Environ Health Perspect

DOI: 10.1289/EHP3986

Note to readers with disabilities: *EHP* strives to ensure that all journal content is accessible to all readers. However, some figures and Supplemental Material published in *EHP* articles may not conform to 508 standards due to the complexity of the information being presented. If you need assistance accessing journal content, please contact ehp508@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.

Supplemental Material

The Carcinogenome Project: *In Vitro* Gene Expression Profiling of Chemical Perturbations to Predict Long-Term Carcinogenicity

Amy Li, Xiaodong Lu, Ted Natoli, Joshua Bittker, Nisha S. Sipes, Aravind Subramanian, Scott Auerbach, David H. Sherr, and Stefano Monti

Table of Contents

Table S11A. GSEA analysis of signature DN_GTX_LIVER enrichment in Drugmatrix Cell Culture Low Dose with respect to phenotype Genotoxicity (GSEA results summary).

Table S11B. GSEA analysis of signature DN_GTX_LIVER enrichment in Drugmatrix Cell Culture Low Dose with respect to phenotype Genotoxicity (Gene ranking).

Table S12. Cmap Perturbagen Classes (PCL) with differential connectivity between carcinogens and non-carcinogens.

Table S13. Cmap Perturbagen Classes (PCL) with differential connectivity between genotoxicants and non-genotoxicants.

Table S14. Cmap Perturbagens with differential connectivity between carcinogens and non-carcinogens.

Table S15. Cmap Perturbagens with differential connectivity between genotoxicants and non-genotoxicants.

Figure S1. Overview of Experimental Design and Analysis Aims: (A) Data generation and annotation: Chemicals with long-term in vivo chemical annotation, as annotated by the Carcinogenic Potency Project, were procured. HepG2 cells are exposed to each chemical and followed by gene expression profiling. The number of unique chemicals and unique profiles by category (carcinogen, non-carcinogen, others) were catalogued. (B) Data analysis: analysis of the data consists of 1) analysis of transcriptional bioactivity using the Transcriptional Activity Scores (TAS), 2) prediction of carcinogenicity and genotoxicity, 3) mechanisms of action analysis using differential pathway enrichment analysis, and 4) comparison to other signatures such as signatures of carcinogenicity (Drugmatrix), small molecule perturbations (Connectivity Map) and Aryl hydrocarbon receptor (AhR) Receptor activity (Tox21).

Figure S2. Distribution of Transcriptional Activity Scores (TAS) grouped by chemical genotoxicity within each dose level. P-values indicate the significance of unpaired one-sided two-group TAS comparison between TAS of genotoxic chemicals and TAS of non-genotoxic chemicals within each dose group (* = $p < 0.05$) (see methods). The lower, middle, upper hinges correspond to the 25th, 50th (median), and 75th percentile. The upper and lower whiskers extend to the smaller and largest value at most 1.5 * IQR (inter-quartile range) from the hinge. Data points beyond the whiskers are represented as dots. Following multiple hypothesis testing, the FDR values are reported as follows: Dose rank 1: FDR = 0.12, Dose rank 2: FDR = 0.88, Dose rank 3: FDR = 0.12, Dose rank 4: FDR = 0.24, Dose rank 5: FDR = 0.55, Dose rank 6: FDR = 0.12.

Figure S3. Distribution of Transcriptional Activity Scores (TAS) grouped by chemical genotoxicity within each dose level, separated by different chemical procurement sources: (A) Sigma Aldrich chemicals with max dose of 20uM and (B) NTP chemicals with max dose of 40uM. P-values indicate the significance of unpaired one-sided two-group TAS comparison between TAS of genotoxic chemicals and TAS of non-genotoxic chemicals within each dose group (* = p< 0.05) (see methods). The lower, middle, upper hinges correspond to the 25th, 50th (median), and 75th percentile. The upper and lower whiskers extend to the smaller and largest value at most 1.5 * IQR (inter-quartile range) from the hinge. Data points beyond the whiskers are represented as dots.

Figure S4. Sensitivity and specificity rates of classifiers at threshold of 0.3 in predictive models of carcinogenicity and genotoxicity. Boxplots have the following specifications: the lower, middle, upper hinges corresponding to the 25th, 50th (median), and 75th percentile, the upper and lower whiskers extend to the smaller and largest value at most 1.5 * IQR (inter-quartile range) from the hinge, and data points beyond the whiskers represented as dots.

Figure S5. Prediction probabilities on unlabeled chemicals for prediction of carcinogenicity in (A) all unlabeled profiles (B) profiles with Trascriptional Activity Scores (TAS) > 0.4, and for prediction of genotoxicity in (C) all unlabeled profiles (D) profiles with TAS > 0.4.

