes, exons, or transcripts is normalized and transformed into log2 reads per million (RPM) with a global scaling method as equation:

$$\log_2(RPM) = \log_2[\frac{(C_{raw}+1)\times 1000000}{N_{G+T}}]$$

where C_{raw} is the raw read count of a gene, exon, or transcript; N_{G+T} is the total number of reads mapping either to the genome or RefSeq RNAs (without counting the reads mapping to ERCCs). Accordingly, read counts for ERCCs are normalized and transformed using the same equation with a different denominator of the total number of reads mapping to ERCCs.



RNA-seq data analysis 3: pipeline-3

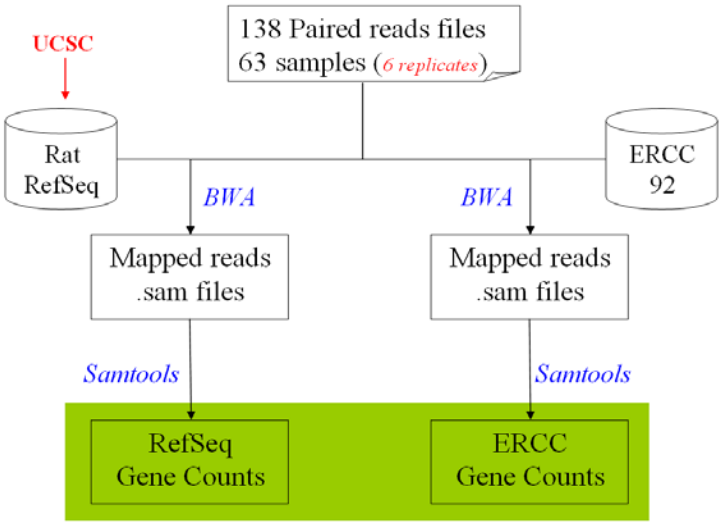
In this pipeline, raw paired-end sequencing data (raw reads) of 63 samples (among them, 6 samples were sequenced twice with different machines) were aligned to reference sequences as depicted in the figure below. First, the raw reads were mapped to the rat RefSeq sequences which were downloaded from http://genome.ucsc.edu/cgi-bin/hgTables and the ERCC sequences that were obtained from SEQC consortium by using a publically available mapping software package BWA (version: 0.5.9-r16). Parameters used in BWA mapping (default if not specified below) are (1) the first 25 bases were used as the seeds; (2) the maximum edit distance in the seed was set to 1;

and (3) the maximum edit distance of alignment was set to 2. The output of the mapping results were in the format of .SAM files that describe the detail of the mapping. To extract the reads mapped to each of the transcripts and genes in the .SAM files, the publically available software package, Samtools (version: 0.1.13 (r926:134)), was used to by the following command lines in the order: (1) samtools view -b –S; (2) samtools sort; (3) samtools index; and (4) samtools idxstats. The raw read count for genes and transcripts is normalized and transformed into log2 reads per million (RPM) with a global scaling method as equation:

$$\log_2(RPM) = \log_2[\frac{(C_{raw}) \times 1000000}{N}], C_{raw} \geq 1$$

$$\log_2(RPM) = \log_2[\frac{0.8 \times 1000000}{N}], C_{raw} = 0$$

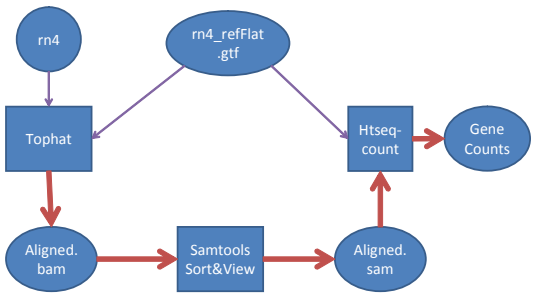
where C_{raw} is the raw read count of a gene or transcript; N is the total number of reads mapping either to the reference (counting the reads mapping to both RefSeq RNAs and ERCC).



