Environ Health Perspect

DOI: 10.1289/EHP3986

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