2019

Deconstructing the carcinogenome: cancer genomics and exposome data generation, analysis, an tool development to further cancer prevention and therapy

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

DECONSTRUCTING THE CARCINOGENOME: CANCER GENOMICS AND EXPOSOME DATA GENERATION, ANALYSIS, AND TOOL DEVELOPMENT TO FURTHER CANCER PREVENTION AND THERAPY

by

AMY LI

B.S., Carnegie Mellon University, 2013

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader _____

Stefano Monti, Ph.D.
Associate Professor of Medicine

Second Reader _____

W. Evan Johnson, Ph.D.
Associate Professor of Medicine and Biostatistics

ACKNOWLEDGMENTS

Stefano, thank you for being the most patient, respectful, and helpful advisor I could ever ask for. You've pushed me to take on many innovative and collaborative projects and gave me tremendous guidance and encouragement to persevere and succeed.

My committee members, Paola, Evan, Dave, and Bob, thank you all for your guidance and your willingness to share knowledge from your individual field of research.

My present and former lab mates, Eric, Anthony, Ali, Steph, Vinay, Dan, Yuxiang, Liye, Francesca, it was a pleasure to work alongside you all not only as colleagues but as friends and confidants.

To the staff and members at the Find the Cause Breast Cancer Foundation, thank you for funding two years of my PhD research, and thank you for all your dedication and commitment towards preventative breast cancer research. We are getting there. Soon, we will not only cure but prevent this horrendous disease.

To the staff and faculties at the Boston University Bioinformatics Program, thank you for believing in me and helping me stay on track, and giving me the resources and opportunities to succeed.

To the Boston University Superfund program, thank you for the financial support and for giving me the opportunity to travel and present at various conferences.

My parents, you encouraged me to come back to Boston after college and I could not be more thankful for helping me in this decision. Of course, also thank you for raising me in the way you did, with unconditional love and support.

My husband, Kurt, you were there before I started this PhD journey, and you were my rock every step of the way. Thank you for always being there through the good and bad, and I look forward to the next chapters of our lives together with you.

DECONSTRUCTING THE CARCINOGENOME: CANCER GENOMICS AND EXPOSOME DATA GENERATION, ANALYSIS, AND TOOL DEVELOPMENT TO FURTHER CANCER PREVENTION AND THERAPY

AMY LI

Boston University Graduate School of Arts and Sciences,

and College of Engineering 2019

Major Professor: Stefano Monti, Associate Professor of Medicine

ABSTRACT

The rise in large-scale cancer genomics data collection initiatives has paved the way for extensive research aimed at understanding the biology of human cancer. While the majority of this research is motivated by clinical applications aimed at advancing targeted therapy, cancer prevention initiatives are less emphasized.

Many cancers are not attributable to known heritable genetic factors, making environmental exposure a main suspect in driving cancer risk. A major aspect of cancer prevention involves the identification of chemical carcinogens, substances linked to increased cancer susceptibility. Traditional methods for chemical carcinogens testing, including epidemiological studies and rodent bioassays, are expensive to conduct, not scalable to a large number of chemicals, and not capable of detecting specific mechanisms of actions of carcinogenicity. Thus, there exists a dire need for improvement in data generation and computational method development for chemical carcinogenicity testing.

Here, we coin the term "carcinogenome" to denote the complete cancer genomic landscape encompassing both its repertoire of environmental chemical exposures, as well as its germ-line and somatic mutations and epi-genetic regulators. To study the carcinogenome, we analyze both the genomic behavior of real human tumors as well as profiles of the exposome, that is, data derived from chemical exposures in human, animal or cell line models.

My thesis consists of two distinct projects that, through the generation and innovative analysis of multi-omics data, aim at advancing our understanding of the molecular mechanisms of cancer initiation and progression, and of the role environmental exposure plays in these processes. First, I detail our effort at data generation and method development for characterizing environmental contributions to carcinogenesis using transcriptional profiles of chemical perturbations. Second, I present the tool iEDGE (Integration of Epi-DNA and Gene Expression) and its applications to the integrative analysis of multi-level cancer genomics data from human primary tumors of multiple cancer types.

These projects collectively further our understanding of the carcinogenome and will hopefully foster both cancer prevention, through the identification of environmental chemical carcinogens, and cancer therapy, through the discovery of novel cancer gene drivers and therapeutic targets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AHR ..... Aryl Hydrocarbon Receptor

aRI ..... Adjusted Rand Index

AUC ..... Area Under the ROC Curve

CCLE ..... Cancer Cell Line Encyclopedia

CMap ..... Connectivity Map

Cmax ..... Estimated Equivalent In-Vitro Dose

CPDB ..... Carcinogenic Potency Database

CN ..... Correlation Network

CTD ..... Comparative Toxicogenomics Database

DepMap ..... Cancer Dependency Map

EPA ..... Environmental Protection Agency

FDA ..... Food and Drug Administration

FDR ..... False Discovery Rate

GSEA ..... Gene Set Enrichment Analysis

iEDGE ..... Integration of (Epi-) DNA and Gene Expression

L1000 ..... Luminex-1000

MDC ..... Module Differential Connectivity Score

NHIS ..... National Health Interview Survey

NTP ..... National Toxicology Program

PCL ..... Perturbagen Class

SCNA ..... Somatic Copy Number Alteration

SFN ..... Scale-free Transformed Networks

TAS ..... Transcriptional Activity Score

TCGA..... The Cancer Genome Atlas

TOM..... Topological Overlap Matrix

WFTM..... Weighted Fraction of Trans Mediation

CHAPTER ONE: INTRODUCTION

1.1 Contributions to cancer susceptibility: environment and genetics

Despite decreasing rates of overall incidence and mortality from cancer over the last two decades, cancer is still a major killer, with a projected 1.7 million new cases and 600k deaths in 2018 alone (Siegel et al. 2018). While cancer is considered a genetic disease caused by changes to genes leading to uncontrolled cell division and growth, genetic changes that confer susceptibility to cancer can not only be hereditary, but also acquired during one's lifetime, due to random errors during cell division or to exposure to environmental toxicants. The exact contributions of hereditary *vs.* non-hereditary factors in cancer has long been a subject of debate in cancer research. While exact estimates vary, studies have agreed that inherited genetic contributions to cancer are minor (Anand et al. 2008, Lichtenstein et al. 2000). DNA replication errors also can cause mutations in key genes involved in cancer development and studies have shown a relationship between numbers of normal stem cell divisions and cancer incidence (Tomasetti and Vogelstein 2015; Tomasetti el al. 2017). Aside from hereditary predisposition and random mutations during stem cell divisions, environmental factors play a major role in cancer development. Lifestyle factors, including diet, alcohol consumption, tobacco smoking, exposure to environmental carcinogens, radiations, air pollutants, and harmful food contaminants, are important risk factors of non-hereditary cancers (Irigaray et al. 2007). If environment carcinogens are identified, limiting exposure can be the key step in prevention of cancer.

Identification of chemical carcinogens for tissue-specific cancers have led to successful policies aimed at prevention. The most prominent example of preventable cancer is lung cancer. Tobacco smoke was identified as the primary cause of lung cancer, with male smokers 23 times more likely and female smokers 17 times more likely to develop lung cancer compared to nonsmokers (American Lung Association, 2018). Among developed countries, the prevalence of smoking and the rate of lung cancer have been declining where accelerated tobacco-control programs were in action. In the U.S., cigarette smoking adults declined from 20.9 percent in 2005 to 15.1 percent in 2015 according to the National Health Interview Survey (NHIS). From 1990 to 2016, an estimated 1.3 million tobacco-related cancer deaths have been avoided (Centers for Disease Control and Prevention, 2016). Other chemical carcinogens have been identified through epidemiological studies. One example is the study of the use of asbestos and its relation to lung cancer and mesothelioma, which has been documented in over 100 epidemiological studies, primarily occupational exposure studies (Lemen et al. 1980). Asbestos fiber was commonly used as a manufacturing material, specifically as building insulation, due to its heat resistant properties. During the first half of the 20th century, evidence linked asbestos exposure among construction and manufacturing workers to increased rates of cancers, particularly mesothelioma. These findings encouraged the enforcement of policies limiting asbestos exposure in many countries (American Cancer Society, 2018). Today, asbestos is banned in more than 55 countries.

Despite limited success stories of identification of chemical carcinogens, there is increased recognition for the necessity to develop more advanced testing methods of

existing but unrecognized environmental carcinogens (Kriebel el al. 2016, Leffall & Kripke, 2010).

1.2 Exposure-based studies

Historically, most research aimed at assessing cancer risk and hazard of chemical exposures in humans relied on epidemiological studies among cancer clusters. Epidemiological studies suffer from inherent shortcomings, such as the reliance on observational data and the lack of control for spurious associations due to confounding effects, the requirement for long follow-up periods, which is not suitable for evaluation of new chemicals on the market, and the high latency periods for certain cancers, making it harder to detect associations between exposure and cancer development in limited time frames.

Starting from the 1970s, following the concern of agricultural products found to be carcinogenic in rodents, several U.S. federal agencies began efforts for testing chemical carcinogenicity in rodents. These efforts were further refined into the 2-Year Rodent Bioassay by the National Toxicology Program (NTP). Since then, over 1500 chemicals were tested for carcinogenicity in this way (Ward 2007). The use of rodent bioassays drew criticisms due to evidence of differences in mechanisms and pathology of tumor development following chemical exposures between rodents and humans.

Many current carcinogenicity testing efforts are aimed at identifying mechanisms of carcinogenesis through the use of high-throughput assays that capture various biological endpoints. These include transcriptional profiling in rodents exposed to selective sets of known carcinogens and controls (Eichner et al. 2013; Ellinger-

Ziegelbauer et al. 2008; Fielden et al. 2007; Gusenleitner et al. 2014; Kossler et al. 2015; Nie et al. 2006; Uehara et al. 2011). Although these studies are better suited to establish mechanistic explanations of chemical-induced carcinogenesis than the two year rodent bioassays, questions still remain about the relevance of rodent models for characterizing human carcinogenicity.

Several initiatives focus on the study of human carcinogenicity using in-vitro screens on human cell line models. For instance, the Toxcast project of the Environmental Protection Agency (EPA) (Judson et al; Richard et al. 2016), and Tox21 initiative in partnership with the Food and Drug Administration (FDA) (Schmidt 2009; Tice et al. 2013) uses hundreds of reporter assays to characterize adverse effects across thousands of in-vitro chemical exposures in human cell lines.

In this thesis, I build on concepts from these past studies to explore the novel use of high-throughput transcriptomic profiling in human cell lines for predicting and characterizing chemical carcinogenicity. This approach makes use of results and methods established in past studies to further accelerate and refine the process of carcinogenicity testing. For instance, labels from the two-year rodent bioassays are used to label our training set of liver carcinogens and non-carcinogens. Results from our study are compared with previously published models of carcinogenicity including results from Drugmatrix and Toxcast.

1.3 Cancer genomic profiling of primary tissues

One aspect of deconstructing the carcinogenome relies on understanding the relationship between chemical exposures and cancer initiation and progression. Another

complementary yet critical piece is the understanding of the genomic landscape of human cancers, i.e., the landscape of genetic and epigenetic alterations, and concomitant transcriptional variation, associated with the disease. In most cancer types, certain mutations in a subset of ~140 cancer driving genes can drive tumorigenesis through altered signaling in pathways that regulate cellular processes involved in cell fate, cell survival and genome maintenance (Vogelstein et al. 2013). Identification of these driver genes and the biological contexts in which mutations in these genes confer a cancerous state has been made possible by the availability of large-scale genomic data repositories such as The Cancer Genome Atlas (TCGA) (Weinstein et al. 2013). TCGA contains more than 11,000 human primary tumor samples spanning many cancer types. Each sample is profiled on multiple genomic platforms including gene expression microarray, RNA-sequencing, somatic copy number and mutation profiling, DNA methylation, microRNA sequencing, and protein expression profiling. Numerous studies used these data repositories for data query and method development for tasks such as patient stratification and biomarker or therapeutic target discovery. In this thesis, I leverage this data repository for tool development aimed at integration of paired gene-expression and epi-DNA profiles, such as copy number alterations, for the purpose of predicting cancer driver genes.

1.4 Dissertation Aims

Cancer continues to be a one of the leading causes of mortality, prompting greater need for increased research in both the areas of cancer therapy and prevention. In this

thesis, I present a two-sided approach to deconstruct the "carcinogenome", that is, to the genomic characterization of cancer through all stages of cancer initiation and progression. We focus on capturing both the impact of environmental exposures and the behaviors of germ-line and somatic mutational signatures and epi-genetic regulators. In the following chapters, I detail two major aims:

Aim 1: Towards cancer prevention – Building liver carcinogenicity and genotoxicity models from in-vitro high throughput transcriptomic assays

Current limitations in carcinogen testing prompted our effort to develop novel predictive models of chemical carcinogenicity based on the generation and computational modeling of high-throughput transcriptional profiles of human cell lines exposed to chemical carcinogens. In Chapter 2, I will describe our experimental and computational approach applied to the transcriptional profiling of HEPG2 cells following exposure with 330 chemicals annotated for rodent liver carcinogenicity.

The computational analysis of this experiment utilizes complementary approaches to understand the transcriptomic effects of the profiled chemical perturbations in human cell lines. I will present my analysis of transcriptional bioactivity of the profiled chemicals and highlight the importance of using the transcriptional bioactivity score for filtering samples for use in downstream analysis such as classification and pathway enrichment. Next, I will describe a predictive model of carcinogenicity and genotoxicity built using supervised classification, which accurately classifies chemicals with profiles with high transcriptional bioactivity. In addition to classification, I show that pathway enrichment analysis reveals gene sets representing known and potentially novel modes of

actions of carcinogenicity. I leverage other studies to cross-reference with my findings, by comparing the signatures of carcinogenicity I derived to external gene signatures from the CMap, Drugmatrix, and Tox21 database, highlighting areas of consistencies and offering explanations for discrepancies in results. Finally, I demonstrate the capacity of our profiling effort to support the interrogation of particular mechanisms of carcinogenic response. Specifically, I investigate AHR receptor-mediated gene expression response in our profiles and demonstrate the similarity of AHR mediated response in our profiles to AHR active profiles in Tox21. Lastly, I will conclude by highlighting the importance of this study in the context of carcinogen testing and showcase an online portal based on this work (https://carcinogenome.org) and offer suggestions for future directions.

This chapter also features a complementary analysis method that utilizes gene expression profiles from chemical perturbation experiments. The method implements a gene regulatory network-based approach for analyzing transcriptional profiles from large datasets of chemical perturbations. In particular, correlation networks are derived from expression profiles of distinct groups of chemicals and are then compared based on the loss or gain of connectivity of the corresponding network modules. This method is evaluated in the context of characterizing chemical-induced carcinogenesis but has broad applicability such as drug discovery and repositioning.

Aim 2: Towards cancer therapy – Molecular characterization of the cancer genome and epi-genome using integrative analysis

This aim is focused on computational method development for integrative analysis of multi-level cancer genomics data. In Chapter 3, I will outline the development

of the tool iEDGE (Integration of Epi-DNA and Gene Expression) and will demonstrate its application to the analysis of copy number and gene expression data in the TCGA.

iEDGE is a framework for the analysis of paired sets of genomic data, e.g., copy number alternation and gene expression data from a set of matched samples. This tool identifies important cis and trans genes of each user-defined (epi-)genetic regulators such as Somatic Copy Number Alterations (SCNA). Furthermore, iEDGE ranks cis genes based on the number of trans genes it mediates and makes predictions of the most likely cis driver genes of each SCNA. I applied iEDGE to the analysis of copy number and gene expression in 19 cancer types using data from the TCGA, with a particular focus on breast cancer, and successfully demonstrate the ability of iEDGE to predict cis driver genes that are significantly enriched for known cancer driver genes from benchmark cancer driver databases. Furthermore, iEDGE-predicted driver genes contain several putative novel oncogenes and tumor suppressors.

Finally, I will demonstrate the performance of the mediation test used in iEDGE using simulated datasets.

CHAPTER TWO: Towards Cancer Prevention – Characterization of transcriptomic profiles from chemical perturbations

2.1 Building liver carcinogenicity and genotoxicity models from in-vitro high-throughput transcriptomic assays

2.1.1 Introduction

Despite significant investments into cancer research over the last decades, approximately 1.7 million new cancer cases and 600,000 cancers deaths were estimated in the U.S. in 2017 alone (American Cancer Society 2017). Of these, 90-95% are not attributable to known heritable genetic factors, thus making environmental exposures a major suspect in driving cancer (Anand et al. 2008), notwithstanding recent studies pointing to the rate of cell replications as an important determinant of cancer development among different tissue types (Tomasetti and Vogelstein 2015; Tomasetti et al. 2017). Most research aimed at assessing cancer hazard from chemical exposure has primarily relied on epidemiological studies of past human exposures to suspected carcinogens in cancer clusters, and on carcinogen screening based on the 2-year rodent-based bioassay. Epidemiological studies rely on observational data, and as such they have limited ability to rule out the possibility of spurious associations due to confounding effects. They also require that exposure to a suspected carcinogen is documentable. Even when the nature of the chemical exposure and the exposure dose is known, epidemiological studies require long follow-up periods, hence are not appropriate for the evaluation of new chemicals on the market. Similarly, the 2-year rodent bioassay, the

gold standard for carcinogen testing, is time-consuming and requires up to $4 million and more than 800 animals per compound. As a result, less than 2% of the ~85,000 chemicals registered in the TSCA Chemical Substance Inventory have been tested by this approach (Bucher and Portier 2004; Gold et al. 2005; Huff et al. 2008).

High-throughput transcriptional profiles from short-term chemical exposures have proven useful for predicting long-term carcinogenicity and for capturing multiple biological MoAs of long-term carcinogenicity. Many studies have explored the use of high-throughput transcriptional profiling in rodent models (Eichner et al. 2013; Ellinger-Ziegelbauer et al. 2008; Fielden et al. 2007; Gusenleitner et al. 2014; Kossler et al. 2015; Nie et al. 2006; Uehara et al. 2011). However, questions remain about the relevance of rodent models for characterizing human carcinogenicity, and most importantly, they are still excessively time-consuming and expensive for large-scale testing. In-vitro-based screens would help address the time and cost constraints of carcinogen testing through automated high-throughput plating, exposure treatment, and assaying. EPA's Toxcast (Judson et al. 2010; Richard et al. 2016) and Tox21 initiatives (Schmidt 2009; Tice et al. 2013) have used various reporter assays to characterize adverse effects across thousands of in-vitro chemical exposures. However, while these efforts use high-throughput techniques with carefully selected gene, pathway and adverse-response-centric endpoints, the number of assays and the diversity of endpoints are limited. For instance, ToxCast used 624 in-vitro endpoints mapped to 315 genes in Phase I (Judson et al. 2010) and an additional ~200 new endpoints in Phase II (Richard et al. 2016). Studies utilizing this data for the assessment of chemical carcinogenicity have emphasized the need to expand

the assay set to better characterize diverse MoAs of certain carcinogens (Kleinstreuer et al. 2013). mRNA profiling, by assaying the entire transcriptome, or a large portion of it, represents a promising solution to this need by providing an agnostic view of which genes and pathways are relevant to chemical-induced carcinogenesis.