RNA-seq data analysis: pipeline-4

Raw paired-end sequencing data were mapped to the rn4 genome using Tophat 1.4.0. Specifically, Tophat first filters out reads that map to many places on the genome (-M),

then maps the remaining reads to the transcriptome (-G), and finally map any reads that do not map to the transcriptome to the genome. The Tophat output (accepted_hits.bam) was sorted and converted to SAM format with samtools 0.1.18 , then gene counts were generated with htseq-count 0.5.3p3. The combination of restricting Tophat to uniquely aligned reads, and requiring that the read fall entirely within the gene in htseq-count should result in a very conservative gene count. Gene counts for different TGx samples were compiled into one table by a custom awk script. (please specify the normalization method). The detail analysis is illustrated in the figure below and explained in Supplementary Method.



Tophat

Switch	Description
-g 1	Use only uniquely aligned reads (1 alignment)
--library-type fr-unstranded	Library is unstranded (standard Illumina)
--phred64-quals	Use Phred quality + 64 (same as –solexa1.3-quals)
-G rn4_refFlat	Align reads to the rn4 transcriptome first
-M rn4	Exclude multi-mapped reads from transcriptome mapping

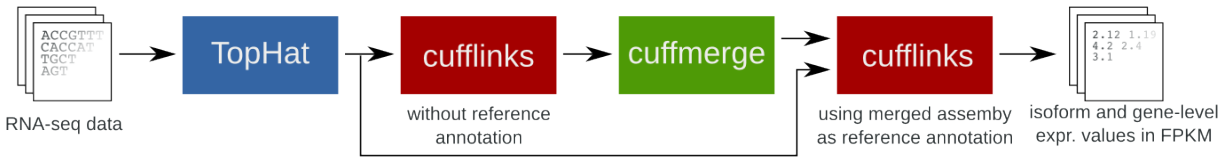
Htseq-count

Switch	Description
--stranded=no	Assay is not strand specific
--mode=intersection-strict	Entire read must fall within the gene

RNA-seq data analysis: pipeline-5

The pipeline consists of three steps: alignment, transcripts assembly, and quantification. In the first step, RNA-seq reads per sample were mapped to the reference using TopHat v2.0 [1]. The reference consists of the UCSC RN4 rat genome, the UCSC "known genes" track for rat transcriptome, and the 92 ERCC sequences, as downloaded from http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_095047.txt. TopHat was run with library type set to Illumina, inner distance between mate pairs set to 160 bp and maximum number of mapping per read set to 10. Moreover, the reads were first mapped to the rat transcriptome; only reads that could not be aligned to the transcriptome were mapped to the rat genome and ERCC sequences. The mapping rate per sample was 0.808 on average, with a standard deviation of 0.027, a minimum of 0.722, and a maximum of 0.875. In the second step of the pipeline, the mapped reads were processed by Cufflinks v1.3.0 [2] to assemble transcripts; in order to detect possible novel transcripts, Cufflinks was run without any

reference annotation. On average, 19309 transcripts were identified per sample. The resulting assemblies were merged together by cuffmerge, obtaining a merged assembly consisting of 21536 transcripts. In the last step, the merged assembly was used as reference annotation for Cufflinks to estimate isoform and gene-level expression values normalized in terms of Fragments Per Kilobase of exon per Million fragments mapped (FPKM). The pipeline was run on the FBK high-performance computing facility KORE. *The analysis is illustrated in the figure below.*