Figure S6. Heatmap of pathway enrichment scores (GSVA) for top 40 upregulated and downregulated differential pathways of carcinogenicity (A) and genotoxicity (B) for profiles with TAS > 0.2. Columns are clustered using the ward method with Euclidean distances. Rows are ordered by the frequency of the pathway categories among the top 40 (direction sensitive).

Figure S7. Pathway enrichment (pathways in Reactome 2016) of directionally inconsistent signatures between Drugmatrix and L1000 using Enrichr (Chen et al. 2013; Kuleshov et al. 2016).

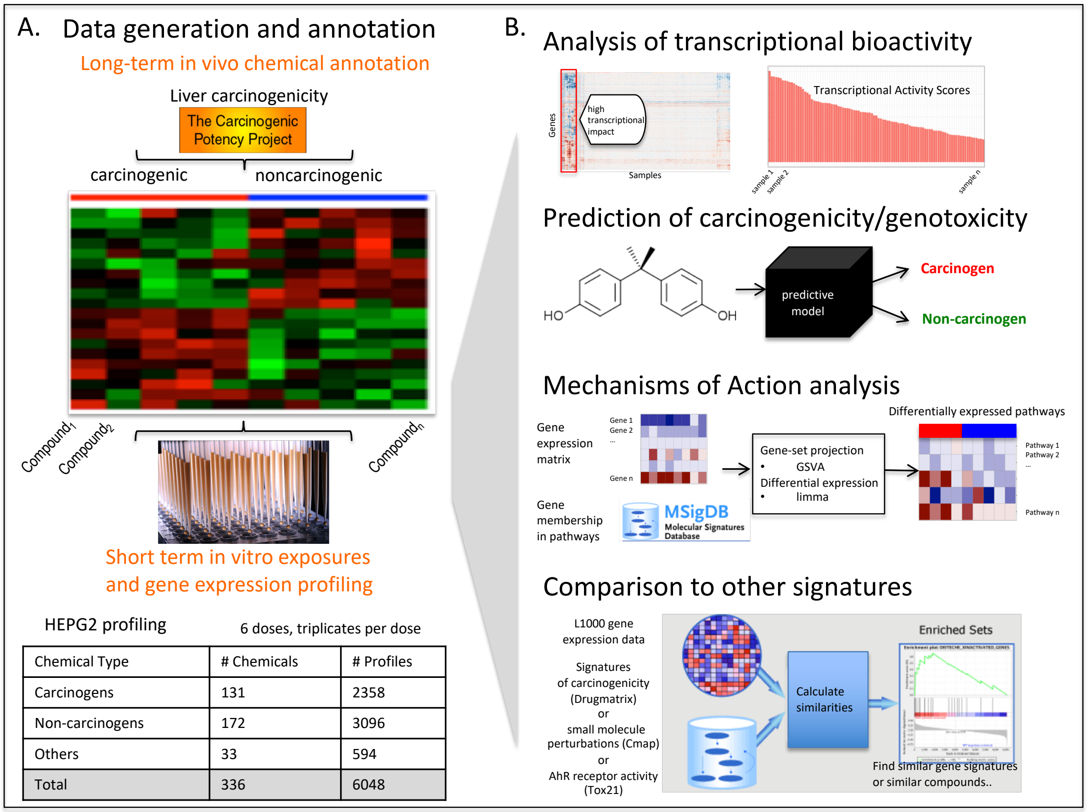
References Cited in Supplemental Material

Additional File- Excel Document

1. Supplemental Tables (see Supplemental Excel File)

2. Supplemental Figures

Figure S1


A.
Data generation and annotation
Long-term in vivo chemical annotation
Liver carcinogenicity
The Carcinogenic Potency Project
carcinogenic
noncarcinogenic
Compound$_1$
Compound$_2$
Compound$_n$
Short term in vitro exposures
and gene expression profiling
HEPG2 profiling
6 doses, triplicates per dose
B.
Analysis of transcriptional bioactivity
Genes
high transcriptional impact
Samples
Transcriptional Activity Scores
sample 1
sample 2
sample n
Prediction of carcinogenicity/genotoxicity
HO
OH
predictive model
Carcinogen
Non-carcinogen
Mechanisms of Action analysis
Gene expression matrix
Gene 1
Gene 2
...
Gene n
Gene membership in pathways
MSigDB
Molecular Signatures Database
Gene-set projection
• GSVA
Differential expression
• limma
Differentially expressed pathways
Pathway 1
Pathway 2
...
Pathway n
Comparison to other signatures
L1000 gene expression data
Signatures of carcinogenicity (Drugmatrix)
or
small molecule perturbations (Cmap)
or
AhR receptor activity (Tox21)
Calculate similarities
Enriched Sets
Find similar gene signatures or similar compounds..