Given the technological advances in gene expression profiling and the development of cost-effective sequencing platforms, opportunities arise for their use in large-scale toxicological screenings. One such solution is the Luminex-1000 (L1000) platform (Peck et al. 2006), a low-cost, high-throughput bead-based platform that measures the expression of ~1000 landmark genes and infers the remaining genes in the transcriptome by imputation. This platform was used in the creation of the Connectivity Map (CMap) (Subramanian et al. 2017), which now includes 1.3 million perturbation profiles of drugs and small molecules and has been instrumental in the discovery of small molecule MoAs. Due to its cost-effectiveness and appropriateness for large-scale perturbation screening, we adopted it for the profiling of chemical carcinogens.

We applied the L1000 platform to study the effects of chemical perturbations of previously validated rat liver carcinogens and non-carcinogens in HEPG2 cell lines. The central hypothesis underlying our study design was that the long-term carcinogenicity of chemicals can be accurately predicted from gene expression profiles of short-term in vitro models. Our approach used machine-learning techniques to build predictive models of the long-term carcinogenicity of chemicals based on L1000-derived gene expression profiles of human cell lines exposed to the studied chemicals. Furthermore, we annotated the in-vitro-derived gene signatures by performing pathway enrichment of carcinogens *vs.* non-

carcinogens, to identify MoAs associated with chemical-induced carcinogenesis. Signatures derived from this study were also compared to external gene signatures and chemical annotations from knowledge bases such as Drugmatrix, CMap, and Tox21, to verify the consistency of results and expand the interpretation of findings. An overview of our experimental design and analysis aims is presented in Figure S2.1.1.

2.1.2 Methods

Chemical selection and annotation

In the chemical selection process, we prioritized chemicals with long-term rodent liver carcinogenicity annotation for inclusion in this experiment. Long-term carcinogenicity annotations were derived from the Carcinogenic Potency Database (CPDB) (Fitzpatrick 2008). Additional chemicals without carcinogenicity annotation were included on the basis of interest to the Superfund Research Program (environmental toxicants) presence in controversial commercial products (included for predictive purposes) and evidence of binding to the aryl hydrocarbon receptor (AhR), as the AhR is an important mediator of xenobiotics, including carcinogens. A complete list of chemicals and their annotations is provided in Table S1. For CPDB annotations, the final carcinogenicity labels denote "+" if carcinogenic in rat liver (female or male) or "-" if non-carcinogenic in both rat and mouse (in female and male) across all tested organs in the CPDB. Genotoxicity labels denote "+" if mutagenic or weakly mutagenic in the Salmonella assay, and "-" otherwise. In total, 330 unique chemicals were used in the analysis including 128 carcinogens, 168 non-carcinogens, 100 genotoxicants, and 161 non-genotoxicants.

Chemical procurement and data generation

Chemicals were procured from the Tox21 library of the National Toxicology Program (NTP) when available, or from Sigma-Aldrich otherwise. Compound purity and identity were confirmed by UPLC-MS (Waters, Milford, MA). Purity was measured by UV absorbance at 210 nm or by Evaporative Light Scattering (ELSD). Identity was determined on a SQ mass spectrometer by positive and/or negative electrospray ionization.

Detailed cell culture, plating, treatment and lysis protocols are described in https://assets.clue.io/resources/sop-cell.pdf (Subramanian et al. 2017). Briefly, HepG2 (human hepatocellular carcinoma cell line; ATCC HB-8065) was used with medium RPMI1640 (Mediatech 10040CV) supplemented with 10% v/v fetal bovine serum (Sigman F4135), 1x penicillin-streptomycin-glutamine (Invitrogen 10378-016), and incubated at humidified 5% CO_2 atmosphere at 37°C. Cell cultures were plated with 4,000 cells (45ul of growth medium) per well on 384-well plates (Corning 3707) and incubated for 24 hours before treatment. Cells were treated with 5uL of 1:100 diluted 1000x stock compound plates to final volume of 50 μL and incubated for 24 hours before lysis.

Each chemical perturbation was administered at 6 doses in triplicate wells per dose and chemical combination, starting from 40μM maximum dose (40mM stock diluted 1:1000) for NTP chemicals and 20μM for chemicals procured from Sigma-Aldrich, in series of two-fold dilutions. The sole exception to the standard dosage was 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD), which had a starting dose of 50nM due to

its extreme potency. The vehicle control used was DMSO. Four positive controls were used (vorinostat, geldanamycin, mitoxantrone, withaferin-a). Four wells on each plate were reserved for L1000 pipeline assay controls. These include A01: bead only control (negative control), B01: POSAMP control (hybridization/staining positive control), A02 & B02 (reference RNA control).

For cell lysis, 30µL of medium was aspirated and 25µl TCL Lysis Buffer (Qiagen 1031576) was added. Plates were sealed and maintained at room temperature for 30 minutes and frozen in −80°C freezer. Following treatment and lysis, the gene expression of the HEPG2 cells was profiled using the L1000 platform, a high-throughput assay that measures the expression of ~1000 landmark genes and computationally infers the expression of non-measured transcripts (Subramanian et al. 2017).

For each perturbation and landmark gene, we computed the change in gene expression following the perturbation using a moderated z-score procedure as described in the CMap-L1000 workflow. Differential expression values were calculated as moderated z-scores for each landmark gene and each unique perturbation (chemical and dose combination) perturbation, collapsed to a single value across replicates.

Assessing the transcriptional strength of a perturbation

We used the *transcriptional activity score* (TAS) as a summary measure of the impact of a chemical perturbation on landmark gene expression. TAS integrates *signature strength*, defined as the number of genes up-regulated or down-regulated by a particular perturbation above a given moderated z-score threshold, and *replicate correlation*, a measurement of similarity among triplicate profiles corresponding to the same

perturbation (unique combination of chemical, dose, cell line, time). Formally, *TAS* is quantified as the geometric mean of the signature strength (SS_{ngene}) and the replicate correlation (*CCq75*) in eq. (1). SS_{ngene} is defined as the number of landmark genes (referred to as *card*) with $ModZ_{adj}$ greater than 2 in eq. (2). *ModZ* is defined as the 978-element vector of replicate collapsed z-scores of landmark genes and *nrep* is the number of replicates in eq. (3). *CCq75* is the 75th percentile of the Spearman's correlation between replicates in landmark space.

$$TAS = \frac{\sqrt{SS_{ngene} * \max(CC_{q75}, 0)}}{\sqrt{978}}$$

[1]

$$SS_{ngene} = card(|modz_{adj}| >= 2) \quad [2]$$

$$modz_{adj} = modz * \sqrt{nrep} \quad [3]$$

TAS was calculated for each aggregated profile (one unique score per chemical and dose combination). This metric takes value in the [0,1] range, with higher values of TAS taken to represent a higher level of chemical bioactivity.

Statistical tests for comparison of TAS across profiles

We tested for the difference in TAS values among adjacent dose groups using a one-tailed Wilcoxon Signed-Rank Test (paired difference test), with the pairing determined by the unique chemical IDs to determine the statistical significance of strictly increased TAS levels between adjacent and increasing dose groups.

We next tested for difference in TAS between chemicals. In particular, for each dose rank, two-group comparisons of TAS scores between carcinogens and non-

carcinogens, and between genotoxicants and non-genotoxicants, were conducted using one-tailed unpaired two-samples Wilcoxon test, to determine the presence and significance of increased TAS for the carcinogenic compared to non-carcinogenic group, or for the genotoxic compared to non-genotoxic group.

Equivalent In-vitro dose (Cmax) estimation and association with TAS

Finding the relationship between in-vitro gene expression responses and adverse phenotypes in-vivo is an important goal of this study. To this end, we assessed the relationship between in-vitro transcriptional bioactivity (TAS) and corresponding in-vivo dose used in the rodent bioassay from which carcinogenicity labels were derived. Using a toxicokinetic model, we estimated the *equivalent in-vitro dose* (Cmax) corresponding to the in-vivo dose tested in the rat bioassay. Cmax values, maximum plasma concentrations, were estimated using a 3-compartment model in the R package HTTK v1.8 (Pearce et al. 2017). For carcinogenic compounds, these values were derived from the CPDB-reported median toxic dose (TD50) administered in rats. For non-carcinogenicity compounds, Cmax values were derived from the CPDB-reported maximum dose administered in rats. Chemicals with missing TD50 (if carcinogenic) or maximum dose (if non-carcinogenic) were omitted from this analysis. It was assumed that dosing was once per day for 365 days. While these Cmax values were not used in the in-vitro dosing scheme, they can be used in the interpretation of the aberrant behavior of some of our in-vitro profiles.

Supervised learning for prediction of carcinogenicity and genotoxicity

To build classifiers for the prediction of carcinogenicity and genotoxicity, we used the moderated z-scores of landmark genes as predictive features. The Random Forest classifier was used, as implemented in the R package *caret* (Kuhn 2008). The performance of the classifier was evaluated using a resampling scheme consisting of 25 random repeats of training on 70% of the samples and testing on the remaining 30%. The training and test set split was performed at the chemical level, so that all replicates of each chemical were only included either in the train or the test set, to avoid "information leakage" (over-fitting). To assess the effect of chemicals' bioactivity on the performance of the classifier, the evaluation was repeated on different subsets of profiles corresponding to different TAS thresholds (all profiles, TAS>0.2, >0.3, >0.4). Area under the ROC curve (AUC) was used for the assessment of a classifier performance, as it is a well-established metric that captures the trade-off between sensitivity and specificity across multiple thresholds.

Final predictions of carcinogenicity and genotoxicity were made using leave-one-(chemical)-out (LOCO) cross-validation (CV); that is, at each CV iteration, a single chemical's profiles across multiple doses are left out and a classifier is trained based on all remaining chemicals, then applied to the prediction of the left-out chemical's profiles. This procedure is repeated with each of the TAS subsets.

Deriving pathway signatures of carcinogenicity

We derived pathway activity scores using the R Bioconductor package GSVA (Hänzelmann et al. 2013). GSVA is a competitive test of gene set enrichment that takes

as input a gene-by-sample expression matrix and generates a geneset-by-sample enrichment score matrix, with its entries representing the pathway enrichment of each sample with respect to each of a user specified list of gene-sets. Pathway enrichment scores were calculated for pathways in the MsigDB C2 Reactome pathway compendium (Croft et al. 2014; Fabregat et al. 2016; Liberzon et al. 2011). The geneset-projected matrix was then used as input for differential analysis with respect to sample phenotype labels (carcinogenicity or genotoxicity) using the R Bioconductor package *limma* (Ritchie et al. 2015; Smyth 2005) to identify pathways with differences in activity levels between chemical groups. This differential analysis was repeated from data inputs with various TAS thresholding (TAS > 0, 0.2, 0.3, 0.4). One-sided p-values consistent with the direction of change in pathway activity scores were estimated. The p-values across analyses from multiple TAS subsets were combined using the Fisher's method, and adjusted for multiple hypothesis testing using False Discovery Rate (FDR) procedure (Benjamini and Hochberg 1995).

Comparison to Drugmatrix signatures

Using gene set enrichment analysis (GSEA) (Subramanian et al. 2005), we compared how well our profiles recapitulated external signatures of carcinogenicity and genotoxicity extracted from the NTP Drugmatrix database (Ganter et al. 2006). The Drugmatrix is a compendium of microarray profiles of short-term chemical exposures in intact rat organs (liver samples used only) and in cell cultures (primary rat hepatocytes). The Drugmatrix-derived signatures were defined as the lists of genes in the Drugmatrix significantly associated with long-term carcinogenicity and genotoxicity. Data processing

of the Drugmatrix data was consistent with methods described in Gusenleitner et al. (2014). Gene features were mapped from rat Ensembl gene identifiers to human gene symbols using Biomart (Durinck et al. 2005). Differential expression analysis was conducted using *limma* (Ritchie et al. 2015; Smyth 2005) to identify markers of carcinogenicity and genotoxicity after correcting for the effect of dose and duration of exposure. For each comparison, a list of significant genes was derived using a FDR cutoff of 0.01 and absolute value of log fold change of 0.2, up to a maximum of 300 genes as ranked by FDR. Signatures of carcinogenicity and genotoxicity (direction sensitive: upregulated/downregulated) were derived for three Drugmatrix subsets: liver profiles, cell culture profiles, and low-dose cell culture profiles (< 50μM), the latter consistent with the range of doses used in our experiment. These gene signatures were tested for enrichment against our L1000 profiles in various subsets (TAS > 0, 0.2, 0.3, 0.4), using the binary phenotypes of carcinogenicity and genotoxicity and the GSEA method, with empirical p-values estimated based on 10,000 gene-set permutations.

Comparison with CMap signatures

We performed a systematic comparison of our signatures to those in the CMap database. To this end, we computed the *connectivity score*, a measure of similarity, between pairs of signatures, in this case, between each of our signatures and each of the perturbation signatures in the CMap, which comprises ~1.3 million profiles corresponding to 19,811 drugs and small molecules, and 5,075 molecular (gene-specific knockdown and over-expression) perturbations across 3 to 77 cell lines (Subramanian et al. 2017). The connectivity scores are expressed as percentile values in the [-100, 100]

range, wherein a score of 100 represents maximum signature overlap, -100 represents maximum signature reversal and 0 represents lack of concordance between signatures in either direction. Connectivity scores were computed with respect to both individual CMap perturbagens, and Perturbagen Classes (PCLs), defined as sets of perturbagens with similar MoAs or gene target annotations. Next, we performed differential connectivity analysis with respect to our chemical groups (carcinogens *vs.* non-carcinogens, genotoxicants *vs.* non-genotoxicants) using a one-tailed Wilcoxon rank-sum test to test for presence of increased connectivity in the positive class (carcinogenic or genotoxic). These tests were repeated for each TAS-based subset of our data, and false discovery rate (FDR) values were calculated. A minimum mean connectivity score of 60 for the positive class was used to filter out differential connectivity hits with low base connectivity scores.

Investigation of AhR activation in L1000 profiles

To examine the behavior of AhR-related chemicals included in the study, we tested whether these chemicals exhibit enriched activity of AhR-related gene-sets compiled from independent sources. Lists of chemicals with known AhR activity were identified using multiple AhR-related Tox 21 reporter assays extracted from the tool Tox21 Enricher, or using custom chemical annotation with expert knowledge (referenced as "Sherr_AHR_agonist" in Figure 2.1.6A). Lists of AhR target genes were compiled from literature, as annotated in Table S5.

A one-directional weighted Kolmogorov-Smirnov (KS) test was performed to test for the enrichment of "AhR-positive" samples (profiles corresponding to AhR-related

chemicals) among the top-ranked profiles sorted by descending AhR geneset activity scores. The activity scores represent the median scores across four individual AhR geneset scores calculated using GSVA.

Profiles corresponding to AhR-related chemicals in the list "Sherr_AHR_agonist" were clustered using the similarity matrix derived from the connectivity scores of the selected profiles (see previous section for the calculation of connectivity scores).

Statistical Reporting

All statements indicating significance are based on threshold of multiple hypothesis corrected $\alpha < 0.05$, unless otherwise specified.

2.1.3 Results

TAS analysis and chemical "bioactivity"

We used the *transcriptional activity score* (TAS) as a proxy for chemical bioactivity. Subsequent analyses are based on subsets of profiles at different TAS thresholds (TAS > 0, 0.2, 0.3, 0.4). TAS > 0.2 is the standard cutoff for sufficient bioactivity adopted by the CMap-L1000 workflow (Subramanian et al. 2017), while TAS > 0.3 and TAS > 0.4 represent more stringent thresholds we used to assess the effect of increasing bioactivity on downstream analysis such as classification and gene-set enrichment. While the majority of our profiles showed low transcriptional bioactivity, a substantial percent of profiles achieved sufficient TAS. Among 330 chemicals represented across 1972 replicate collapsed profiles, 133 chemicals (40.3%) achieved

TAS > 0.2 in at least one dose, 89 chemicals (26.97%) achieved TAS > 0.3 and 63 chemicals (19.09%) achieved TAS > 0.4.

The effect of chemical dose on transcriptional bioactivity

We performed statistical tests to compare TAS of adjacent dose groups and to evaluate how bioactivity is affected by dose. Statistically significantly higher TAS were found when comparing dose rank 3 with rank 2 (FDR < 0.01), rank 4 with 3, rank 5 with 4 and rank 6 with 5 (FDR < 0.001)(Figure 2.1.1A). The consistent significance of TAS differences between adjacent dose groups implies that increasing dose is effective at increasing the transcriptional bioactivity of profiles, with the maximum dose used in this experiment yielding the highest range of TAS scores. When binned by TAS range (Figure 2.1.1B), the monotonically increasing dose response of TAS was apparent across all bins and stronger for higher TAS ranges.

The effect of carcinogenicity and genotoxicity on transcriptional bioactivity

Next, we evaluated whether the level of a chemical bioactivity as captured by TAS had any association with that chemical's long-term carcinogenicity or genotoxicity. Remarkably, carcinogenicity showed no effect on TAS in all dose groups (Figure 2.1.1C). On the other hand, genotoxicity showed a marginally significant effect on TAS among profiles with dose rank 1 (lowest dose group) and dose rank 6 (highest dose group) where genotoxic chemicals had nominally significantly higher TAS compared to non-genotoxic chemicals (p-value cutoff > 0.05), although following multiple hypothesis testing (FDR method), no groups showed significance at FDR < 0.05 (Figure S2.1.2).

Comparison of in-vivo rat bioassay dosage with in-vitro bioactivity

The lack of association between TAS and carcinogenicity motivated us to further investigate the relationship between the L1000 doses and the in-vivo doses used in the rodent bioassay. To this end, we tested the association between in-vitro bioactivity (TAS) and the estimated *equivalent in-vitro dose*, Cmax (see Methods), where Cmax represents the estimated in-vitro dose corresponding to the in-vivo dose tested in the rat bioassay. Cmax estimates could be calculated for 183 of the 330 chemicals included in our screen. The mean TAS of profiles for each chemical were plotted against the same chemical's Cmax/40μM (the ratio of estimated equivalent dose to the max in-vitro dose) (Figure 2.1.1D).