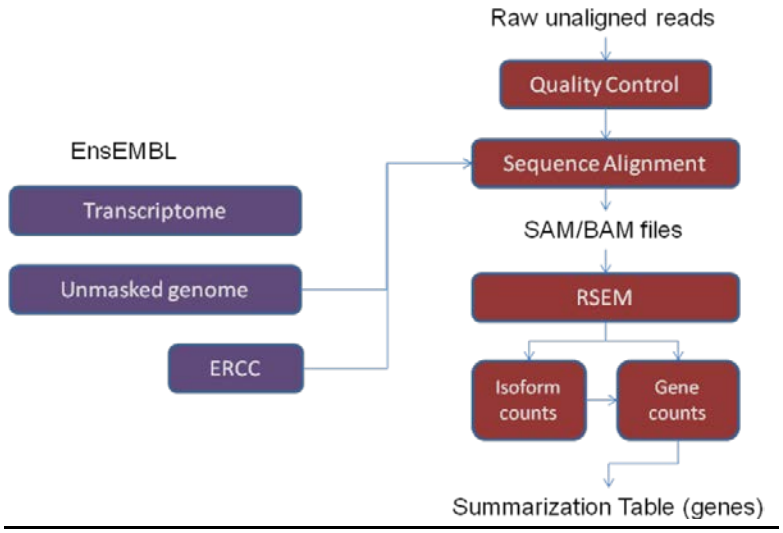


[1] C. Trapnell, L. Pachter, and S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25 (2009) 1105-1111.
[2] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech 28 (2010) 511-515.

RNA-seq data analysis: pipeline-6

The pipeline consists of 2 steps: alignment by bowtie and quantification by RSEM v1.1.18-modified (Li and Dewey, 2011). Prior to sequence alignment, the quality of the sequences was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The RNA-seq reads were aligned to the unmasked rat EnsEMBL genome (build 66 - ftp://ftp.ensembl.org/pub/release-66/fasta/rattus_norvegicus/dna/Rattus_norvegicus.RGSC3.4.66.dna.toplevel.fa.gz) using bowtie (v0.12.7). Prior to aligning the reads, ERCC sequences were added to the EnsEMBL database. For bowtie the following parameters were used: -q –phred64-quals -n 2 -e 99999999 -l 25 -I 1 -X 1000 -p 46 -a -m 200 -S. RSEM was chosen to handle both gene and isoform level multi-reads. RSEM re-evaluates ambiguously mapped reads if they uniquely align within the same gene but different isoforms by estimating the expression of that isoform using a maximum likelihood estimate with confidence

intervals. In a last step two output files for each sample were generated -- one at the isoform level and the other at the gene level. The TPM (Transcripts per Million) values of all samples were combined into a large table for further processing.



Li, B, Dewey CN, RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome, BMC Bioinformatics. 2011 Aug 4;12:323.

Supplementary Note 2. Classifier Development

1. Classifier development – methodology M1

The classification of the rats by chemical and by MOA is a complex network problem. The first task is to recognize the triplets of rats who received identical chemical. The next task is to group the chemicals per MOA. But some chemicals elicit a much stronger response than other chemicals in the same MOA (e.g. LEF in AhR). Functional overlap across chemicals' response is also common and most of the perturbations in gene expression appear to reflect a cascade of secondary effects rather than the molecular initiating event. A semi-automatic method is proposed to regroup the rats iteratively using all relative comparisons until an optimal pattern emerges. In contrast to conventional methods, this analysis is based on all genes expressed in each rat rather than on the top differentially expressed genes. Each rat is compared at once to the three control varieties (by route) or to the whole group of controls, as well as to ~40 candidate groups of rats, by chemical, by MOA, or even by histopathology and biological panel measures. The most important comparisons are between the treated rat and the controls, but the distance to all other treatments enriches the picture and helps to unambiguously identify the pattern of each chemical or MOA.

In practice, each individual is represented by an N-dimensional vector with the expression index of each gene as coordinates, and the cosine of the angles is used to measure the distance between pairs of vectors, i.e. the scalar product of the 2 vectors divided by the product of their lengths. To avoid being swamped by normalization issues for the lowly expressed genes, all indexes below a given threshold t are set to t and the scalar product is computed only over the genes with index above t in at least one of the 2 vectors. The threshold t was chosen at 8 for this study. Groups of rats are hand selected and their average vectors are computed. 18 groups were constructed for the training set, one for each chemical and control type. Each individual rat is then compared to these named groups and the cosines could be plotted to see the grouping by MOA.