Chemical Type	# Chemicals	# Profiles
Carcinogens	131	2358
Non-carcinogens	172	3096
Others	33	594
Total	336	6048

Figure S1: Overview of Experimental Design and Analysis Aims: (A) Data generation and annotation: Chemicals with long-term in vivo chemical annotation, as annotated by the Carcinogenic Potency Project, were procured. HepG2 cells are exposed to each chemical and followed by gene expression profiling. The number of unique chemicals and unique profiles by category (carcinogen, non-carcinogen, others) were catalogued. (B) Data analysis: analysis of the data consists of 1) analysis of transcriptional bioactivity using the Transcriptional Activity Scores (TAS), 2) prediction of carcinogenicity and genotoxicity, 3) mechanisms of action analysis using differential pathway enrichment

analysis, and 4) comparison to other signatures such as signatures of carcinogenicity (Drugmatrix), small molecule perturbations (Connectivity Map) and Aryl hydrocarbon receptor (AhR) Receptor activity (Tox21).

Figure S2

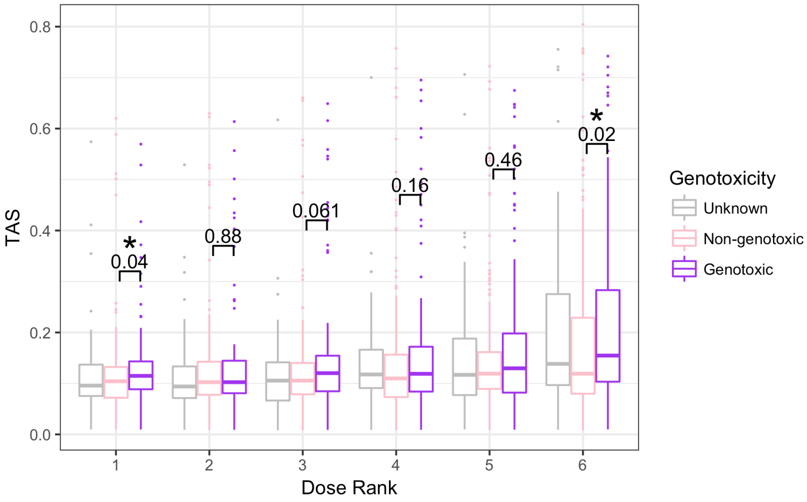


Figure S2: Distribution of Transcriptional Activity Scores (TAS) grouped by chemical genotoxicity within each dose level. P-values indicate the significance of unpaired one-sided two-group TAS comparison between TAS of genotoxic chemicals and TAS of non-genotoxic chemicals within each dose group (* = p< 0.05) (see methods). The lower, middle, upper hinges correspond to the 25th, 50th (median), and 75th percentile. The upper and lower whiskers extend to the smaller and largest value at most 1.5 * IQR

(inter-quartile range) from the hinge. Data points beyond the whiskers are represented as dots. Following multiple hypothesis testing, the FDR values are reported as follows: Dose rank 1: FDR = 0.12, Dose rank 2: FDR = 0.88, Dose rank 3: FDR = 0.12, Dose rank 4: FDR = 0.24, Dose rank 5: FDR = 0.55, Dose rank 6: FDR = 0.12.