To determine the association between TAS, in-vivo carcinogenicity, and Cmax, we used the following linear regression model:

$$\log_{10}(\text{Cmax}) \sim \alpha + \beta_{\text{TAS}} \times \text{TAS} + \beta_{\text{CARC}} \times \text{CARC} + \beta_{\text{T:C}} \times \text{TAS:CARC}$$

TAS denotes the mean TAS for each chemical (across 6 doses), and CARC denotes the carcinogenicity status of the chemical in the rodent bioassay. We tested for significance of the coefficients $\beta_{\text{TAS}}, \beta_{\text{CARC}}, \beta_{\text{T:C}}$ under the null hypotheses of zero-valued coefficients (no effect). We found significant effects of TAS (β_{TAS}= -4.49, p-value = 0.01), and CARC (β_{CARC}= -1.22, p-value = 0.001) and non-significant effect of the interaction of TAS and CARC ($\beta_{\text{T:C}}$ = 3.1, p-value = 0.16).

As expected, TAS was negatively associated with Cmax. In other words, chemicals that required a low equivalent dose to elicit a carcinogenic response in the rodent bioassay tended to be more transcriptionally active in the in-vitro assay. On the

other hand, there were exceptions, as carcinogenic chemicals with low TAS and non-carcinogenic chemicals with high TAS were observed and can be explained by several unique pharmacokinetic properties.

We annotated the carcinogenic chemicals with low TAS based on their structural group membership, in-vivo dose requirement for carcinogenicity labeling, and requirements for metabolic activation in HEPG2. Carcinogenic chemicals with low TAS tend to fall in one or more of the following categories: (1) small nitrosamines and other alkylating agents that form DNA adducts but are not adequately recognized by the DNA repair machinery (enriched in Group 2, Figure 2.1.1D), (2) require bioactivation by CYP2E1 and other p450s that are not present at high levels in HEPG2 cell culture (also enriched in Group2, Figure 2.1.1D), or (3) require high equivalent In-vitro dose (Cmax) to be carcinogenic, thus likely under-dosed in our in-vitro assay (enriched in Group 1 in Figure 2.1.1D).

Among non-carcinogenic chemicals with high TAS, we generally noted lower overall doses used in the rodent bioassays due to dose limiting toxicity or early deaths at higher doses in the cancer bioassay, e.g., Cyclosporin A (immune suppression and kidney toxicity)(Ryffel 1992), Pyrimethamine, Rhodamine 6G and Rotenone (bone marrow suppression)(Abdo et al. 1988; National Toxicology Program 1978; National Toxicology Program 1989), and hexachlorocyclopentadiene (point of contact pulmonary toxicity)(National Toxicology Program 1994). Thus, if higher doses were tolerated in rodent bioassays, it is possible that some of these chemicals would elicit a carcinogenic response in liver.

The effect of transcriptional bioactivity on prediction of carcinogenicity and genotoxicity

While chemical bioactivity levels were not associated with long-term carcinogenicity, the most relevant question was whether a chemical's bioactivity affected the ability of its expression profile to be predictive of carcinogenicity (and genotoxicity). To answer this question, we built multiple classifiers based on profiles with TAS values within various ranges and used a random resampling scheme to assess their prediction performance. Datasets corresponding to different TAS ranges were randomly split into train (70%) and test (30%) sets multiple times (n=25), classifiers were built on the train sets, and predictions were made on the test sets. The average Area Under the Curve (AUC), sensitivity, and specificity were then estimated over the 25 random resamples. The prediction AUC improved with higher stringencies of TAS (Figure 2.1.2). We achieved the highest predictive accuracy within the most stringent TAS subset (TAS > 0.4), with 72.2±2.7% (mean±se) AUC for prediction of carcinogenicity (Figure 2.1.2A), and 82.3±1.6% AUC for prediction of genotoxicity (Figure 2.1.2B). These results suggest that short in-vitro gene expression profiles of chemical perturbations, given sufficient transcriptional bioactivity, can accurately predict long-term chemical carcinogenicity and to a greater extent, genotoxicity.

In addition to the AUC, we report the sensitivity (true positive rate) and specificity (true negative rate) at the cutoff of 0.5 in Figure S2.1.3. Higher specificities were observed at the expense of lower sensitivities in most TAS groups for both

classifiers. This outcome is desirable for a preliminary screening strategy in which a higher false-positive rate can be tolerated.

Gene markers for prediction of carcinogenicity and genotoxicity

Final predictive models of carcinogenicity, genotoxicity, and genotoxicity within carcinogens were built using the entire set of profiles with TAS > 0.4. Landmark genes were ranked by variable importance and the top 20 genes for each model were reported in Figure 2.1.3. Variable importance was measured by the Mean Decrease in Gini Index (MeanDecreaseGini) as defined in the function *importance* in the R package *randomForest* (Liaw and Wiener 2002). In the carcinogenicity prediction model, top genes included BLCAP, an apoptosis inducing gene, and SESN1, a target of p53 in response to DNA damage and oxidative stress (Figure 2.1.3A). Among the top 20 landmark genes for prediction of genotoxicity were pro-apoptotic regulators such as BLCAP and BAX (Figure 2.1.3B). BAX is regulated by p53 and has been shown to be involved in p53-mediated apoptosis, a hallmark of DNA damage response to genotoxic chemical exposure. Of note, the absolute magnitude of the variable importance of top markers is model dependent – thus not comparable between models – and is not informative about the comparative performance of different classifiers.

The markers in Figure 2.1.3 were the most predictive features of carcinogenicity and genotoxicity in the restricted space of the L1000 landmark genes, and as such, are not necessarily the most relevant to define chemicals' MoAs. For a more thorough MoA analysis, see section *"Pathway enrichment analysis for characterizing MoAs of*

carcinogenicity and genotoxicity", where gene-set scores were derived from the expression of *all* genes, including the L1000-inferred ones.

Final predictions of carcinogenicity and genotoxicity in bioactive profiles

Final predictions of carcinogenicity and genotoxicity were made using a leave-one-chemical-out cross-validation scheme, in which predictive models were trained based on all but one chemical and predictions were made on the profiles of the left-out chemical (see methods). This procedure was repeated for all unique chemicals in profiles with TAS > 0.4 to derive probability measurements of the profile being "Positive" for either carcinogenicity or genotoxicity (see methods) using a probability threshold of 0.5. Prediction probabilities for carcinogenicity and genotoxicity were reported along with the true class labels denoted by the dot colors (Figure 2.1.4). From this representation, we observed that predictions tend to be consistent across profiles of varying doses of the same chemical. Several exceptions exist in chemicals whose prediction probabilities were close to 0.5. In addition, prediction probabilities monotonically increasing as a function of dose were observed for some compounds, e.g., 3'-Methyl-4-dimethylaminoazobenzene showed increased probability of genotoxicity prediction with increasing dose. However, this pattern did not generalize to all chemicals.

Predictions of unlabeled chemicals

Using the final predictive models trained on all profiles with TAS > 0.4, predictions of carcinogenicity and genotoxicity were made for the chemicals without known CPDB annotation (Figure S2.1.4). The majority of unlabeled profiles were predicted "Positive" for both carcinogenicity and genotoxicity using a probability

threshold of 0.5. This is likely due to bias in chemical selection. Sources of unknown chemicals include chemicals of interest to the Superfund Research Program (likely environmental toxicants), chemicals that were tested for either carcinogenicity or genotoxicity in the CPDB but whose labels cannot be determined, and controversial chemicals in commercial use (triclosan, Glycel). Many profiles have predicted probabilities between 0.5-0.65, indicating low confidence in prediction, potentially attributable to low bioactivity of profiles. When restricting predictions to unlabeled profiles with TAS > 0.4 to be consistent with the subset used for model training, the separation of ranges of prediction probabilities becomes clearer (Figure S2.1.4B and S1.1.4D). The top two ranked predicted carcinogens, benzo(a)pyrene and 7,12-Dimethylbenz(a)anthracene, are two polyaromatic hydrocarbons (PAHs) that have been shown to manifest carcinogenic and genotoxic properties.

The top ranked predicted genotoxicant, indoxyl sulfate, is an endogenous tryptophan metabolite, which has been shown to activate p53 expression through reactive oxygen species (ROS) production and is a source of endogenous oxidative DNA damage (Shimizu et al. 2013). While indoxyl sulfate may not necessarily be considered a genotoxicant as it is a uremic solvent found in low concentrations (1-5μM) in the human serum normally, it activates the AhR, inducing cytochrome P450 enzymes which metabolize other substrates, including mutagenic intermediates (see Result section "*Characterizing AhR-mediated response in L1000 gene expression profiles"*). Thus, prediction of indoxyl sulfate as a genotoxicant may be due to transcriptional activation of shared pathways involved in metabolism of genotoxic chemicals.

Pathway enrichment analysis to characterize MoAs of carcinogenicity and genotoxicity

To identify pathway-level differences between carcinogens and non-carcinogens, and similarly, between genotoxicants and non-genotoxicants, we performed differential pathway enrichment analysis and ranked pathways according to the significance of their differential enrichment between chemical groups. In accordance with the breakdown of TAS subsets used in classification analysis, and based on the observation that increasing thresholds of TAS yield better classification performance, the differential pathway enrichment analysis was repeated for each of the TAS subsets previously considered (Table S2 and Table S3). With no TAS threshold (e.g., inclusion of all profiles), only a few pathways were differentially scored between carcinogens and non-carcinogens and between genotoxicants and non-genotoxicants. With increasing thresholds of TAS, the number of significantly expressed pathways increased. At TAS 0.2 and above, the identity of significant pathways became more stable, particularly for genotoxicity-related pathways, with many significant pathways shared across TAS > 0.2, 0.3, and 0.4. To quantify the similarity of significant pathways across TAS subsets, we measured by Jaccard index the overlapping proportion of significant ($p<0.05$) pathways among all possible TAS subset pairs, and then computed the mean Jaccard index of each TAS subset with respect to all other TAS subsets. The mean Jaccard index for carcinogenicity was 0.14, 0.36, 0.42, and 0.41 for TAS > 0, 0.2, 0.3, and 0.4, respectively. For genotoxicity, it was 0.23, 0.54, 0.54, and 0.52. The increase in the number and in the overlap of significant pathways at higher TAS is likely due to the associated stronger

signal. At lower TAS, the larger number of false positives likely increased the noise level and heterogeneity of the transcriptional response, and the consequent reduction in the number of pathways found to be significantly enriched.

We derived an aggregated ranking score of differential pathway enrichment by combining p-values across all the TAS subsets (see methods) and the lists of differentially enriched pathways (combined FDR < 0.05) with respect to carcinogenicity and genotoxicity are included in Excel Tables S6 and S7, respectively. When comparing carcinogens to non-carcinogens, we observed up-regulation of immune-related pathways (interferon-α/β), cell death (apoptosis induced DNA fragmentation), DNA repair (nucleotide excision repair), transcriptional regulation (RNA polymerase I, II, and III related activity), and cell cycle checkpoints (p53-dependent G1 DNA damage checkpoint), and down-regulation of various metabolism related pathways (phase II conjugation, phase I functionalization, peptide hormone biosynthesis), cell-cell organization and communication (cell-cell junction organization, integrin cell surface interactions, tight junction interactions), and G-protein signaling. Among genotoxicants compared to non-genotoxicants, upregulated pathways include DNA repair (nucleotide excision repair, formation of incision complex in GG-NER), AKT signaling, programmed cell death, G1/S DNA damage checkpoints, innate immune response (interferon signaling, toll-like receptor signaling). Down-regulated pathways include xenobiotic metabolism (phase I and phase II metabolism), peptide hormone biosynthesis, cell-cell organization and cell-cell communication, innate immune response (complement cascade), and various hemostasis and metabolism related pathways.

From the differentially scored pathways of carcinogenicity, we identified a reduced set consisting of the top 40 up-regulated and top 40 down-regulated pathways with Reactome categories as ordered by the aggregated rankings and visualized their enrichment scores across profiles with TAS > 0.2 in Figure S2.1.5A (top pathways differentially enriched with respect to carcinogenicity) and Figure S2.1.5B (genotoxicity). Hierarchical clustering of samples revealed stratification by carcinogenicity status, with Cluster 1 significantly enriched for carcinogens compared to Cluster 2 (Fisher test p-value = 0.0073), and an even stronger stratification by genotoxicity status, with Cluster 1 significantly enriched for genotoxicants compared to Cluster 2 (Fisher test p-value = 7.36e-7).

Comparison of L1000 signatures of carcinogenicity and genotoxicity with signatures from Drugmatrix

We tested for enrichment of the Drugmatrix-derived signatures of carcinogenicity and genotoxicity against our L1000-based differential signatures of carcinogenicity and genotoxicity (Table S4). Both the directional concordance of signatures (column "direction_match", e.g., are the genes upregulated by carcinogens in Drugmatrix also upregulated in L1000?) and the significance of signature enrichment (column "FDR.q.val") were measured. Significant similarities were observed between signatures derived from Drugmatrix low dose rat primary hepatocyte cell cultures and our L1000 profiles. For example, the signature of up-regulated genes in response to low-dose carcinogens in cell cultures (UP_CARC_CELL_LOWDOSE) was enriched in the L1000-profiled carcinogen subsets at TAS > 0.4, 0.3, and 0.2 (FDR<0.05). Conversely, the

signature of down-regulated genes in response to low-dose carcinogens in cell cultures (DN_CARC_CELL_LOWDOSE) was enriched in the L1000-profiled non-carcinogen subsets at TAS > 0.2 and 0 (FDR<0.05). Similarly, signature of genotoxicants in the Drugmatrix cell cultures ("UP_GTX_CELL_LOWDOSE") was enriched in the L1000-profiled genotoxicant subsets at TAS > 0.4, 0.3, and 0. When repeating the analysis for signatures derived from *all* Drugmatrix cell culture profiles (including high doses), signatures of genotoxicity were generally directionally consistent with L1000 profiles (in all 8 relevant tests), but signatures of carcinogenicity were inconsistent, and in fact sometimes behaved in the opposite direction (directions matched according to expectation in 2 out of 8 relevant tests). For example, the Drugmatrix signature "UP_CARC_CELL" was enriched among non-carcinogens in the L1000 TAS > 0.4 subset. This inconsistency is likely due to the use of extremely high doses for some of the chemicals in the Drugmatrix cell culture profiles. For reference, the mean dose in Drugmatrix cell culture profiles was ~3,000uM and the max dose was 180mM. In contrast, the max dose among L1000 profiles was 40uM.

Next, we compared signatures derived from the Drugmatrix in-vivo rat liver profiles to the L1000 profiles. For carcinogenicity, the signature of down-regulated genes in response to carcinogens ("DN_CARC_LIVER") was correctly enriched among non-carcinogens in L1000 TAS < 0.4, 0.3, 0.2 and 0 with FDR < 0.05. Similarly, the signature of up-regulated genes in response to carcinogens ("UP_CARC_LIVER") was marginally enriched among L1000 TAS < 0.4 (FDR = 0.06), and TAS < 0.3 (FDR = 0.09) carcinogens. On the other hand, the signatures of genotoxicity were largely not enriched

in the right direction (e.g., "DN_GTX_LIVER" shows enrichment among genotoxicants of TAS 0.4).

To rule out the possibility that the observed signatures' inconsistency was due to platform differences – since the Drugmatrix data were microarray-based while our data were L1000-based – we compared Drugmatrix cell culture to Drugmatrix liver signatures of genotoxicity (both microarray based). We found that the downregulated genotoxicant signature in liver was also behaving in the opposite direction compared to the cell culture signature. This finding suggests that the signatures' inconsistency between liver and cell line was likely due to differences between in-vitro and in-vivo responses to exposure rather than to differences in the profiling platform. Upon detailed inspection of the Drugmatrix liver signatures, we identified an enrichment of genes related to metabolism in both the up- and down-regulated gene signatures (lipid metabolism, cholesterol biosynthesis, Phase I metabolism in "UP_GTX_LIVER", amino acid metabolism, fatty acid metabolism in "DN_GTX_LIVER"), supporting the conclusion that there may be substantial differences between metabolic activities in in-vitro and in-vivo models (Figure S2.1.6).

In summary, L1000-derived signatures of carcinogenicity and genotoxicity were concordant with Drugmatrix low dose cell culture signatures, but inconsistent with Drugmatrix liver signatures, with the differences largely driven by discrepancies in the expression of certain metabolism-related genes between in-vitro and in-vivo exposures.

Comparison of L1000 signatures of carcinogenicity and genotoxicity with drug perturbation signatures in the CMap

The availability of the CMap offered the opportunity to compare our profiles to a much larger database of pharmacologically annotated signatures and allowed us to predict MoAs or pharmacological properties based on signature similarity. To this end, we first computed the similarity of our signatures to each signature in the CMap using connectivity scores (see Methods). We then identified the CMap signatures that showed significant difference in connectivity scores (FDR < 0.05) between carcinogens and non-carcinogens, and between genotoxicants and non-genotoxicants. The top CMap hits are summarized at the level of Perturbagen Classes (PCLs) in Figure 2.1.5.

Focusing on the significantly differential PCLs across all TAS subsets (TAS > 0.2, 0.3, 0.4), we found that carcinogens, compared to non-carcinogens, were significantly more connected to drug classes consisting of topoisomerase inhibitors, DNA synthesis inhibitors, and ribonucleotide reductase. Genotoxicants, compared to non-genotoxicants, were significantly more connected to the three aforementioned drug classes, as well as to CDK inhibitors, aurora kinase inhibitors, and ubiquitin specific peptidases (Figure 2.1.5).

Characterizing AhR-mediated response in L1000 gene expression profiles

Carcinogens and genotoxicants are sometimes recognized by cellular receptors such as the aryl hydrocarbon receptor (AhR). Given that the AhR is an important mediator of the toxicity of many chemicals represented in our dataset, we sought to investigate the effects of AhR-activated chemicals in terms of known AhR-regulated

gene expression and the similarity of transcriptomic profiles among sub-groups of AhR agonists.

The L1000 profiles exhibited consistent enrichment of AhR-related gene-set activity among chemicals labeled as AhR-active in several Tox21 reporter assays, namely, HTS_ACTIVE.agonism_AhR (p-value: 2.9e-7), HTS_ACTIVE.cytotoxicity_AhR/agonism (pvalue: 1.5e-4) and ATG_Ahr_CIS_up (p-value: 0.006)(Figure 2.1.6A). This finding validated the ability of unbiased gene expression profiling to accurately capture endpoints from more specific and targeted assays such as those in the Tox21 library.