To classify the blinded test set, the 48 new runs were similarly compared to the training group, the technical replicas from the training set, treated with aflatoxin or controls, as well as a few new controls were recognized. 12 groups of 3 rats treated with a given chemical were easily and all correctly identified. The grouping by MOA was more difficult due to pleiotropy, as well as activation or repression of genes in common secondary cascades, leading to some overlap in profiles. In particular, CAR/PXR-like effects are strong in the ER MOA, making the identification of the new to the training CAR|PXR an ambiguous choice. This led to an initial switch in the CAR/PXR MOA identification in the first predictions, which was then corrected in the second.

Applied to the microarray data, however, the covariance plots were more cluttered and the topology less crisp, so that confident identification of all groups using the covariance method could not be achieved.

2. Classifier development – methodology M2

2.1 Data sets and endpoint

Two data sets from 111 samples (training and validation) were used in the classifier development. The first data set was generated from RNA-seq platform (pipeline-3 described above), which contains 16978 genes (RefSeq). The second data set was generated from Affymetrix microarray platform, containing 31042 probe sets. The endpoint of each data is the mode of toxicological action (MOA).

2.2. Data Pre-processing

2.2.1 RNA-seq data

The raw counts in the RNA-seq data sets were normalized using RPM transformation. For each gene (RefSeq), the mean value of the control samples were calculated. Then the ratio values of the other samples were calculated. The genes with any control mean value equals “0” were filtered out. The zero value was replaced by a smaller value of the minimum (min/1.2). Afterward, the data set were transformed using log2.

2.2.2 microarray data

Similarly, the mean intensities of the matched control samples were calculated first and were then used to calculate the ratios for the treated samples by dividing the intensities of the treated samples using the mean intensities of the matched controls.

2.3 Feature selection

Information Gain was used as the feature selection method. Before modeling, the features were sorted by the information gain values in descending order. The candidate number of features ranged from 1 to 22 with step size 1.

2.4 Classification method

Support vector machine (SVM) with radial basis function (RBF) kernel was used in this study. There are two parameters for an RBF kernel: C and γ. In this study, we used grid searching method to determine the best combination of C and γ. The average accuracy of 10 times 5-fold cross validation was calculated as the evaluation criteria. The searching range for C was -5 to 15 with a step of 2, and the range of γ was -15 to 5 with a step of 2. The SVM module was provided by LibSVM.

2.5 Predictions

The endpoint of MOA is not binary, and there are 5 different MOAs in the training sets. Therefore, 5 binary classification models were built instead of 1 multi-case prediction model. Each model was associated with 1 MOA, and gave a possibility range from 0 to 1 of whether the predicted samples belong to the associated MOA. Final determination was made according to the 5 probabilities calculated by the 5 models:

$$\mathrm{Pred} = [P_1, P_2, P_3, P_4, P_5]$$

$$\mathrm{MOA}_{\mathrm{Pred}} = \begin{cases} \mathrm{MOA}_i & (P_i = \mathrm{Max}(\mathrm{Pred}) \ \mathbf{and} \ P_i > 0.5) \\ new \ \mathrm{MOA} & (\mathrm{Max}(\mathrm{Pred}) \leq 0.5) \end{cases}$$

3. Classifier development – methodology M3

The other DEGs are derived from ArrayTrack's t-test, where RNA-seq data (pipeline-2 described above) and microarray data (MAS5) were used (1) to calculate ratio for each

gene by pairing the treated sample divided by the average of the 6 vehicle matched controls. Each control is divided by the average of 6 controls as well (using reference average comparison normalization in ArrayTrack); (2) to conduct a t-test for each of MOA: using 9 treated samples vs. 12 samples of vehicle matched controls and identified signatures for each platform RNA-seq and microarray, respectively; A) top 100 P value ranking; B) top 100 fold change ranking with $P<0.01$; C) the common gene signature shared between RNA-seq and microarray.

ArrayTrack's KNN method was used for classifiers' trainings and predictions. The training model was built one MOA at time for PPARA and Car/PXR, respectively. Each model was based on 45 training samples, divided into two groups: group 1 (36 samples) and group 2 (9 samples of the specific MOA) with K=5 External validations were done using the 36 sample unknown testing set.