Figure S3

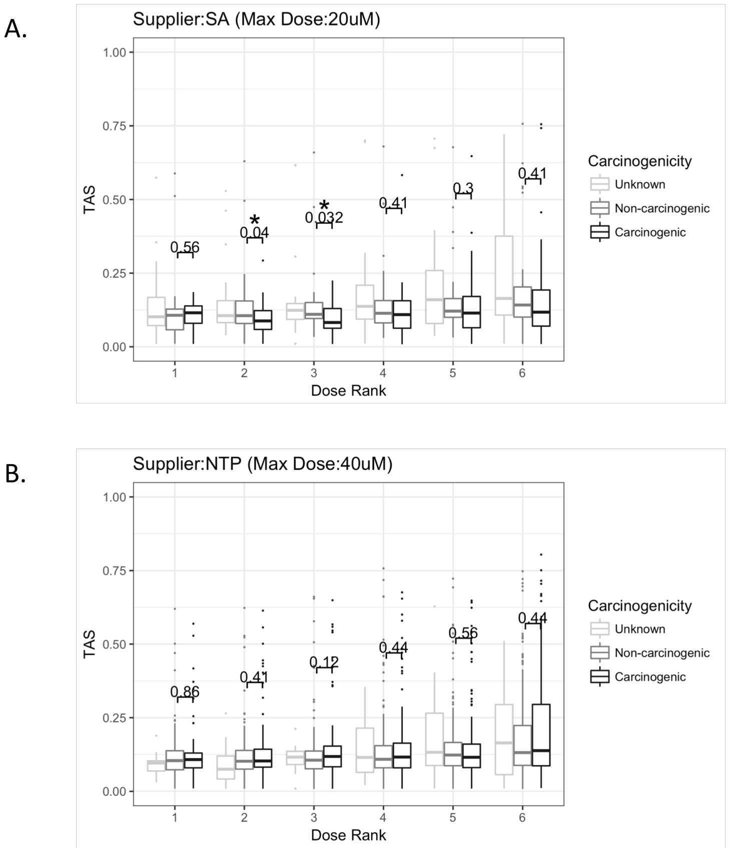


Figure S3: Distribution of Transcriptional Activity Scores (TAS) grouped by chemical genotoxicity within each dose level, separated by different chemical procurement sources: (A) Sigma Aldrich chemicals with max dose of 20uM and (B) NTP chemicals with max dose of 40uM. P-values indicate the significance of unpaired one-sided two-group TAS comparison between TAS of genotoxic chemicals and TAS of non-genotoxic chemicals within each dose group (* = p< 0.05) (see methods). The lower, middle, upper hinges correspond to the 25th, 50th (median), and 75th percentile. The upper and lower

whiskers extend to the smaller and largest value at most 1.5 * IQR (inter-quartile range) from the hinge. Data points beyond the whiskers are represented as dots.

Figure S4

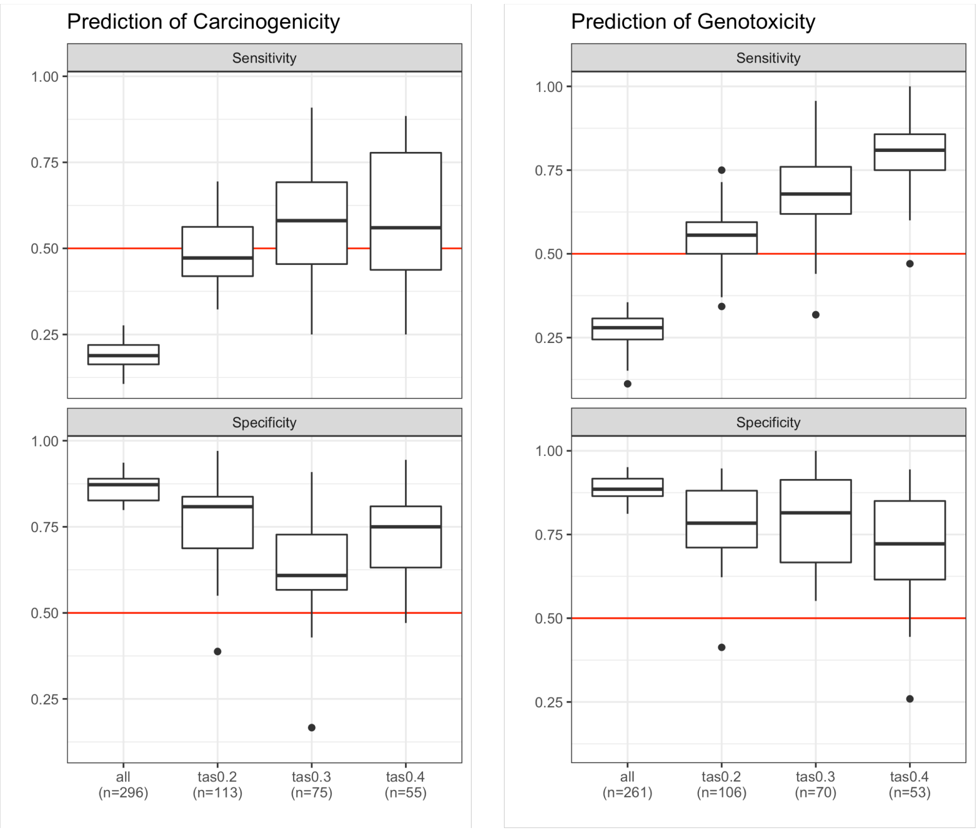


Figure S4: Sensitivity and specificity rates of classifiers at threshold of 0.3 in predictive models of carcinogenicity and genotoxicity. Boxplots have the following specifications: the lower, middle, upper hinges corresponding to the 25th, 50th (median), and 75th percentile, the upper and lower whiskers extend to the smaller and largest value at most 1.5 * IQR (inter-quartile range) from the hinge, and data points beyond the whiskers represented as dots.