Next, we examined an expert-curated set of AhR-related chemicals (Group: Sherr_AHR_agonist). Based on the similarity of their gene expression profiles as measured by the connectivity scores, we found two functionally distinct classes of AhR-related chemicals (Figure 2.1.6B). Cluster 1 contains 5 profiles (out of 6) of perturbation by benzo(a)pyrene, a strong AhR agonist and known genotoxicant. Cluster 2 is enriched with profiles of strong exogenous AhR ligands, most with potent toxic effects, including 7, 12-dimethylbenz(a)anthracene (DMBA) and TCDD. It is not surprising that many of these chemicals also had high in-vitro transcriptional bioactivity (high TAS). Interestingly, profiles of indoxyl sulfate clustered with the group of strong AhR agonists. While indoxyl sulfate is an endogenous AhR ligand, it can be considered a uremic toxin that is observed at elevated levels in patients with chronic renal failure (Niwa et al. 1999). Cluster 3 contains endogenous AhR ligands (l-kynurenine, indole-3-carbonyl, kynurenic acid, xanthurenic acid, and cinnabarinic acid). Since l-tryptophan is not an AHR ligand,

its presence in this latter group suggests that it is metabolized to one of the kynurenine pathway metabolites that are AhR ligands (l-kynurenine, kynurenic acid, xanthurenic acid, and cinnabarinic acid). These results show promise for our platform to be used not only as a general predictor of active transcriptional pathways such as the AhR signaling pathway, but also to distinguish, with finer granularity, classes of AhR agonists according to the transcriptomic profile they induce.

Carcinogenome Portal – a framework for data query and visualization

All data described in this manuscript are available for public access. Data processed under the standard CMap-L1000 pipeline are available under https://clue.io/data/CRCGN_ABC. To facilitate the interactive querying of the downstream analysis results produced by this study, we developed a web portal (https://carcinogenome.org/HEPG2). The query and visualization functionalities supported by the portal include differential expression, gene-set enrichment, and connectivity analysis against CMap signatures. This interface supports both marker-centered (genes, pathways, CMap signatures) and chemical-centered queries. For instance, one can find the top gene markers and pathways regulated by a particular perturbation; identify the top chemicals that up-regulate a particular gene or pathway of interest; or find CMap chemicals or chemical groups that are most similar to the profiles of a particular perturbation. In addition, the portal supports bulk query and visualization of groups of perturbations in the form of heatmaps.

2.1.4 Discussion

Prediction of carcinogenicity and genotoxicity

The results from the prediction of carcinogenicity and genotoxicity experiments provide strong evidence that transcriptional bioactivity as captured by TAS had a high impact on the classifier performance. In fact, while absolute levels of bioactivity were not associated with carcinogenicity in our experiments, a sufficiently high bioactivity was necessary to elicit enough transcriptional signal to use a chemical's expression profile for carcinogenicity prediction. Thus, when limiting to profiles with high TAS, the performance of our predictive models drastically improved. Among highly bioactive profiles (TAS>0.4), our classifiers yielded mean AUC of 72.2% for prediction of carcinogenicity (Figure 2.1.2A), and 82.3% for prediction of genotoxicity (Figure 2.1.2B). To boost the effective sample size used in classification, we outlined the following dose selection strategy for improving bioactivity of in-vitro gene expression profiles.

In-vitro dose recommendation

The selection of doses in short-term acute exposures for prediction of long-term in-vivo phenotypes is a challenging task. In this experiment, we chose to adopt a standard 6-dose titration, starting from 40μM or 20μM depending on source of chemicals. The sole exception to the standard dosing was TCDD, whose starting concentration is 50nM due to its extreme potency. The choice of standard dosing was made for a couple of reasons: 1) lack of commercial availability of certain chemicals at higher stock concentrations; 2) scarcity of in-vitro dose recommendations from publicly available

data, e.g., dose recommendations derived from MTT assays; and 3) cost efficiency of standardized dosing using the L1000 platform.

One alternative dosing scheme is to determine unique doses for each chemical using the MTT assay. For instance, a previous study of genotoxicity prediction based on in-vitro experiments selected doses based on a MTT assay resulting in 80% viability at 72h incubation, or maximum dose of 2mM in the case of lack of cytotoxicity (Magkoufopoulou et al. 2012). Some chemicals used in that study were administered at doses that vastly exceeded the 40μM or 20μM dose limit adopted in our experimental setup. Furthermore, the lack of plateau effect in dose response as a function of TAS (proxy for bioactivity) suggests that doses exceeding the 40μM or 20μM threshold may indeed yield profiles with higher bioactivity and increase the power to detect gene and pathway markers for prediction of carcinogenicity and genotoxicity without experiencing saturation effects (response plateauing) or excessive cell death. Although standardizing dosage across chemicals was the logistically and cost-effective solution for this experiment, going forward, MTT assays are highly recommended for maximizing biological signal across transcriptional profiles. Estimation of the appropriate in-vitro dose from toxicokinetic modeling of the in-vivo doses tested in animal bioassays, when available, is another viable alternative, as shown in Figure 2.1.1D and associated discussion. We offer these dose recommendations in the context of accurate hazard prediction, for which this study has shown that sufficient signal (transcriptional bioactivity) is necessary. For effective risk assessment and translation, human relevant doses should be considered.

Acute vs. chronic response

Through analysis of transcriptional activity scores between carcinogens and non-carcinogens (Figure 2.1.1C), we observed that long-term carcinogenicity, as established from long-term in-vivo rodent studies, had no effect on transcriptional bioactivity in our short-term assay (Figure 2.1.2A). This observation supports the conclusion that bioactivity as defined by TAS at less than 40μM is not associated with carcinogenicity, and consequently, a short-term chemical perturbation with minimal transcriptional response cannot be assumed "safe".

While TAS alone was not predictive of carcinogenicity, it was instrumental to the selection of those compounds with sufficient bioactivity to allow us to build an accurate gene expression-based classifier of carcinogenicity (up to 72.2% AUC). It was also instrumental to capturing important MoAs of carcinogenicity, as shown by our pathway enrichment analysis, which highlighted the upregulation of interferon-α/β response, cell death, DNA repair, and transcriptional regulation (RNA polymerase I, II, III) pathways in response to carcinogen exposure, as well as downregulation of phase I and phase II metabolism and cell-cell organization and communication pathways. Overall, we observed a stronger signal of genotoxicity compared to carcinogenicity, which is to be expected, as the latter is a more heterogeneous phenotype and thus harder to capture as a binary distinction; this is also evidenced by the higher accuracy of the genotoxicity classifier (82.3%) as well as the by the higher TAS among genotoxicants compared to non-genotoxicants.

Implication of findings in context of tumor initiation and promotion

Chemical carcinogens can be classified into tumor initiators and promoters. Initiators cause changes to the DNA (mutagens) and promoters drive the proliferation of the cell, typically by interacting with receptors to affect pathways leading to cell proliferation. We derived labels of carcinogenicity based on long-term rodent studies, which includes both tumor initiators and promoters. However, it is important to understand that we used short-term human cell line gene expression patterns to predict long-term rodent carcinogenicity. Pathways relevant to tumor initiation were accurately captured by the short-term in-vitro gene expression data (DNA repair, DNA damage, etc). As for tumor promotion, promoters typically interact with receptors to mediate cell proliferation, and our cell culture model contains a subset of receptors that mediate these processes. However, one limitation is that culture conditions are already a promotion environment (high growth) that might limit the detection of promoting agents. Another limitation is that tumor promoters mediated by receptors not expressed in a culture system may elicit reproducible but not biologically accurate patterns of gene expression in the short-term in-vitro assay, although they may be correctly classified by our machine learning approach. Mechanistic expert judgement will need to be applied to evaluate the relevance of these findings to human carcinogenicity.

Interfacing with the Connectivity Map

One of the important features of the perturbation experiment data we generated is in their support for “guilt by association” inference of chemical function by signature-

based comparison to the Connectivity Map's Perturbagen Classes (PCLs), as illustrated in Figure 2.1.5.

For example, we showed that carcinogens are significantly more connected than non-carcinogens to the PCL consisting of topoisomerase inhibitors. These represent a specific class of DNA synthesis inhibitors, which are mainly recognized as chemotherapeutic drugs that preferentially inhibit the topoisomerase enzymes (commonly topoisomerase I or II) in cancer cells to slow their rate of replication. Topoisomerase I or II introduce single- or double-strand DNA breaks in cells undergoing replication, and form topoisomerase-DNA complexes. Most topoisomerase inhibitors function by trapping these complexes, leading to increased strand breaks but incomplete DNA replication, subsequently provoking DNA damage response and DNA repair (Pommier 2006; Pommier 2013; Wang et al. 2002). Thus, DNA damage response induced by topoisomerase inhibitors is expected to mimic the response to genotoxic carcinogens.

Other relevant PCLs also exhibit shared MoAs with carcinogens and genotoxicants. Aurora kinase inhibitors play a major role in cell cycle regulation through the induction of G1 arrest and apoptosis (Bavetsias and Linardopoulos 2015). Ubiquitin specific peptidases, specifically USP24, have been shown to play a role in DNA damage response (Zhang and Gong 2015).

Challenges and future developments

This experiment aimed to accelerate short-term in-vitro testing approaches to predict long-term chemical carcinogenicity. We showed that short-term in-vitro gene expression profiling is not only capable to accurately predict carcinogenicity and

genotoxicity, but is also useful to characterize important mechanisms of carcinogenic response, particularly DNA damage and repair, and changes in cell cycle and cell-cell organization and communication. Other general biological processes that may be relevant for carcinogenic response, including inflammatory response, immune dysfunction, metabolic disruption and endocrine disruption, require further investigation in other in-vitro contexts.

The choice of HEPG2 as our primary cell line model was driven by the abundance of chemical annotations for liver carcinogenicity and the appropriateness of HEPG2 for the study of liver toxicity. However, there are limitations in its use. Firstly, the expression of genes involved in phase I and phase II metabolism vary between passages and results relating to xenobiotic metabolism may be difficult to determine (Soldatow et al. 2013); this is also seen in the comparison of our genotoxicity-related signatures to Drugmatrix liver signatures. One potential contribution to this effect is the low bioactivation capacity in HEPG2 compared to in-vivo. As an alternative, the hepatoma cell line, HepaRG, which has a liver-like bioactivation, could be used as an in-vitro liver model for studying carcinogens and genotoxicants. One study has shown that while HEPG2 performs better in discriminating signatures between genotoxic and non-genotoxic carcinogens, HepaRG is a more suitable in-vitro liver model for biological interpretation of effects of chemical exposures (Jennen et al. 2010). Secondly, since HEPG2 is a cancer cell line, the exposures of carcinogens in this line may show differences as compared to a non-transformed cell line. For the purpose of predictive modeling, these cell line-specific nuances may be overlooked as long as the performance of the classifier is adequate.

Other cell line candidates for follow up studies should include more realistic hepatocyte models, such as induced pluripotent stem cells (iPSC)-derived hepatocytes, or organoids (Davidson et al. 2015; Underhill and Khetani, 2018). Alternatively, hepatic stem cells such as oval cells could be considered given that the stem cell theory of cancer initiation and maintenance is well supported (Fábián et al. 2013; Tan et al. 2006).

While liver carcinogenicity prediction was the adverse phenotype of choice for this study, this experiment provided us with many valuable insights to facilitate future experiments, including the logistics of large chemical panels procurement, and chemical and dose selection for tissue-specific carcinogenicity. It also set the stage for in-vitro based exposure studies of additional adverse phenotypes. For instance, we initiated the in-vitro screening of mammary gland carcinogenicity through the use of a non-tumorigenic human mammary epithelial cell line, MCF10A and p53-deficient MCF10A. The experimental and computational pipeline we established, paired with the cost-effective technology we used for chemical exposure and gene expression profiling, paves the way for the screening of large chemical panels for exposure-based experiments in other organ, disease, and adverse outcome contexts.

2.1.5 Conclusions

Long term tests for chemical carcinogens based on epidemiology and rat studies are expensive and time consuming and not feasible for scaling to a large number of chemicals. In this study, we detailed a high-throughput gene expression profiling of more than 300 liver carcinogens and non-carcinogens in a short term in-vitro exposure model. These gene expression profiles, given sufficient transcriptional bioactivity, were capable

of accurate prediction of long-term carcinogenicity and even more accurate prediction of genotoxicity. Pathway enrichment analysis revealed similarities between pathway level response captured by the short term in-vitro exposures and known MoAs of carcinogenesis, particularly genotoxic mechanisms such as DNA damage and repair.

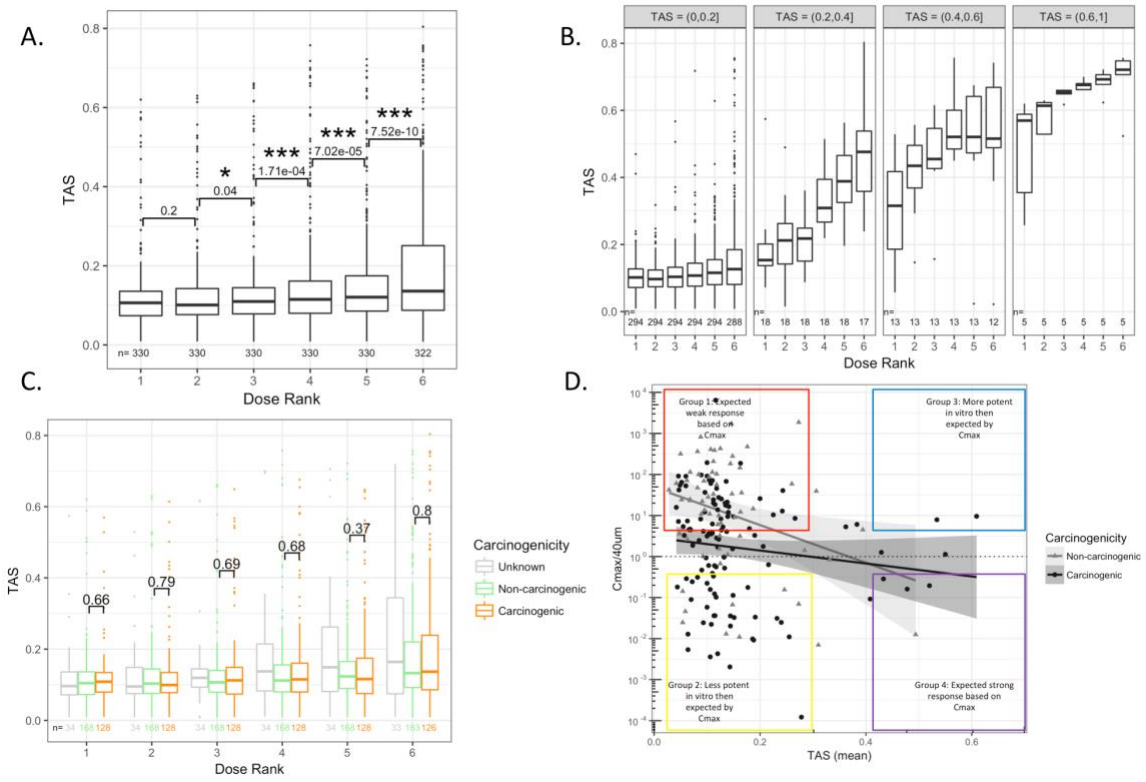


Figure 2.1.1 **Boxplot of Transcriptional Activity Scores (TAS) by sample subsets**

A. Boxplot of TAS distributions for each dose level (rank = 1 lowest dose, rank 6 = highest dose). Numeric labels indicate the significance of paired one-sided two-group TAS comparison between adjacent dose groups, adjusted for multiple comparisons across doses using the False Discovery Rate method (FDR) (* = FDR < 0.05, *** = FDR < 0.001) (see methods). **B.** Boxplot of TAS distribution for each dose level, binned by TAS subsets. **C.** Distribution of TAS grouped by chemical carcinogenicity within each dose level. P-values indicate the significance of unpaired one-sided two-group TAS comparison between TAS of carcinogenic chemicals and TAS of non-carcinogenic chemicals within each dose group (* = p< 0.05, ** = p< 0.01, *** = p< 0.001) (see methods). **D.** Scatter plot of mean TAS per chemical and the ratio of *equivalent in-vitro dose* (Cmax) over maximum in-vitro dose (40uM) (see methods for Cmax calculation). Boxplots in Panel A, B, and C 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.

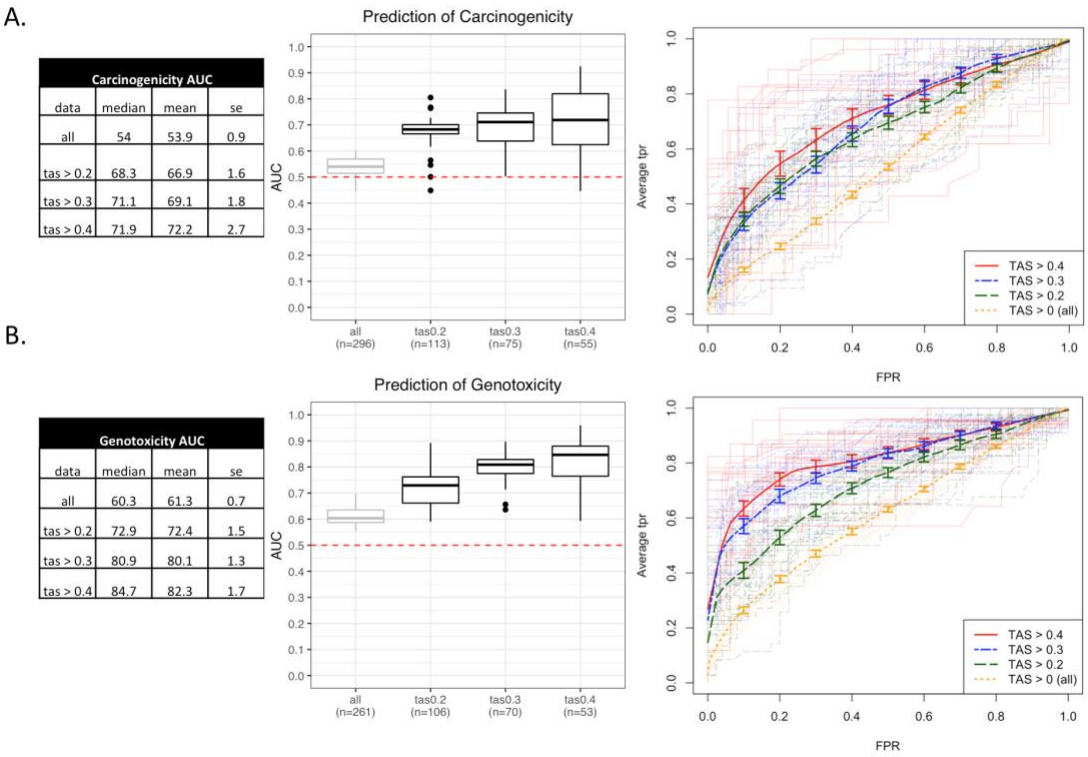
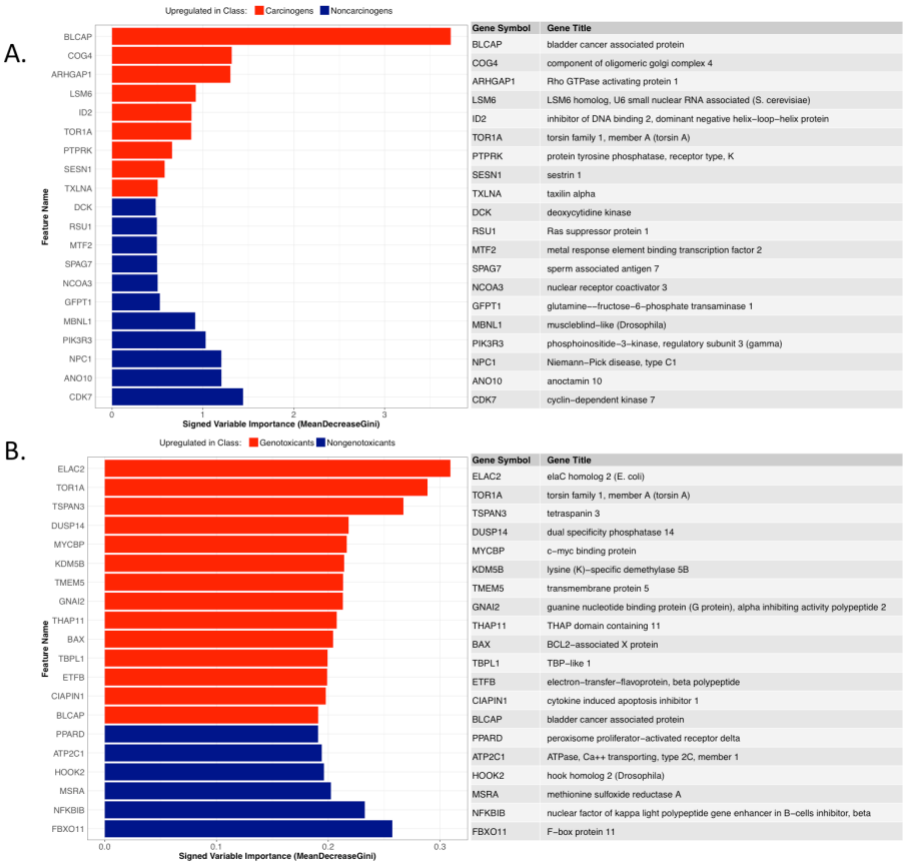