4. Classifier development – methodology M4

Recognizing that different MOAs have transcriptome response to different strength as indicated by the number of DEGs, we devised a feature selection algorithm named by "sequential MOA removal ANOVA based voting". Briefly, ANOVA test was done first for all five MOAs in the training set to derive a list of MOA differentiating gene signatures. PPARA, Cytotoxicity, and DNA Damage were then sequentially removed and ANOVA tests were done at each step. Those genes selected by all 4 ANOVA tests were used to classify the test samples. We classified the test samples based on a visual reading of the Hierarchical Clustering Analysis (HCA) and Principal Component Analysis (PCA) of both training and test samples. ANOVA, HCA and PCA were conducted on the ratio data. Prior to feature selection, a list of gene candidates was produced and ratio data were generated based on the matched controls with the following criteria: valid expression value for the treated sample, number of matched controls with valid expression values greater than 3 (out of 6), the standard deviation (SD) of log2 expression value among the matched controls less than 1. For each MOA, those DEGs shared by at least two chemicals in that MOA were designated as DEGs for the MOA. All 5 MOAs in the training set were then ranked by the number of their DEGs, which determined the sequential

order for removal during ANOVA testing and voting. A union of all MOA's DEGs served as gene candidates for ANOVA tests. R/limma was adopted for per-chemical DEG test with p-value threshold of 0.05 and fold change cut-off of 1.5. This classifier approach was applied on both microarray data (MAS5) and RNA-seq data (pipeline-1 described above).

5. Classifier development – methodology M5

A general graph-based semi-supervised learning (GGSSL) algorithm with novel class discovery (Nie et al., 2010) was adopted for prediction and novelty detection. GGSSL training was based on 20 random subsampling (Monte Carlo) runs on the training data set, stratified over the five classes, splitting the data into internal training and internal test sets (75-25% training-test proportion). In details:

1) For each run, the internal training set was scaled to [-1, 1] and a ranked list of features, weighted for importance (Robnik-Sikonja and Kononenko, 2003), was used to build a sequence of KNN models (k=3) with an increasing number of such features. Models were evaluated on the internal test set, with classification performance assessed by the Matthews Correlation Coefficient (MCC).
2) An average MCC curve over the 20 runs was computed for different feature set sizes, together with a unified feature list ranked by average position (Borda count list). A feature set size *s** maximizing the average MCC was then selected.
3) Training and validation data were used together to build a GGSSL classification model in semi-supervised mode (labels available only for the training) on the *s** selected features. For each model, class labels (six classes: the original five MOA classes and class UNKNOWN) were thus predicted for the validation data.
4) To define the optimal GGSSL model over a grid of parameters, a KNN classifier was run in a 10x random subsampling schema (75-25% training-test proportion) on all data, using given labels for training and predicted labels for validation. The GGSSL model with best KNN performance in terms of MCC was selected and the corresponding labels used as predictions.

The classifier was developed by using the Python package mlpy v3.5.0 (http://mlpy.fbk.eu) on the FBK high-performance computing facility KORE.

Nie F, Xiang S, Liu Y, Zhang C. A general graph-based semi-supervised learning with novel class discovery. Neural Computing and Applications (2010), 19(4):549-555
Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning (2003), 53:23-69

Supplementary Note 3. Housing of animals

Upon delivery to testing laboratory male Sprague-Dawley (Crl:CD® (SD)|GS BR) were housed individually in hanging, stainless steel, wire-bottom cages in a ventilated room (temperature, 22±3°C; humidity, 30–70%; 12-h light:12-h dark cycle per day, 6:00 a.m.–6:00 p.m.) and received Certified Rodent Diet #5002 (PMI Feeds Inc.. Richmond, IN); chlorinated tap water was available ad libidum. Animals were acclimated for one week to the testing laboratory prior to dosing. After dosing, all liver samples were harvested as 100 mg punches using six millimeter disposable biopsy punches. An appropriately staggered schedule was employed so that the harvest times were accurate to ±30 min of the designed dose-to-harvest interval.