Figure S5

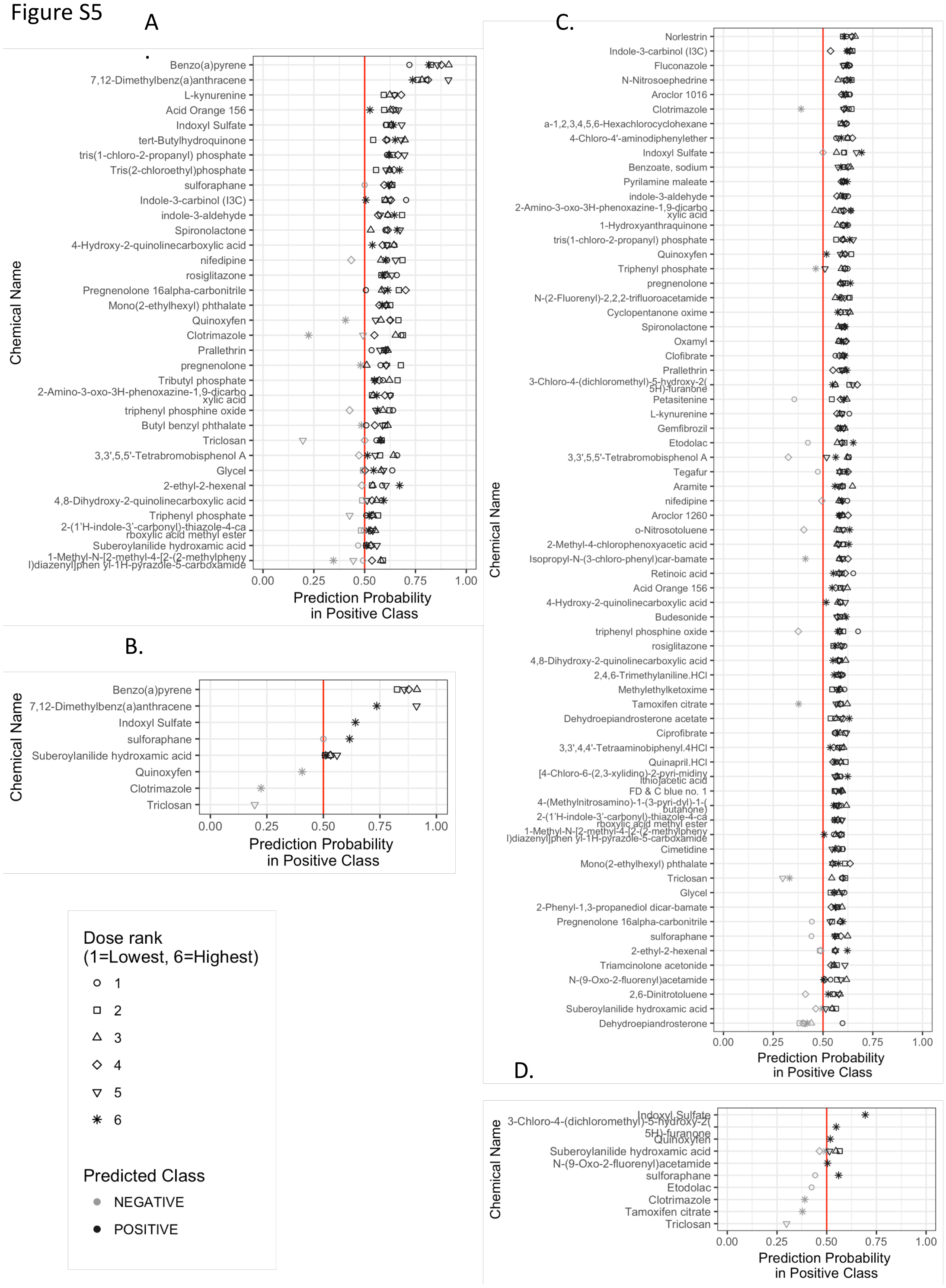


Figure S5: Prediction probabilities on unlabeled chemicals for prediction of carcinogenicity in (A) all unlabeled profiles (B) profiles with Trascriptional Activity

Scores (TAS) > 0.4, and for prediction of genotoxicity in (C) all unlabeled profiles (D) profiles with TAS > 0.4