Figure 2.1.2 **Performance of classifiers in predictive models**

A. of carcinogenicity, and **B.** genotoxicity. From left to right: (1) Summary statistics tables of Area Under the Curve (AUC) for each Transcriptional Activity Score (TAS) subsets; data represented are the median, mean and se (standard error) of the AUC scores. (2) Boxplots of AUC across resamples (N = 25) for each TAS subset with the lower, middle, upper hinges corresponding to the 25th, 50th (median), and 75th percentile, the upper and lower whiskers extending 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. Dotted line at 0.5 represents the expected AUC of a random classifier. Labels in each TAS group ("n=") represent the number of unique chemicals in the model training and validation step. (3) Receiver operating characteristic (ROC) curves (False Positive Rate (FPR) vs. Average True Positive Rate (TPR)). Thick lines represent vertical averaging of ROC curves across resamples in each TAS group shown with bars denoting the standard errors. Thin semi-transparent lines represent ROC curves of individual resamples in each TAS group.



Gene Symbol	Gene Title
BLCAP	bladder cancer associated protein
COG4	component of oligomeric golgi complex 4
ARHGAP1	Rho GTPase activating protein 1
LSM6	LSM6 homolog, U6 small nuclear RNA associated (S. cerevisiae)
ID2	inhibitor of DNA binding 2, dominant negative helix–loop–helix protein
TOR1A	torsin family 1, member A (torsin A)
PTPRK	protein tyrosine phosphatase, receptor type, K
SESN1	sestrin 1
TXLNA	taxilin alpha
DCK	deoxycytidine kinase
RSU1	Ras suppressor protein 1
MTF2	metal response element binding transcription factor 2
SPAG7	sperm associated antigen 7
NCOA3	nuclear receptor coactivator 3
GFPT1	glutamine–-fructose–6–phosphate transaminase 1
MBNL1	muscleblind–like (Drosophila)
PIK3R3	phosphoinositide–3–kinase, regulatory subunit 3 (gamma)
NPC1	Niemann–Pick disease, type C1
ANO10	anoctamin 10
CDK7	cyclin–dependent kinase 7

Gene Symbol	Gene Title
ELAC2	elaC homolog 2 (E. coli)
TOR1A	torsin family 1, member A (torsin A)
TSPAN3	tetraspanin 3
DUSP14	dual specificity phosphatase 14
MYCBP	c–myc binding protein
KDM5B	lysine (K)–specific demethylase 5B
TMEM5	transmembrane protein 5
GNAI2	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2
THAP11	THAP domain containing 11
BAX	BCL2–associated X protein
TBPL1	TBP–like 1
ETFB	electron–transfer–flavoprotein, beta polypeptide
CIAPIN1	cytokine induced apoptosis inhibitor 1
BLCAP	bladder cancer associated protein
PPARD	peroxisome proliferator–activated receptor delta
ATP2C1	ATPase, Ca++ transporting, type 2C, member 1
HOOK2	hook homolog 2 (Drosophila)
MSRA	methionine sulfoxide reductase A
NFKBIB	nuclear factor of kappa light polypeptide gene enhancer in B–cells inhibitor, beta
FBXO11	F–box protein 11

Figure 2.1.3 **Top 20 landmark gene features of predictive models**

A. of carcinogenicity and **B.** genotoxicity as ranked by variable importance (Mean Decrease in Gini Index) in the predictive models of Transcriptional Activity Scores (TAS) > 0.4 subset.

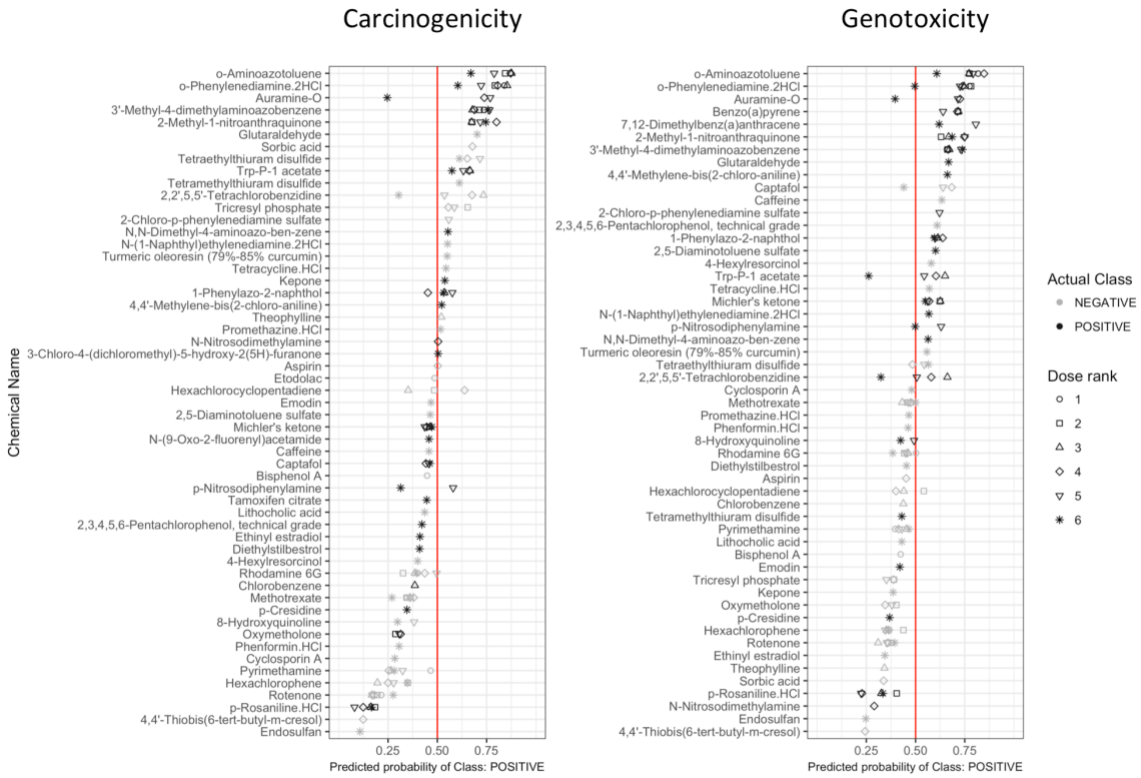


Figure 2.1.4 **Dot plot of probabilities of predicted classes for hold-out chemicals in the Transcriptional Activity Score (TAS) > 0.4 subset**

Point outline colors represent actual class labels (carcinogenic vs. non-carcinogenic, genotoxic vs. non-genotoxic). Point shapes represent dose ranks (dose rank 6 represents the highest dose level for each chemical). X-axis positions of points represent predicted probability of class "Positive" (carcinogenic in left column or genotoxic in right column), e.g. at the cutoff of 0.5 (red line), instances with values greater than 0.5 are predicted "Positive" and those with less than 0.5 are predicted "Negative".

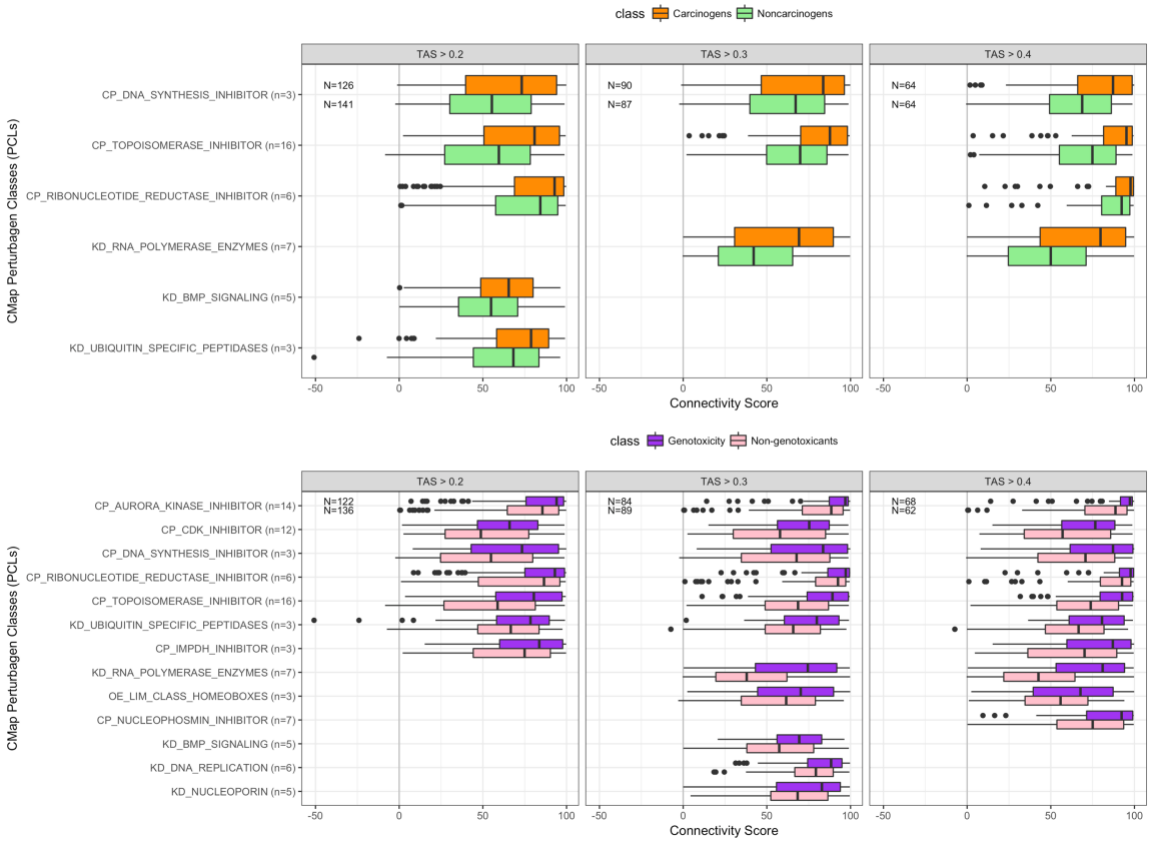


Figure 2.1.5 **Connectivity scores of top CMap Perturbagen Classes with differential connectivity (FDR < 0.05) to Carcinogens vs. Non-carcinogens and Genotoxicants vs. Non-genotoxicants grouped by Transcriptional Activity Scores (TAS) subsets**

The lower, middle, upper hinges of boxplots 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, and data points beyond the whiskers represented as dots.

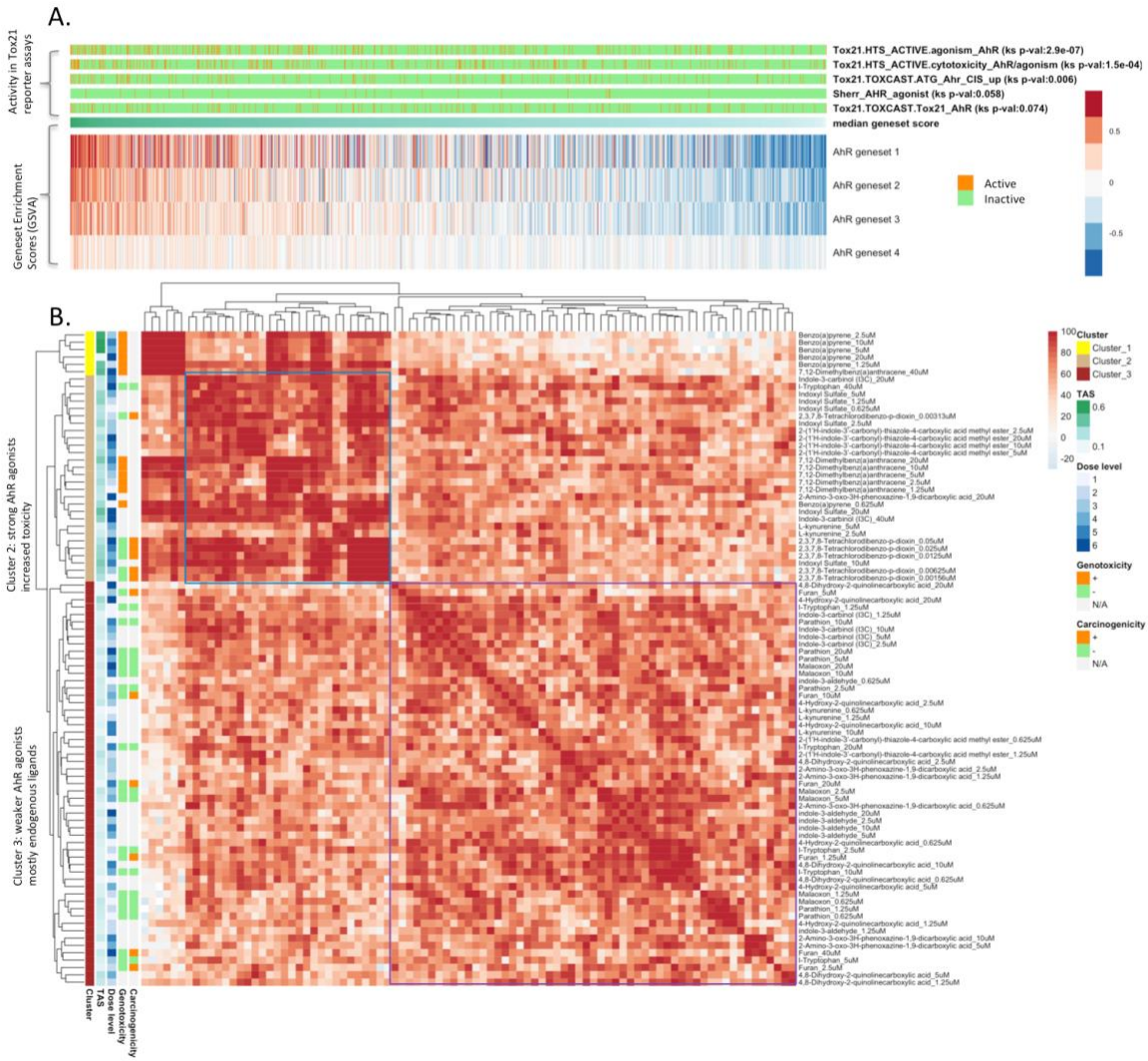


Figure 2.1.6 **Investigation of profiles of AhR related chemical perturbations**

A. Profiles with AhR activity ranked by median geneset scores of AhR target gene lists.
B. AhR-related profiles clustered by connectivity scores.

2.2 Network-based analysis of transcriptional profiles from chemical perturbations

The work in section was published:

Mulas F, Li A, Sherr DH, Monti S. 2017. Network-based analysis of transcriptional profiles from chemical perturbations experiments. BMC Bioinformatics 18: 130.

2.2.1 Introduction

High-throughput screening of gene expression data of chemical perturbations can help identify patterns of similarly behaving perturbations in terms of their biological effects in vivo or in vitro. One method of constructing and comparison biological networks under multiple conditions is through transcriptional network inference and comparison. Network models – with genes represented by nodes and gene-gene interaction represented by edges – are inferred from experimental data and manually curated repositories. One instance of network models is gene co-expression networks, where genes are connected if the corresponding gene is significantly co-expressed across a set of samples of a specific context. These networks are particularly useful as they represent a snapshot of gene co-regulation in the experiment under study (Zhang and Horvath, 2005).

In the context of co-expression network inference, one popular method widely used in the study of protein interactions is scale free networks (SFN), in which degree of connection of member nodes follow the power law (Barabasi, 2009). The construction of these networks usually relies on the computation of gene-gene correlations across replicate experiments, and on the subsequent thresholding of the absolute correlation

values so as to define as connected only those genes with correlation above a chosen threshold (Butte el al. 2000; Carter et al. 2004). Although this approach has proven extremely useful in identifying key hub genes in multiple biological conditions (Margolin et al. 2006), the high sensitivity of the obtained networks to the choice of threshold raises questions about the reproducibility of the obtained results, as well as about their biological meaning (Zhang et al. 2005, Carter et al. 2004). An alternative approach is to use Correlation Networks (CN), where all pairwise gene associations are considered, to avoid loss of information in those cases where the analysis focuses on the identification of groups of tightly connected genes (modules), rather than on the identification of single key nodes (hubs).

Regardless of the methodology used to infer the graph, network-derived gene modules can be investigated experimentally in order to gain insights into their biological function, or with the help of gene and pathway annotation resources. Additionally, the comparison of correlation networks from different conditions (e.g., different disease stages, or perturbations with different chemicals) may help identify modules whose connectivity is significantly altered in the compared conditions (Zhang et al. 2013). Connectivity-based comparisons may thus help identify “aggregate changes” that could be missed by standard methods of differential analysis comparing individual genes (Davis et al. 2009).