Supplementary Note 4. Sample selection

The following criteria were applied in the selection of treated samples: act via one of 7 modes of toxicological action, have existing Affymetrix whole genome GeneChip® Rat Genome 230 2.0 array data available via the DrugMatrix Database, be derived from repeat dose toxicity studies (3, 5, or 7 days) and be from liver. Sets of 6 control samples were selected to match three different route (oral gavage) and vehicle (nutritive or non-nutritive) combinations employed for test chemical delivery. As with the test chemical treated samples, all control samples needed to have existing Affymetrix whole genome GeneChip® Rat Genome 230 2.0 array data. It should be noted that due to sample selection during the creation of the DrugMatrix database Affymetrix data for set control samples were not derived from the same toxicity study as the treated samples.

Supplementary Note 5. Paired-end RNA-seq analysis

RNA-sequencing (RNA-seq) of 63 training and 42 test set samples was performed according to the manufacturer's protocol using the Illumina TruSeq RNA Sample Preparation Kit and SBS Kit v3 (San Diego, CA) by the Functional Genomics Core Facility at City of Hope National Medical Center & Beckman Research Institute. RNA quality was determined with an Agilent Bioanalyzer (RIN > 7.4 for all samples). Briefly, 500ng of each

total RNA sample was mixed with I µl 1:100 diluted ERCC RNA spike in control mix1 or mix2 (cat 4456740, 4456739, Life Technologies), then it was used for polyA mRNA selection and fragmentation, followed by first and second strand synthesis, end repair, adenylation of 3' ends, and adapter ligation. Each library was enriched by 15 cycles of PCR and the size distribution was validated on the Agilent Bioanalyzer using a DNA 1000 kit. The final library is made from a band between 200-500bp with a peak at approximately 260bp. All libraries were quantitated with Qubit 2.0 Fluorometer (Life Technologies, Grand Island, NY) and ran at a concentration of 8.6 pM on HiScanSQ or HiSeq2000 Sequencing Systems (Illumina) using four samples per lane. A total of 123 runs were made, including 18 replicas meant for quality control: those reran the same library on different platforms, or distinct libraries from the same sample. Depths of 30 - 130 million of paired 100 bp reads were generated for each sample.

Supplementary Note 6. RNA preparation and array hybridization

Automated RNA isolation was performed according to the manufacture's protocol using the RNeasy Kit from Qiagen (Germantown, MD). cRNA target preparation and fragmentation was carried out using the GeneChip® One-Cycle Target Labeling and Control Reagents Kit (P/N 900493) from Affymetrix (Affymetrix Inc., Santa Clara, CA) according to the Affymetrix GeneChip® Expression Analysis Technical Manual (subsections: Total RNA and mRNA isolation for one-cycle target labeling, one-cycle cDNA synthesis, cleanup of double-stranded cDNA for both the one-cycle and two-cycle target labeling assays, synthesis of biotin-labeled cRNA for both the one-cycle and two-cycle target labeling assays, cleanup and quantification of biotin-labeled cRNA, and fragmenting the cRNA for target preparation) using 20 µg of isolated RNA. cRNA fragmentation mix was as follows: 20 µg RNA, 8 µL 5X Fragmentation Buffer, RNase-free water to a final volume of 40µL.

The array hybridization cocktail was as follows: 15 µg cRNA, 5 µL control oligo B2 (3 nM), 15 µL 20X eukaryotic hybridization controls (bioB, bioC, bioD and cre), 150 µL 2X hybridization mix, 10 µL DMSO, and water to a final volume of 300 µL. After 16 hours of

array hybridization in the GeneChip® hybridization oven 640, the arrays were washed, stained and scanned in accordance with protocols outlined in Affymetrix GeneChip® Expression Analysis Technical Manual (subsection: Eukaryotic Arrays: Washing, Staining, and Scanning) using a GeneChip® Fluidics Station 450.