Figure S6

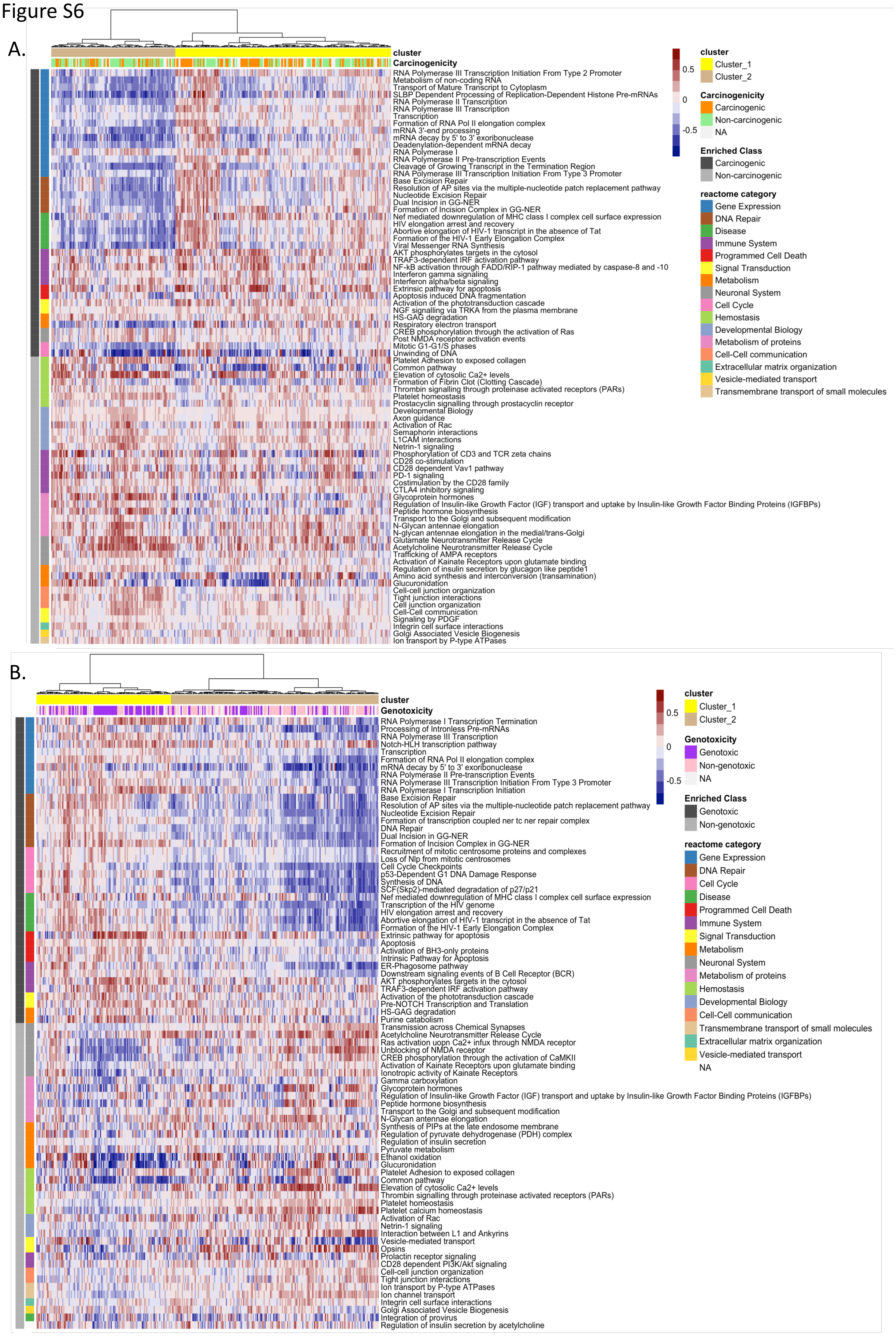
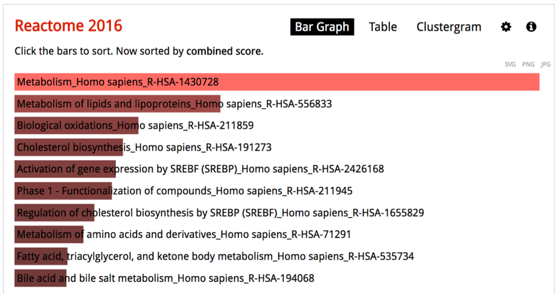