In this study, we describe the development of a network-based analysis pipeline and its application to gene expression datasets from chemical perturbation experiments, with the goal of elucidating the modes of actions of the profiled perturbations. We apply

our pipeline to the analysis of the DrugMatrix dataset from the National Toxicology Program (NTP) (Ganter et al. 2005), one of the largest toxicogenomics datasets available, which contain organ-specific gene expression measurements for model organisms exposed to hundreds of chemical compounds with varying carcinogenicity and genotoxicity.

Previous studies have shown that it is possible to infer highly accurate predictive models of chemical-associated long-term cancer risk from rat-based short-term toxicogenomics data, and to identify genes significantly associated with carcinogenesis (Gusenleitner et al. 2014). Here, we aim to go beyond the inference of predictive models and the identification of single biomarker genes, towards the identification of gene modules or pathways significantly associated with the profiled chemical perturbations and the induced adverse phenotypes. We do so by comparing the connectivity of gene modules in the networks derived from the control samples ("Control network") to those obtained from samples collected after the exposure to specific chemical compounds. To this end, we reconstruct chemical-specific transcriptional networks, and show that by grouping chemicals based on the similarity of their associated networks we can identify groups of chemicals or drugs with similar functions and similar carcinogenicity and genotoxicity profiles. We also show that the in-silico annotation by pathway enrichment analysis of the gene modules with a differential connectivity (i.e. showing a gain or loss of connectivity for specific groups of compounds) can point to the main molecular pathways induced by specific chemicals.

2.2.2 Methods

Data resources

The DrugMatrix (Ganter et al. 2005), available through the Gene Expression Omnibus (GEO) with the accession number GSE57822, contains gene expression profiles from male rat primary tissues (liver, kidney, heart and thigh muscle) and cultured rat hepatocytes, corresponding to treatments with 376 chemicals, and including 994 control samples from rats kept in matched conditions. Each compound was administered at multiple doses and durations (6 h - 7 days), and each combination of tissue, compound, time and dose was profiled in triplicate. Of the 376 chemicals tested, 255 were annotated with either carcinogenicity or genotoxicity information in the Carcinogenic Potency Database (CPDB) (Gusenleitner et al. 2014), corresponding to 3448 profiles. In our study, only the samples from liver were considered, both for controls (279 samples) and for chemical perturbations represented by at least 10 samples and including all doses and durations available.

Data processing

Both Affymetrix datasets were normalized using the R Bioconductor package frma and frmaTools [32]. The Median Absolute Deviation (MAD) was used as the variation filter to select the 7000 best-ranked probes whose expression was then considered for inference of transcriptional networks. Data normalization, gene selection, network inference and other analyses were performed with custom scripts developed using the programming languages R, and several Bioconductor packages.

Network inference and modules analysis

Transcriptional network inference starts by defining an adjacency matrix $A = \{a_{ij}\}$, with weight a_{ij} denoting the strength of the relation of genes i and j in the expression data. Scale-free transformations (thresholding) can then be applied to the correlation measurements to achieve a scale-free topology typical of biological networks, characterized by relatively few highly connected nodes (hubs) among a larger number of sparsely connected neighbors (Zhang and Horvath, 2005). In this work, we explored both the direct use of non-transformed correlation networks (CN) as well as of scale-free transformed networks (SFN).

In order to obtain a correlation matrix, Pearson correlation measures between all pairs of gene expression profiles were computed. In the CN approach, the correlation matrix was directly used as the adjacency matrix A. Conversely, in the SFN approach two additional steps were required: i) only those edges with correlation values exceeding a specific threshold were retained, with the threshold selected so that the resulting distribution of connectivities fitted a scale-free topology; and ii) the adjacency matrix A was computed by transforming the thresholded correlation values into a topological overlap matrix (TOM), which takes into account the indirect interactions between each couple of genes in the network (Zhang and Horvath, 2005). In both approaches, hierarchical clustering with Ward's method was then applied to the obtained adjacency matrix, and a dynamic tree cutting algorithm was applied to determine the number and composition of gene clusters, henceforth referred to as gene modules. We used the R package cutreeDynamic with minimum cluster size set to 10 genes, method set to "hybrid" and "deepSplit" parameter set to 4, which allows a higher number of more

homogeneous clusters if compared to other parameter settings. Finally, a Module Differential Connectivity score (MDC) was used to compare the connectivity of gene modules between networks (Zhang et al. 2013). For a specific module composed of N genes and an edge set EN, the MDC measures the ratio of the weighted cardinalities of EN in the two network, i.e.:

$$MDC(X,Y) = \frac{|EN_x|}{|EN_y|}$$

where $|EN_X|$ and $|EN_Y|$ denote the average connectivities a_{ij} X and a_{ij} Y among the module's genes within networks X and Y. MDC values below 1 represent a loss of connectivity of the module in X with respect to Y, while values exceeding 1 indicate a gain of connectivity.

Inference of Compound and Aggregate Compound Networks

The inference approach (CN) showing the best validation results was used to analyze how different compounds (or groups of compounds) affect the connectivity patterns of specific gene modules. All non-treated liver samples available were used to construct the Control Network, while for the treatment-related networks we relied only on chemical compounds for which at least ten replicate experiments (animals) were available, obtaining 62 Compound Networks.

In order to build Aggregate Networks representing multiple chemical compounds, the similarity of Compound Networks based on their module composition was measured by the adjusted Rand index (aRI) (Rand, 1971). The aRI is a well-accepted measure that allows for the comparison of clustering results even when these yield different numbers

of clusters. Groups of chemical compounds were then identified by applying dynamic tree cutting (Langfelder et al. 2008)(cutreeDynamic hybrid, minimum cluster size set to 3 compounds, "deepSplit" set to 4) to the hierarchical tree obtained by merging compounds, with the Ward method, based on their aRI similarity. For each of the 13 groups detected, an aggregate network was built using partial correlation, in place of simple correlation, as the adjacency measure, so as to control for the potential confounding effect of the chemicals grouped. The sample sizes of these groups ranged from 41 to 154, with an average of 86 samples used to infer correlation values. Internal similarity of compounds in each group was assessed by evaluating the overlap of their interacting proteins, as retrieved through the CTD database (Davis et al. 2015), by means of the Fisher exact test. Specifically, a p-value was obtained for each pair of compounds in a group and the median p-value was used as the score of internal similarity among the aggregated set of compounds. The significance of these measures was assessed by randomly selecting equally sized groups of compounds from the CTD database and computing their internal similarity, with the procedure repeated 1000 times.

Selection and annotation of significant modules

For each Control Network-specific module, a confidence interval for the value of the corresponding MDC with respect to each Aggregate Network (or Individual Compounds network) was computed by randomly selecting with replacement the same number of samples from the replicates of non-treated samples. After 1000 iterations, the standard value of each log-transformed MDC was compared with the obtained estimates

of MDC confidence interval. The resulting p-values assess the significance of the deviation of the MDC from 1 (with 1 denoting lack of differential connectivity).

The same bootstrap procedure was adopted for modules extracted from each Aggregate Network (or Individual Compounds network), this time by randomly selecting with replacement the same number of samples from the entire set of perturbations.

In order to rank the control modules most specifically altered by each compound group, the changes in connectivity of each module m measured for a compound group g i with respect to the Control Network c were compared to those obtained for the other compound groups (Tawa et al. 2014). First, the absolute value of the difference in the MDC scores between two groups of compounds gi and gj was computed for each module m as:

$$\Delta m_{gi,gj} = \left| logMDC_m\,(g_i, c) - \log MDC_m(g_j, c) \right|$$

This was used to compute the specificity of module m to compound group g $_i$, as:

$$Sp(m)_{gi} = \sum_{k=1}^{N} \Delta m_{gi,gk}$$

where N denotes the total number of compound groups different from g_i. For each compound group, modules with score exceeding the top 5th percentile of the overall distribution of specificity scores were selected for enrichment analysis.

Specificity of modules inferred from the Aggregate Compounds Networks was assessed based on two alternative criteria, depending on the frequency of observation of the same module (or a highly overlapping module, Fisher test p-value < 0.01) in the networks. Modules identified in more than 50% of the aggregate networks were labeled

as "high frequency" and a score $Sp(m)_{gi}$ for module m to compound group g_i, was computed as described for the Control Network modules. The Specificity score was obtained as $\frac{Sp(m)_{gi}}{N_{tot}}$ where N_{tot} denotes the total number of compound groups. For each compound group, modules with score exceeding the top 5th percentile of the overall distribution of specificity scores were selected as "high frequency" specific modules.

Modules identified in less than 50% of the networks were labelled as "low frequency" and selected based on significance of their correspondent MDC values ($p < 0.01$).

Both Control Network-derived and Aggregate Network-derived (low and high frequency) specific modules were annotated by enrichment of the hallmark gene sets part of the MSigDB compendium (Liberzon et al. 2011). Significance of pathway enrichment was computed using a hyper-geometric distribution-based test and corrected for multiple hypothesis testing across multiple pathway gene sets via the false discovery rate (FDR) estimation. Hallmark pathways with FDR ≤ 0.25 was reported for each specific gene module in both Control Networks and Aggregate Networks.

Selected genes and Hallmark gene sets were compared with the "Perturbational Transcriptome", a list of genes identified as significantly differentially expressed (with respect to matched controls) in at least five compounds. Each module was tested for enrichment of genes included in the Perturbational Transcriptome by a hyper-geometric test.

2.2.3 Results

Differential connectivity analysis of chemical perturbations

As shown in Figure 2.2.1, the network-based pipeline for differential connectivity analysis starts by inferring chemical-specific Compound Networks, obtained from samples collected after the exposure to specific chemical compounds (Figure. 2.2.1.a), and a network from the control samples, hereafter named "Control Network" (Figure. 2.2.1.b). Groups of compounds are then identified based on the similarity of their individual network structures. For each group, a new "Aggregate Compound Network" is inferred by pooling all the samples across the clustered compounds (Figure. 2.2.1.c-e). Next, modules of tightly connected genes are identified in each of the constructed networks and compared between conditions (i.e., control vs. compound group) in terms of Module Differential Connectivity (MDC) (Figure. 2.2.1.f). Given a module identified in one of the two networks under comparison (e.g., the aggregate compound network), the MDC score is computed as the ratio of the average connectivity across all the genes within the module in the aggregate network (numerator) and in the control network (denominator). This score represents changes of connectivity in the Compound group with respect to the Control. MDC scores are computed for all modules in the networks, and tests of statistical significance and module specificity are performed to identify modules that will be further investigated through enrichment analysis based on pathway repositories and additional annotation sources (Figure. 2.2.1.g-h).

Reproducibility analysis for network inference

The inference and comparison methods were evaluated on networks derived from independent sample subsets extracted from the same dataset by a resampling approach. In this case, we attained more reproducible results, with the distribution of MDC values correctly centered at 1, and with a lower variance when using the correlation network (CN) approach rather than the Scale-Free Network (SFN) approach. An additional advantage of the simpler CN approach is that it does not require the selection of a threshold for the correlation values, a choice that is highly sensitive to the samples analyzed and strongly influences the subsequent calculation of MDC values. Based on these results, the subsequent analyses were all based on the CN approach as the network inference method of choice.

Groups of similar chemicals can be inferred by network analysis

Using the CN approach, we analyzed how different groups of compounds affect the connectivity pattern of specific gene modules. First, we inferred the network from the non-treated liver samples, hereafter named “Control Network”. The Control Network was clustered into 60 gene modules, with sizes ranging from 21 to 551 genes.

Next, chemical-specific Compound Networks were inferred for each of the 62 chemicals for which at least ten replicate experiments (animals) were available. We aimed at identifying chemicals with similar network structure, which were then grouped to infer “Aggregate Compound Networks”. This step had two main goals: i) to study how well groups of chemicals with similar known features (e.g., similar mechanism of action) can be identified through networks; and ii) to increase the sample size available for

network inference. For the aggregation of the chemicals, the Compound Networks were compared pairwise based on the similarity of their respective modules as measured by the adjusted Rand index (aRI) (Rand, 1971). The aRI is a score specifically devised for the comparison of clustering results even when the two networks have different number of clusters (i.e., modules). Hierarchical clustering was then applied to the matrix of aRI's to induce a similarity-based partial ordering and grouping of the chemicals (Fig. 2.2.2.a).

The aRI-based clustering yielded a clear separation between non-genotoxic carcinogens (left sub-dendrogram in Fig. 2.2.2.b) and genotoxic non-carcinogens (right sub-dendrogram). Of notice, genotoxic compounds were not as well separated when we applied alternative, more standard clustering approaches, such as one based on the direct similarity of the chemicals' expression profiles.

Module differential connectivity highlights chemicals' modes of action

Samples related to the 13 groups identified (Figure 2.2.2.b) were used to infer Aggregate Compound Networks, each representing the partial correlation among genes across all replicate experiments from the group of compounds considered. As shown in Fig. 2.2.1, we aimed at comparing the Control Network with multiple perturbations by analyzing the changes in gene modules connectivity. This analysis was repeated twice, first using the modules identified in the Control Network, and then using the modules identified in each of the Aggregate Compound Networks.

Control network-centered analysis

For each of the modules identified in the Control Network, the MDC score was computed to measure the change in connectivity (gain or loss) among the module's genes

due to the action of each group of chemicals. Significance of the MDC values was assessed by a bootstrap approach, whereby a confidence interval for each MDC is estimated by performing network inference on bootstrapped (i.e., sampled with replacement) versions of the original dataset.

Since the modules are defined in the control network, hence are the same for each pairwise comparison, the results can be represented as a matrix and an associated color-coded heatmap, with each row corresponding to a Control Network module, and each column corresponding to a compound group. Several modules manifest a remarkable change of connectivity, as captured by their MDC scores, and as confirmed by their estimated q-values.

In order to focus on compound-specific effects, we computed a "Specificity Score" for each Control module (Tawa et al. 2014). The specificity score quantifies the uniqueness of a gain or loss of connectivity to a given compound group. Briefly, for a given group of compounds and a given module, the differences in MDC between that group of compounds and all the other groups are computed. Specificity is then defined as the sum of all the differences, with higher values identifying modules with a high MDC absolute value relative to all the others. Modules with scores exceeding a top percentile of the distribution of Specificity values were subject to enrichment analysis and significant pathways were selected with FDR-corrected p-values. A bipartite graph in Fig. 2.2.4 is used to graphically represent the obtained associations between compound groups and enriched gene sets. All groups except one (G10) showed at least one top

specific module significantly enriched for Hallmarks gene sets (Enrichment FDR-corrected p-value < 0.25).

Aggregate network-centered analysis

Taking an approach complementary to the one adopted in the control network-centered analysis, here a set of gene modules is defined for each of the aggregate compound networks. That is, for each aggregate network, a set of densely connected modules is identified, and their change of connectivity with respect to the control network is calculated by MDC. Since a potentially distinct set of modules is identified in each aggregate network, this precludes the representation of the differential connectivity analysis results across the aggregate networks in matrix form.

In this analysis, we first identified high frequency (HF) modules as those modules for which similar grouping of genes (i.e., similar composition) was found across multiple aggregate networks (Fisher test, $p < 0.01$). A Specificity Score was then computed to highlight those with MDC values specific to a particular compound group. We next identified, low frequency (LF) modules, i.e., modules whose composition was unique to only one or few Aggregate Networks. Both HF and LF modules are graphically represented in Figure. 2.2.5, where Hallmarks gene sets have been used to investigate their biological function. Taken together, the findings described below confirm that the approach is capable of identifying known modes of action (Waters et al. 2010), and of grouping compounds based on their coordinated effect on molecular pathways.

Comparison of network-based approach to standard differential expression analysis

A comparison of our network-based analysis results with those from standard differential expression analysis highlighted the complementarity of the two approaches. Differentially expressed genes identified as belonging to the "Perturbational transcriptome" in our previous study (Gusenleitner et al. 2014), and their correspondent enriched Hallmarks, were evaluated in terms of their overlap with genes and gene sets identified by the present approach. First, each of the Control and Compound-related modules was scored for its enrichment in terms of differentially expressed genes. As expected, a majority of the modules (26 out of 32) were significantly enriched for differentially expressed genes (Fisher test, $p < 0.01$). Table S6 shows a summary of this analysis where all genes contained in at least one of the modules identified were compared with the Perturbational transcriptome. Despite the significant overlap yielded by the two approaches, a considerable number of genes were identified only by one of the methods, pointing to the complementarity of the approaches. Interestingly, many of the genes and Hallmarks identified only with the network-based approach were associated to pathways previously implicated in mediation to chemical responses, and described in the following section, including Heme Metabolism, Myc targets, and inflammation signaling.

2.2.4 Discussion

Both Control-centered and Compound-centered network based analyses yielded biologically meaning findings that reflects known or novel mechanisms of the chemical perturbations, as discussed in detail below.

Alcohol-induced liver inflammation

As shown in Fig. 2.2.3, G1 has a specific effect on a high number of pathways, mostly associated with inflammation and tumorigenesis. Given that liver samples have been considered in this study, the significant impact of this group on the entire pathway set could be explained by the high number of alcohols included in group G1. In fact, alcohol-mediated activation of inflammation signaling pathways in the liver is known to increase tumorigenesis in mice and to activate pro-inflammatory cytokines, such as tumor necrosis factor alpha (TNF-α), interleukin 6 (IL-6), and nuclear factor kappa B (NFκB) (Wang et al. 2010). This liver damage response has been reported for several members of the "G1-Solvents" group, including allyl alcohol (Lee et al. 1996), Lipopolysaccaride, known as an endotoxin, (Wang et al. 2010), and chloroform (Gupta et al. 2003).

Hypolipidemic compounds induce cholesterol metabolism and inflammation

Groups G3 (Statins) and G5 (Fibrates), both including hypolipidemic compounds, show a specific alteration of modules related to cholesterol and fatty acid metabolism. In particular, the highest Specificity score is obtained by the Fibrates on module "thistle1", enriched for Fatty Acid Metabolism. While all the other groups of chemicals cause a LOC of the "thistle1" module, groups G5 and G3 are the only ones to produce a GOC, with a higher MDC obtained by Fibrates. Statins have a more significant effect on the module "honeydew", enriched for cholesterol homeostasis (Table S7), which has not a high specificity ranking for the Fibrates, confirming different actions of those two distinct classes of drugs. Statins are also associated with stress response pathways, including oxidative phosphorylation, UV response and activation of TNF-α in both Control (Figure.

2.2.3) and compounds-related modules (Figure. 2.2.4). While hypoxia-inducible factors were found to play a role in the inhibition of cholesterol synthesis (Schröder et al. 2011), an inflammatory response of the liver has not been clearly reported in the literature. However, the indication of liver damage as a rare side effect by FDA might suggest that tissue-specific network analysis could capture most of the possible mechanisms induced by drugs exposure.

Effect of estrogens, steroids and cancer drugs on cellular replication

Both modules related to estrogens and to cancer treatment have an effect on the connectivity of module "maroon", enriched for pathways related to cellular replication. While this module is gaining connectivity as an effect of proliferation-inducing drugs contained in group G4 (estrogens), a loss of connectivity is observed for G9 (chemotherapeutics) and G11, pointing to the disruption of the cellular replication machinery caused by anti-cancer drugs. The gain of connectivity of G2M checkpoint induced by G10 (Alkylating-cancer) and p53 pathway in the compound-related modules (Fig. 2.2.4) also confirms known mechanisms in the treatment of tumors and DNA damage (Kastan and Bartek, 2004).

The loss of connectivity of inflammation-related pathways observed on G6 perturbation could be explained by the known action of some steroids-related compounds.

Non-homogeneous groups of compounds have known common effects

Interesting results can be observed for non-homogeneous groups of compounds for which a predominant pharmacological action could not be assigned. In particular,

group G13 (Antiseptics-Estrogens) shows a double effect, clearly visible in Fig. 2.2.3, by acting on inflammation-related pathways and on mitosis gene sets. Possible key players of this group are Safrole, which has been shown to induce liver DNA damage (explaining the action on stress response pathways) (Ding et al. 2015) and Methyl salicylate, shown to have an estrogenic potential (Zhang et al. 2012), thus increasing the coordinated activity of genes related to cellular replication pathways. Another interesting example is group G7, containing four chemicals with apparently different functions. As suggested by the specific loss of connectivity induced on module "coral1", related to inflammatory response, the majority of compounds included in this group have antioxidant properties (Aviram et al. 1998; Jamdade et al. 2016; Pigoso et al. 2002). Among these, Atorvastatin is a compound belonging to the class of Statins, which was not grouped with the other Statins in G3. Interestingly, no other statins except for one were demonstrated to have antioxidant effects in vitro, confirming the grouping of compounds found with aRI (Aviram et al. 1998; Schröder et al. 2011).

2.2.5 Conclusion

We have presented a pipeline for transcriptional network inference and comparison that was primarily designed for the analysis of chemical perturbations from high-throughput transcriptional screening experiments. Here, we applied it to the analysis of gene expression profiles from rat-based chemical exposure experiments. We show that groups of chemicals with similar functions and carcinogenicity/genotoxicity profiles can be identified through our proposed pipeline. In addition, modules with altered connectivity due to the action of specific compounds were enriched for pathways actually

related to the chemicals' action. These findings highlight potential advantages in the application of this network-based approach. In the context of drug discovery (or repositioning), the methods presented here could help assign new functions to novel (or existing) drugs, based on the similarity of their associated network with those built for other known compounds. Additionally, networks with patients as nodes could be compared with the same tools in order to identify groups with a similar response to a set of drugs. In fact, the proposed methodology has broad applicability beyond the uses here described and could be used as an alternative or as a complement to standard approaches of differential gene expression analysis.

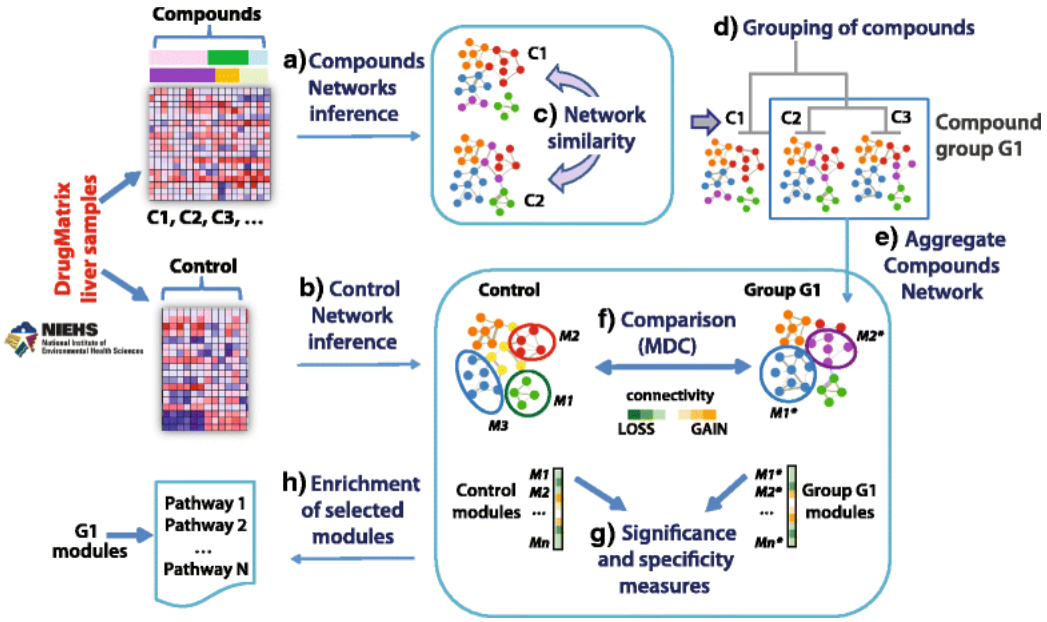


Figure 2.2.1 **Workflow of network-based analysis of transcriptional profiles from chemical perturbations**

DrugMatrix liver samples are used to infer chemical-specific Compound Networks (**a**) and a Control Network (**b**). Similarity in terms of network structure is evaluated to identify groups of compounds, whose samples are pooled to infer Aggregate Compound Networks (**c-e**). Modules of tightly connected genes both in the Control network and in

Compound Aggregate networks are identified and compared across conditions in terms of Module Differential Connectivity (MDC) (**f**). Modules with a change in connectivity that is highly specific to each compound group are investigated through pathway enrichment analysis (**g-h**)

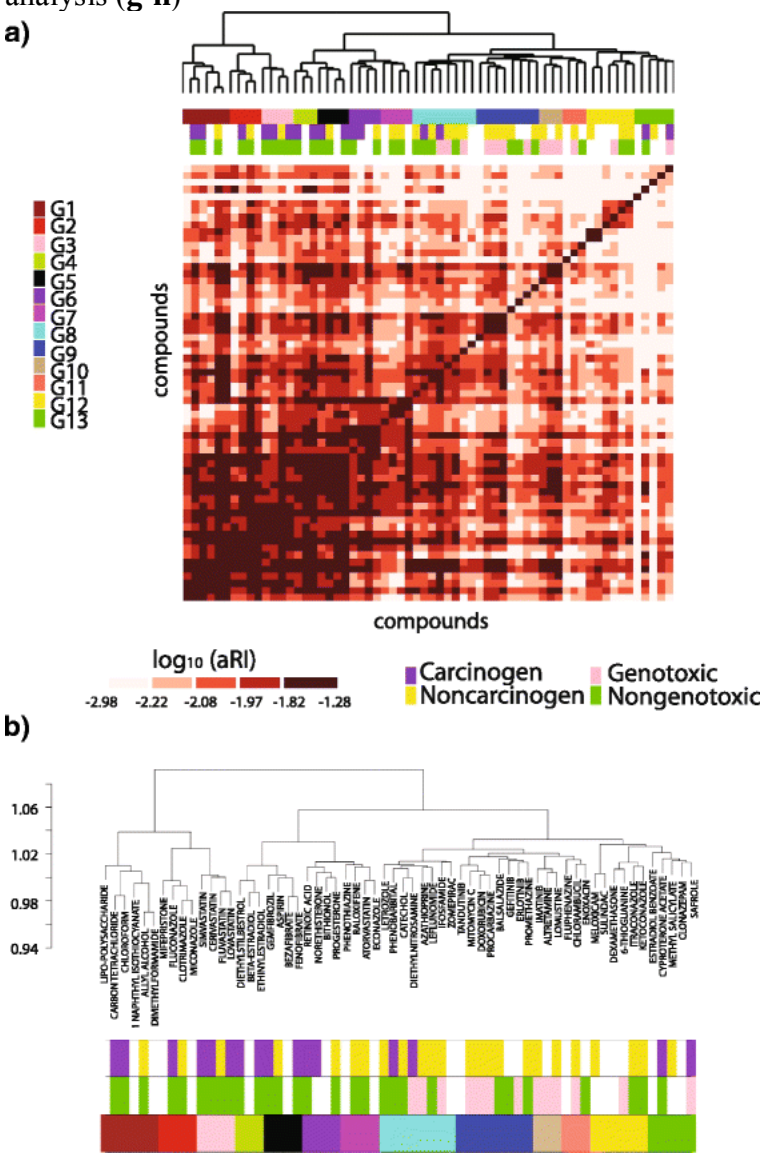


Figure 2.2.2 **Compounds aggregation**

Similarity of 62 chemical compounds based on adjusted Rand Index (aRI). **a.** Heatmap of aRI and grouping of compounds with similar networks structure. **b.** Zoom-in on the compounds grouping

Figure 2.2.3 **Enrichment of specific Control modules**

Bipartite graph representing associations between compound groups and enriched Hallmarks gene set corresponding to specifically altered modules extracted from the Control Network

Figure 2.2.4 **Enrichment of specific Compounds-related modules**

Bipartite graph representing associations between compound groups and enriched Hallmark gene sets corresponding to specifically altered modules extracted from each Aggregate Compound Network

CHAPTER THREE: Towards cancer therapy - Molecular characterization of the cancer genome and epi-genome using integrative analysis

3.1 Introduction

A central goal of cancer genomics research is to identify the key genetic and epigenetic fingerprints that promote the initiation and progression of cancer. These fingerprints can manifest across multiple biological levels quantifiable by high-throughput profiling technologies, such as RNA and protein expression, copy number, DNA methylation and microRNA profiling.

Large-scale cancer genomics data compendia such as The Cancer Genome Atlas (TCGA) have collected comprehensive multi-omics datasets for tens of thousands of patients across ~30 types of cancer (Cancer Genome Research Network, 2013). The availability of such large-scale datasets provides an opportunity for method development to integrate data from multiple types of profiling platforms towards the discovery of novel diagnostic and prognostic biomarkers and therapeutic targets.

Past research using integrative approaches have shown success in discovering novel cancer drivers (Akavia et al. 2010; Xie et al. 2012). However, the existing methods still rely on ad-hoc, albeit sophisticated, analysis methods and scripts not accessible to analysts other than those responsible for their development. Furthermore, the generated analysis results are often static, and not accessible in an interactive fashion. The approach here presented aims to address both these shortcomings.

The central hypothesis behind integrative approaches is that the key molecular fingerprints of cancer manifest through multiple layers of genetic and epigenetic markers.

Integration of these layers provides greater power to detect relevant disease markers, such as cancer drivers or potential therapeutic targets. An important type of genetic alteration in cancer is somatic copy-number alteration (SCNA). SCNAs harbor many known cancer drivers and play an important role in cancer initiation and progression through activation of oncogenes and inactivation of tumor suppressors (Zack et al. 2013). Identification of novel SCNA-associated cancer drivers is complicated by the fact that each SCNA contain many genes, the majority of which are likely not to have any critical functional effects. One approach to tackle this problem is to prune the set of candidate drivers based on their association with other types of paired omics data, such as gene expression profiles. For example, one can prioritize genes found in frequent SCNA peaks whose gene expression changes are associated with corresponding copy number changes (Lai et al. 2017). Even after this pruning step, the set of remaining candidate drivers might still yield too many testable hypotheses for use in functional validation studies. More importantly, association between SCNA and gene expression alone may not be the best metric for ranking potential cancer drivers.

To address this problem, we present a methodology, and an associated software package, to identify SCNA associated genes and to perform prediction of cis gene drivers prioritized by their capability of mediating downstream gene expression trans effects, that is, by combining information from trans genes (genes outside the SCNA of interest) whose expression is also associated with a particular SCNA event of interest. The approach is predicated on the hypothesis that SCNA-related drivers of tumorigenesis will mediate a larger proportion of the downstream effect observable by trans gene expression

than non-drivers. Using this heuristic, we identify putative SCNA-associated cancer drivers as the cis gene mediating the most trans gene expression, although the method allows users to customize the set of trans genes considered for mediation analysis.

We developed a corresponding software tool, *integration of Epi-DNA and Gene Expression* (iEDGE), for prediction of (epi-)DNA related cancer drivers. Preliminary versions of iEDGE have shown success in uncovering SCNA-associated cis driver genes in diffuse large B cell lymphoma (Chapuy et al. 2018, Monti et al. 2012). Here, we utilized iEDGE for predicting SCNA-associated cancer drivers across 19 cancer types using the data from TCGA, with a particular focus on analysis of TCGA breast cancer. Our list of candidate drivers is highly enriched for known oncogenes and tumor suppressors and additionally implicates many suspected drivers as well as novel candidate genes with potential prognostic or therapeutic importance in cancer.

3.2 Methods

An overview of the iEDGE approach is summarized in Figure 3.1. Briefly, iEDGE integrates samples quantified from a gene expression profiling assay paired with another genomic or epi-genomic assay capturing information upstream of gene expression, such as SCNAs, DNA methylation, or microRNA expression. First, we identify the features mapping from the epi-DNA assay to gene expression. In the case of SCNAs, cis genes of each SCNA are defined as genes within the focal peaks of the SCNA, and trans genes are defined as genes outside of the focal peaks. Then, iEDGE performs differential expression analysis to identify cis and trans genes significantly associated with each SCNA and, optionally, pathway enrichment analysis of each

significant cis and trans gene sets (Figure 3.1A). Next, iEDGE predicts cis driver genes using mediation analysis, wherein each differentially expressed cis gene is ranked by the number of differentially expressed trans genes it mediates as determined using the Sobel test of mediation (Figure 3.1B).

Copy number and gene expression pre-processing

We utilized the dataset of somatic copy number alterations and RNA-seq from the TCGA breast cancer cohort, which was preprocessed using Firehose v0.4.13 and downloaded from the Broad Institute TCGA GDAC repository (http://gdac.broadinstitute.org/runs/).

The SCNA dataset was preprocessed using GISTIC2.0 under the Firehose run release analyses__2015_08_21, which identified 29 significant focal amplifications and 40 significant focal deletions to be considered for integrative analysis (Broad Institute TCGA Genome Data Analysis Center 2015). SCNA status by sample was binarized using the amplitude threshold of 0.1, that is, SCNA status = 1: $t < 0.1$ or 0: $t >= 0.1$ for amplifications and 1: $t< -0.1$ or 0: $t >= -0.1$ for deletions. Cis genes were identified by GISTIC2.0 as genes in the wide peak of each significant SCNA with boundaries selected at the confidence level of 0.99. Trans genes were identified as genes outside the wide peak of each SCNA.

The gene expression data is a RSEM processed gene expression matrix (stddata__2015_06_01). Expression values were $\log_2$-transformed prior to integrative analysis. The samples were categorized into breast cancer subtypes using a combination of the pam50 classifier (Parker et al. 2009) and the *HER2* status. *HER2* status was

determined using *HER2* receptor activity, labeled positive if tested positive by either FISH or IHC method. In *HER2* negative samples, the pam50 classification was used. Samples with gene expression-based membership in one of four major breast cancer subtypes (Luminal A, Luminal B, Her2, Basal) were retained for integrative analysis. Tumors classified as "Normal-like" by pam50 were removed from further analysis. A total of 1050 samples (primary solid tumors only) were found with paired gene expression and SCNA data by matching the sample barcode identifier (combination of patient id and sample type).

Determining significantly expressed SCNA associated cis and trans genes

We performed differential expression of cis and trans genes with respect to each GISTIC2.0 defined significant focal SCNA peak. The significance of differential expression was estimated using limma (Ritchie et al. 2015) for each SCNA with samples split into two groups (amplified vs. normal for amplification peaks and deleted vs. normal for deletion peaks) using FDR < 0.25 and fold change > 1.2 for cis genes, and FDR < 0.01 and fold change > 1.5 for trans genes. One-sided significant levels were reported for cis genes with the rationale that a focal amplification is commonly associated with an increase in gene expression and a deletion is associated with a decrease in gene expression. Two-sided significant levels were reported for trans genes, as indirect downstream effects can occur through either transcriptional repression or activation.

Pathway enrichment analysis of significant cis and trans gene sets

Significantly differentially expressed cis and trans gene sets were tested for pathway enrichment using the MSigDB gene set compendia hallmark (hallmark gene

sets), c2.cp (curated gene sets from online pathway databases), and c3 (motif gene sets), version 5.0 (Liberzon et al. 2011). The significance of pathway enrichment was determined using a hypergeometric distribution-based test and corrected for multiple hypothesis testing using the False Discovery Rate (FDR) method (Benjamini and Hochberg 1995).

For breast cancer-specific pathway enrichment results, pathways with significant enrichment (FDR < 0.25) in any SCNA were reported (Figure 3.2). In addition, each SCNA was labeled according to its over-representation in a particular breast cancer subtype using a one-sided Fisher exact test comparing counts of SCNA occurrence within vs. outside each breast cancer subtype (FDR < 0.05). Subtype-specific SCNAs were subsequently used in conjunction with pathway enrichment results to determine subtype-specific pathway enrichments using a one-sided Fisher exact test (FDR < 0.05).

Mediation testing and prediction of cis drivers

To elucidate which cis genes are likely to mediate the association between copy number alteration and trans gene expression, we used the Sobel test to estimate the mediation effect of each cis gene and its significance (Sobel 1982). Briefly, we model the association for each triplet of SCNA, cis gene, and trans gene using the linear regression models specified in Figure 3.1B. The mediation effect of the cis gene: $\Delta\tau = \tau - \tau'$, represents the change in the magnitude of the effect of the SCNA status on the trans gene expression after controlling for the cis gene expression. The significance of the mediation effect is calculated from the t statistic: $t = \Delta\tau / SE$, where SE is the pooled standard error

term, and is compared to the normal distribution to determine the p-value and FDR (Benjamini and Hochberg 1995).

An important simplifying model assumption is that the association between each SCNA and trans gene is mediated by at most one cis gene, therefore the cis gene mediator for each SCNA-trans gene pair is chosen based on the most significant mediation effect (ranked by the FDR values of the mediation test).

Once the cis genes mediator is determined for each unique trans gene, the mediation effect of cis gene *i* on trans gene *j* can be expressed as either binary (0 or 1) or as the weight $w_{ij} = \frac{\Delta\tau_{ij}}{sign(\tau_{ij})\times\tau_{ij}}$, limited to the range of [0, 1]. Thus, the total mediation effect of cis gene *i* across *m* significantly expressed trans genes, also referred to as the *Weighted Fraction of Trans Mediation* (WFTM), is expressed as $M_i = \sum_1^m w_{ij} \times I_{ij}$, where I_{ij} denotes the indicator variable taking the value 1 if cis gene *i* is has the most significant mediation effect on trans gene *j* among all cis genes, 0 otherwise.

Next, for each SCNA, we rank each cis gene based on its total mediation effect M_i. The cis gene with the highest value of M_i is denoted as the "Rank-1 cis gene" for the given SCNA, the candidate driver gene of the alteration.