A.
cluster
Carcinogenicity
RNA Polymerase III Transcription Initiation From Type 2 Promoter
Metabolism of non-coding RNA
Transport of Mature Transcript to Cytoplasm
SLBP Dependent Processing of Replication-Dependent Histone Pre-mRNAs
RNA Polymerase II Transcription
RNA Polymerase III Transcription
Transcription
Formation of RNA Pol II elongation complex
mRNA 3'-end processing
mRNA decay by 5' to 3' exoribonuclease
Deadenylation-dependent mRNA decay
RNA Polymerase I
RNA Polymerase II Pre-transcription Events
Cleavage of Growing Transcript in the Termination Region
RNA Polymerase III Transcription Initiation From Type 3 Promoter
Base Excision Repair
Resolution of AP sites via the multiple-nucleotide patch replacement pathway
Nucleotide Excision Repair
Dual Incision in GG-NER
Formation of Incision Complex in GG-NER
Nef mediated downregulation of MHC class I complex cell surface expression
HIV elongation arrest and recovery
Abortive elongation of HIV-1 transcript in the absence of Tat
Formation of the HIV-1 Early Elongation Complex
Viral Messenger RNA Synthesis
AKT phosphorylates targets in the cytosol
TRAF3-dependent IRF activation pathway
NF-kB activation through FADD/RIP-1 pathway mediated by caspase-8 and -10
Interferon gamma signaling
Interferon alpha/beta signaling
Extrinsic pathway for apoptosis
Apoptosis induced DNA fragmentation
Activation of the phototransduction cascade
NGF signalling via TRKA from the plasma membrane
HS-GAG degradation
Respiratory electron transport
CREB phosphorylation through the activation of Ras
Post NMDA receptor activation events
Mitotic G1-G1/S phases
Unwinding of DNA
Platelet Adhesion to exposed collagen
Common pathway
Elevation of cytosolic Ca2+ levels
Formation of Fibrin Clot (Clotting Cascade)
Thrombin signalling through proteinase activated receptors (PARs)
Platelet homeostasis
Prostacyclin signalling through prostacyclin receptor
Developmental Biology
Axon guidance
Activation of Rac
Semaphorin interactions
L1CAM interactions
Netrin-1 signaling
Phosphorylation of CD3 and TCR zeta chains
CD28 co-stimulation
CD28 dependent Vav1 pathway
PD-1 signaling
Costimulation by the CD28 family
CTLA4 inhibitory signaling
Glycoprotein hormones
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Peptide hormone biosynthesis
Transport to the Golgi and subsequent modification
N-Glycan antennae elongation
N-glycan antennae elongation in the medial/trans-Golgi
Glutamate Neurotransmitter Release Cycle
Acetylcholine Neurotransmitter Release Cycle
Trafficking of AMPA receptors
Activation of Kainate Receptors upon glutamate binding
Regulation of insulin secretion by glucagon like peptide1
Amino acid synthesis and interconversion (transamination)
Glucuronidation
Cell-cell junction organization
Tight junction interactions
Cell junction organization
Cell-Cell communication
Signaling by PDGF
Integrin cell surface interactions
Golgi Associated Vesicle Biogenesis
Ion transport by P-type ATPases
0.5
0
-0.5
cluster
Cluster_1
Cluster_2
Carcinogenicity
Carcinogenic
Non-carcinogenic
NA
Enriched Class
Carcinogenic
Non-carcinogenic
reactome category
Gene Expression
DNA Repair
Disease
Immune System
Programmed Cell Death
Signal Transduction
Metabolism
Neuronal System
Cell Cycle
Hemostasis
Developmental Biology
Metabolism of proteins
Cell-Cell communication
Extracellular matrix organization
Vesicle-mediated transport
Transmembrane transport of small molecules
B.
cluster
Genotoxicity
RNA Polymerase I Transcription Termination
Processing of Intronless Pre-mRNAs
RNA Polymerase III Transcription
Notch-HLH transcription pathway
Transcription
Formation of RNA Pol II elongation complex
mRNA decay by 5' to 3' exoribonuclease
RNA Polymerase II Pre-transcription Events
RNA Polymerase III Transcription Initiation From Type 3 Promoter
RNA Polymerase I Transcription Initiation
Base Excision Repair
Resolution of AP sites via the multiple-nucleotide patch replacement pathway
Nucleotide Excision Repair
Formation of transcription coupled ner tc ner repair complex
DNA Repair
Dual Incision in GG-NER
Formation of Incision Complex in GG-NER
Recruitment of mitotic centrosome proteins and complexes
Loss of Nlp from mitotic centrosomes
Cell Cycle Checkpoints
p53-Dependent G1 DNA Damage Response
Synthesis of DNA
SCF(Skp2)-mediated degradation of p27/p21
Nef mediated downregulation of MHC class I complex cell surface expression
Transcription of the HIV genome
HIV elongation arrest and recovery
Abortive elongation of HIV-1 transcript in the absence of Tat
Formation of the HIV-1 Early Elongation Complex
Extrinsic pathway for apoptosis
Apoptosis
Activation of BH3-only proteins
Intrinsic Pathway for Apoptosis
ER-Phagosome pathway
Downstream signaling events of B Cell Receptor (BCR)
AKT phosphorylates targets in the cytosol
TRAF3-dependent IRF activation pathway
Activation of the phototransduction cascade
Pre-NOTCH Transcription and Translation
HS-GAG degradation
Purine catabolism
Transmission across Chemical Synapses
Acetylcholine Neurotransmitter Release Cycle
Ras activation uopn Ca2+ infux through NMDA receptor
Unblocking of NMDA receptor
CREB phosphorylation through the activation of CaMKII
Activation of Kainate Receptors upon glutamate binding
Ionotropic activity of Kainate Receptors
Gamma carboxylation
Glycoprotein hormones
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Peptide hormone biosynthesis
Transport to the Golgi and subsequent modification
N-Glycan antennae elongation
Synthesis of PIPs at the late endosome membrane
Regulation of pyruvate dehydrogenase (PDH) complex
Regulation of insulin secretion
Pyruvate metabolism
Ethanol oxidation
Glucuronidation
Platelet Adhesion to exposed collagen
Common pathway
Elevation of cytosolic Ca2+ levels
Thrombin signalling through proteinase activated receptors (PARs)
Platelet homeostasis
Platelet calcium homeostasis
Activation of Rac
Netrin-1 signaling
Interaction between L1 and Ankyrins
Vesicle-mediated transport
Opsins
Prolactin receptor signaling
CD28 dependent PI3K/Akt signaling
Cell-cell junction organization
Tight junction interactions
Ion transport by P-type ATPases
Ion channel transport
Integrin cell surface interactions
Golgi Associated Vesicle Biogenesis
Integration of provirus
Regulation of insulin secretion by acetylcholine
0.5
0
-0.5
cluster
Cluster_1
Cluster_2
Genotoxicity
Genotoxic
Non-genotoxic
NA
Enriched Class
Genotoxic
Non-genotoxic
reactome category
Gene Expression
DNA Repair
Cell Cycle
Disease
Programmed Cell Death
Immune System
Signal Transduction
Metabolism
Neuronal System
Metabolism of proteins
Hemostasis
Developmental Biology
Cell-Cell communication
Transmembrane transport of small molecules
Extracellular matrix organization
Vesicle-mediated transport
NA

Figure S6. Heatmap of pathway enrichment scores (GSVA) for top 40 upregulated and downregulated differential pathways of carcinogenicity (A) and genotoxicity (B) for profiles with TAS > 0.2. Columns are clustered using the ward method with Euclidean distances. Rows are ordered by the frequency of the pathway categories among the top 40 (direction sensitive).

Figure S7

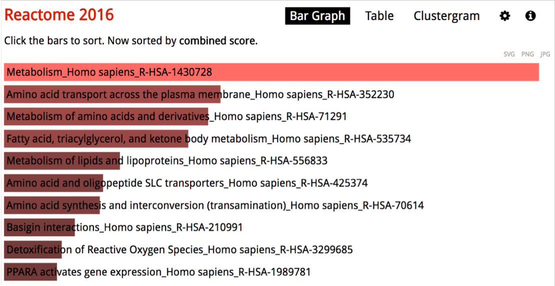
UP_GTX_LIVER



Click the bars to sort. Now sorted by **combined score.**

Metabolism_Homo sapiens_R-HSA-1430728
Metabolism of lipids and lipoproteins_Homo sapiens_R-HSA-556833
Biological oxidations_Homo sapiens_R-HSA-211859
Cholesterol biosynthesis_Homo sapiens_R-HSA-191273
Activation of gene expression by SREBF (SREBP)_Homo sapiens_R-HSA-2426168
Phase 1 - Functionalization of compounds_Homo sapiens_R-HSA-211945
Regulation of cholesterol biosynthesis by SREBP (SREBF)_Homo sapiens_R-HSA-1655829
Metabolism of amino acids and derivatives_Homo sapiens_R-HSA-71291
Fatty acid, triacylglycerol, and ketone body metabolism_Homo sapiens_R-HSA-535734
Bile acid and bile salt metabolism_Homo sapiens_R-HSA-194068

DN_GTX_LIVER



Click the bars to sort. Now sorted by **combined score.**

Metabolism_Homo sapiens_R-HSA-1430728
Amino acid transport across the plasma membrane_Homo sapiens_R-HSA-352230
Metabolism of amino acids and derivatives_Homo sapiens_R-HSA-71291
Fatty acid, triacylglycerol, and ketone body metabolism_Homo sapiens_R-HSA-535734
Metabolism of lipids and lipoproteins_Homo sapiens_R-HSA-556833
Amino acid and oligopeptide SLC transporters_Homo sapiens_R-HSA-425374
Amino acid synthesis and interconversion (transamination)_Homo sapiens_R-HSA-70614
Basigin interactions_Homo sapiens_R-HSA-210991
Detoxification of Reactive Oxygen Species_Homo sapiens_R-HSA-3299685
PPARA activates gene expression_Homo sapiens_R-HSA-1989781

Figure S7: Pathway enrichment (pathways in Reactome 2016) of directionally inconsistent signatures between Drugmatrix and L1000 using Enrichr (Chen et al. 2013; Kuleshov et al. 2016).

3. References Cited in Supplemental Material

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles G V., et al. 2013. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14; doi:10.1186/1471-2105-14-128.

Kuleshov M V., Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44:W90–W97; doi:10.1093/nar/gkw377.