Assessing enrichment of predicted cis drivers in databases of known cancer drivers

To investigate the functional impact of putative drivers identified by iEDGE, we tested for the enrichment of iEDGE predicted driver genes in several cancer driver databases. Reference cancer driver genes, denoted as either “oncogenes” or “tumor suppressors” in the original sources, were compiled using data from Tuson Explorer (Davoli et al. 2013), Online Mendelian Inheritance in Man (OMIM) (Hamosh et al.

2005), Cancer Gene Census (CGC) from Catalogue of Somatic Mutations In Cancer (COSMIC) ("Cancer Gene Census" 2018; Forbes et al. 2017), and Uniprot (The UniProt Consortium, 2017). We tested for the overrepresentation among Rank-1 cis genes, compared to non-Rank-1 cis genes, of known drivers from the reference databases (Table 3.2). Enrichment tests were conducted separately for each reference database and driver type, i.e., oncogene ("_OG"), tumor suppressor ("_TN"), or both ("_COMBINED"), as well as using the union of the driver genes across databases (column "ANY" indicates union of drivers across knowledge bases). Enrichment significance was calculated using a one-sided Fisher exact test assessing the overrepresentation of Rank-1 vs. non-Rank-1 cis genes with respect to their membership in the reference driver list, conditional on the direction of change, e.g. amplified cis genes among oncogenes, deleted cis genes among tumor suppressors, or direction insensitive. P-values are adjusted with the FDR procedure to correct for multiple hypothesis testing across 19 tumor types.

Copy number-associated gene dependencies

SCNAs often lead to overexpression of driver oncogenes and confer a tumor-promoting environment. In other words, driver oncogenes are more likely to act as essential genes (increased gene dependency) in an amplified state. To identity such genes, we looked for copy number associated gene dependencies using data available from DepMap, specifically, gene dependency data (McFarland et al. 2018) and cancer cell line genomics data from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al. 2012; Cancer Cell Line Encyclopedia Consortium 2015). In particular, we mined for genes with associations between gene dependency (Combined RNAi screens from Broad, Novartis,

Marcotte) and somatic copy number status across cell lines (CCLE). To do this, we used a linear regression model $Y = \alpha X + \beta$ where Y is the Gene Dependency Score and X is the copy number level (log2 relative to ploidy) across cell lines. Gene dependency scores were calculated using DEMETER2 (McFarland et al. 2018). A negative gene dependency score corresponds to high gene dependency, e.g., an increased gene essentiality. In contrast, a high gene dependency score corresponds to non-essential genes. We looked for genes with significantly negative association between copy number and Gene Dependency Score, that is, genes in which higher copy number is associated with higher gene essentiality, using a one-sided t-test on the coefficient α (alternative hypothesis $\alpha < 0$). Additionally, FDR correction was performed across p-values for all genes. Since gene dependency scores are calculated from only gene knockdowns, we were only able to test amplification-driven gene dependencies, whereas overexpression assays would be needed for detection of deletion-driven gene dependencies.

Finally, to determine if iEDGE was able to uncover an enrichment of amplification-driven gene dependencies, we tested for enrichment of genes with amplification-driven gene dependencies among iEDGE-predicted Rank-1 cis genes using a one-sided Fisher test on the contingency table of counts (rows: membership in Rank-1 vs. non-Rank-1, columns: presence vs. absence of amplification driven gene dependencies).

Pan-Cancer Analysis

TCGA gene expression and copy number (GISTIC2.0) data were retrieved using Firehose v0.4.13 for 19 cancer types as summarized in Table S9. Gene expression data

(RNASeq) correspond to the latest release at the time of retrieval (stddata__2015_02_04 for cancer types ACC, KIRP, THCA and stddata__2016_07_15 for all other cancer types). SCNA copy number data uses the GISTIC2.0 run corresponding to Firehose run release analyses__2016_01_28. Gene expression processing and copy number processing steps for the pan-cancer analysis are consistent with methods used for the BRCA-only analysis. Of note, the BRCA dataset in the pan-cancer analysis includes all TCGA BRCA samples with paired copy number and gene expression data to be consistent with processing of other TCGA cancer types, contrary to the removal of samples without an assigned molecular subtype in the BRCA-only analysis.

We tested for enrichment of known cancer driver genes among Rank-1 cis genes in each of the 19 TCGA cancer types using a one-sided Fisher test (Table S3.3), consistent with the BRCA-only analysis. FDR correction was performed on the nominal p-values across all 19 cancer types for each test (unique combination of database origin and alteration direction, gain or loss). Enrichment tests are direction sensitive ("OG" tests for enrichment of oncogenes in Rank-1 cis genes in amplifications, "TN" tests for enrichment of tumor suppressors in Rank-1 cis genes in deletions, "COMBINED" tests for the union of the two sets).

We also tested for enrichment of amplification-driven gene dependencies in Rank-1 cis genes across the 19 cancer types (see Methods: Copy number-associated gene dependencies) (Table S3.4). Multiple hypothesis correction using the FDR procedure (Benjamini and Hochberg 1995) was used to adjust the significance values across multiple cancer types.

Evaluation of the Reproducibility of Cis Driver Gene Predictions in BRCA

To evaluate the consistency of Rank-1 driver gene predictions, we generated 100 bootstrapped resamples of the original TCGA breast cancer dataset using sampling with replacement with the number of samples equal to the size of the original dataset, derived the predicted Rank-1 cis genes across the 100 bootstrapped datasets, and compared these predictions against the original list of predicted Rank-1 cis genes from the full dataset. A reproducibility score was calculated for each of the original predicted Rank-1 cis gene as the percent of inclusion of the particular gene as a Rank-1 cis gene among the bootstrapped results. To explain the variation on reproducibility scores across genes, we modeled these scores using linear regression models with the dependent variable being either the Weighted Fraction of Trans Mediated (WFTM) or the Entropy of WTFM for the alteration of interest, calculated as the Shannon Entropy of WTFM of all differentially expressed cis genes within the alteration harboring the Rank-1 cis gene of interest. A two-sided t-test on the slope, β_1, of the linear regression, with H_a: $\beta_1 \neq 0$, was used to estimate significance, defined as p-value < 0.05.

Evaluation of Mediation Testing from Simulated Data

The Sobel test of mediation identifies cis genes that mediates SCNA and trans gene expression. To determine the conditions in which mediation is correctly identified, we used a forward simulation approach to generate labeled data of true positives and true negative, and then applied the mediation test to estimate its sensitivity and specificity.

True positive instances of mediation were generated using the following linear regression models:

$$Y_a = \alpha_1 + \beta_1 X_a + N(0, {\sigma_1}^2)$$

$$Z_a = \alpha_2 + \beta_2 Y_a + N(0, {\sigma_2}^2)$$

Here, X_a denotes the independent variable (SCNA status), Z_a is a dependent variable (trans gene expression) and Y_a is a true mediator of X_a and Z_a (cis gene expression).

True negative instances, representing the lack of a mediation effect, were generated using the following models:

$$Y_b = \alpha_1 + \beta_1 X_b + N(0, {\sigma_1}^2)$$

$$Z_b = \alpha_2 + \beta_2 X_b + N(0, {\sigma_2}^2)$$

Here, X_b is the independent variable (SCNA status) and Y_b and Z_b are both dependent variables generated based on separate regression models from X_b.

Variables X, Y, Z are vectors of length n, representing the sample size of the data. X is a binary vector (0s and 1s) corresponding to the binarized SCNA copy number status. σ is the standard deviation of the Gaussian noise term. We fixed β_1 and β_2 at 0.7 based on estimation from real data (TCGA breast cancer). The mediation test was performed on 1000 simulated true positives and 1000 simulated true negatives, and performance was measured in terms of AUC, sensitivity and specificity. For sensitivity and specificity, mediation calls were made based on the Sobel test p-value of 0.05. Test performance was recorded for simulated datasets based on a range of values of n (sample size) and σ (standard deviation of the Gaussian noise in the regression models). The standard deviation σ is a proxy for the correlation strength between dependent and independent variables, as higher noise corresponds to weaker correlation. For

interpretability purposes, values for the parameter σ are converted to the corresponding Pearson correlation estimates using a Loess model (Local Regression).

Software availability

iEDGE is available as an R package for download at https://github.com/montilab/iEDGE.

Data Access

The datasets analyzed in this study are available from the Broad Institute TCGA GDAC (http://gdac.broadinstitute.org/runs/) as described in Methods. iEDGE reports on these datasets are available in an interactive web portal (https://montilab.bu.edu/iEDGE) to allow for exploration and mining of results in user-friendly tabular and graphical formats.

3.3 Results

iEDGE identifies SCNA-associated cis and trans genes and pathway signatures in TCGA breast cancer

To identify cis and trans gene signatures of SCNAs in breast cancer, we performed integrative analysis on paired copy number and gene expression data from TCGA breast cancer primary tumors from four gene-expression based molecular subtypes (Luminal A, Luminal B, Her2 and Basal) using the workflow summarized in Figure 3.1.

Focal SCNAs were identified using GISTIC2.0 (Mermel et al., 2011), including 29 amplifications and 40 deletions. Next, we identified sets of cis and trans genes with significant differential expression with respect to each SCNA, pathway enrichments of

cis and trans gene sets (Figure 3.1A) and made predictions for cis gene drivers of each SCNA using mediation analysis (Figure 3.1B).

We identified a list of cis genes whose expression is significantly associated with each SCNA. This list comprises an average of 20 genes (and a median of 8 genes) per SCNA, totaling 1330 genes (269 in amplifications, 1061 in deletions) out of the original 2003 genes across all SCNAs identified using GISTIC2.0, a number clearly still too large for meaningful consideration for functional validation. The significance cutoff includes a fold change of $\log_2$ gene expression > 1.2 and FDR < 0.25 for the one-directional test of significance of differential expression with respect to the presence/absence of the alteration harboring the gene of interest (gene expression upregulation in amplifications and downregulation in deletions).

Similarly, we compiled a list of significantly differentially expressed trans genes using a fold change cutoff of $\log_2$ gene expression > 1.5 and bidirectional FDR < 0.1. This list contains an average of 865 (a median of 598) significant trans genes per SCNA.

Pathway enrichment analyses of the union of cis and trans genesets yielded interesting and potentially biological meaningful patterns (Figure 3.2). Several pathways were enriched across most SCNAs. For instance, gene sets HALLMARK_ESTROGEN_RESPONSE_LATE, HALLMARK_ESTROGEN_RESPONSE_EARLY, HALLMARK_G2M_CHECKPOINT were significant hits in more than 75% of SCNAs. These gene sets may indicate global patterns of downstream effects related to genomic instability induced by co-occurring SCNAs across tumor samples (Figure S3.1), since co-

occurring SCNAs will tend to have common pathway enrichments in the cis or trans genes. To elucidate breast cancer subtype specific pathway enrichments, we categorized each SCNA by their enrichment in each of four major breast cancer types: Luminal A, Luminal B, Her2, Basal. In addition, for each pathway, we tested if the enrichment across SCNAs tended to occur in subtype-specific SCNAs compared to non-subtype-specific SCNAs (Figure 3.2). Several significant pathways were found to be occurring more frequently in Basal-specific SCNAs, including HALLMARK_SPERMATOGENESIS, HALLMARK_KRAS_SIGNALING_UP, HALLMARK_E2F_TARGETS, HALLMARK_BILE_ACID_METABOLISM, HALLMARK_FATTY_ACID_METABOLISM, HALLMARK_UV_RESPONSE_UP, HALLMARK_MYOGENESIS (FDR < 0.05). The enrichment of KRAS signaling can be explained by published evidence supporting KRAS activation in basal-type breast cancer cells compared to luminal cells (Kim et al, 2015).

iEDGE identifies known cancer drivers in TCGA breast cancer

In addition to identifying cis and trans gene expressions of SCNAs, iEDGE can be used to predict cis gene drivers of each SCNA (Figure 3.1B). For each SCNA in the TCGA breast cancer dataset, we ranked the significant cis genes using the Sobel test of mediation (Sobel 1982) to predict the driver gene of the alteration. Cis genes were ranked by the weighted fraction of significant trans genes they mediate. Rank-1 cis genes, the predicted driver genes for each SCNA, are summarized in Table 3.1.

To verify the functional relevance across all Rank-1 driver gene predictions, we tested the list of Rank-1 cis genes for enrichment in cancer driver databases compared to

non-Rank-1 cis genes (Table 3.2). Significance of enrichment was determined using a one-sided Fisher exact test on the contingency table of counts of Rank-1 *vs.* non-Rank-1 cis genes against the list of known cancer drivers *vs.* unknown genes (P-value < 0.05). Predicted Rank-1 driver genes in amplifications were significantly enriched for known oncogenes in UNIPROT, COSMIC and the combined test (union of driver databases), and predicted Rank-1 driver genes in deletions were significantly enriched for tumor suppressors in TUSON, UNIPROT, COSMIC, and the combined test. Among the 65 Rank-1 cis genes (Table 3.1), known cancer drivers included *MCL1*, *ACTL6A*, *ZNF703*, *MYC*, *CCND1*, *FOXA1*, *ERBB2*, *CCNE1*, *ZNF2017* (oncogenes in amplifications) and *RPL5*, *ZMYND10*, *KMT2C*, *CSMD1*, *CDKN2B*, *PTEN*, *CREBL2*, *FANCA*, *MAP2K4*, and *ARHGAP35* (tumor suppressors in deletions) (Table 3.2).

iEDGE identifies amplification-driven gene dependencies in TCGA breast cancer

In addition to the prediction of known cancer drivers, we assessed whether iEDGE was capable of identifying genes with copy-number driven cancer dependencies, specifically, amplification-driven gene dependencies (see Methods). We identified genes with increased essentiality in an amplified state using DepMap data of genetic screens (RNAi screens) paired with copy number data (CCLE) (Table S8) and tested for their enrichment among iEDGE Rank-1 cis genes. In the TCGA breast cancer dataset, we found a highly significant enrichment of Rank-1 cis genes in genes with amplification-driven gene dependencies (Fisher test one-sided P-value: 3.53e-5). These were *ERBB2*, *CCNE1*, *CCND1*, *FOXA1*, *ANKRD17*, *MCL1*. Multiple literature sources suggest that all of these genes are linked to breast cancer development in the overexpressed state.

ERBB2, *CCND1*, *CCNE1* are well-characterized oncogenes present among our curated set of cancer driver databases. *FOXA1* has been shown to play an important role in promoting ER+ breast cancer (Meyer and Carroll 2012). *ANKRD17* is a cyclin E/Cdk2 substrate which positively regulates cell cycle progression by promoting G_1/S transition (Deng et al. 2009). *MCL1* high expression is linked to poor prognosis in triple-negative breast cancer and targeting of *MCL1* restricts the growth of triple negative breast cancer xenografts, suggesting its potential therapeutic value (Campbell et al. 2018).

In summary, using the cis gene mediation step, we identified known SCNA-associated breast cancer gene drivers and potentially novel genes with amplification-driven gene dependency that are of potential prognostic or therapeutic value.

TCGA pan-cancer analysis

Next, the driver prediction procedure was carried out across 19 cancer types from TCGA (Table S9). We tested for the enrichment of Rank-1 cis genes with known cancer drivers across the 19 cancer types and found significant enrichment (FDR < 0.05) in 15 out of 19 cancer types, which includes all tested cancer types with exception of COAD, ESCA, KIRC, THCA (Table 3.3).

Additionally, we tested for the enrichment of Rank-1 cis genes with genes manifesting amplification-driven dependencies and found significant enrichment (FDR < 0.05) in 8 out of the 19 cancer types analyzed, including BLCA, BRCA, CESC, ESCA, HNSC, LUAD, OV, UCEC (Table 3.4).

To assess the importance of the cis mediating step for the identification of known cancer drivers, we ordered Rank-1 cis genes by the number of cancer types they occur in

and tracked which of these genes were validated cancer drivers (Figure 3.3). The derived ordered Rank-1 gene list was then compared with the ordered list of recurrent top differentially expressed cis genes (top D.E.) in each SCNA, irrespective of their cis-mediating rank as well as with the ordered list of *all* recurrent differentially expressed cis genes in each SCNA. These comparisons confirmed that the recurrent Rank-1 cis genes were more likely to capture known drivers than both the recurrent top D.E. cis genes and all cis genes (Figure 3.3A). Remarkably, among the top 15 Rank-1 cis genes, 14 genes were known cancer drivers (*PTEN*, *WWOX*, *CCNE1*, *CDKN2A*, *MAP2K4*, *EGFR*, *KAT6A*, *MYC*, *KRAS*, *ERBB2*, *WHSC1L1*, *PARK2, CREBBP*, *RB1*) and only 1 was unverified (*ATP9B*) (Figure 3.3B, left) (Fisher's exact test p-value: 7.76e-08). In contrast, in the absence of the mediation step, among the top 15 top D.E. genes in SCNAs, 9 were confirmed drivers (Figure 3.3B, middle) (Fisher's exact test p-value: 0.0047), and among the top 15 cis genes in SCNAs, only 4 were confirmed drivers (Figure 3.3B, right) (Fisher's exact test p-value: 0.21). These results demonstrate the usefulness of the mediation test in further restricting the list of SCNA-associated cis genes to those of functional relevance across multiple cancer types. The mediation test provides a substantial improvement over ranking solely based on the cis gene differential expression, and demonstrates the added value of performing integrative analysis that models downstream biological effects as captured by trans gene expression.

In addition to enrichment for known cancer drivers, Rank-1 cis genes that frequently occur across cancer types (Figure 3.3) include putative or novel genes implicated in cancer initiation or progression. These include: *TRIP13*, a mitosis regulator

that was shown to promote tumor growth in colorectal cancer (Sheng et al, 2018) and is a predictor of poor prognosis in prostate cancer (Dong et al. 2019); *ORAOV1*, a gene overexpressed in many solid tumors that is linked to generation of reactive oxygen species (Zhai et al. 2014); *TPX2*, an interactor and substrate of Aurora-A that is a potent oncogene amplified in many cancers and a promising therapeutic target (Kufer et al. 2002; Yan et al, 2016); and *DUSP22*, which has been shown to behave as a tumor suppressor gene in peripheral T-cell lymphomas (Mélard et al, 2016) and regulates ERα dependent transcription in breast cancer cells (Sekine et al, 2007).

Evaluation of reproducibility of Rank-1 cis genes

To evaluate the reproducibility of cis gene ranking based on mediation testing, we quantified the consistency of Rank-1 cis gene predictions across bootstrapped resamples of the original TCGA breast cancer dataset (Figure 3.4). The majority of Rank-1 cis genes (69.2%) was consistently found as Rank-1 cis genes in bootstrapped datasets with greater than 0.75 fraction of inclusion, and the fraction of inclusion is not biased by SCNA type (amplification or deletion) (Figure 3.4A). We further revealed that the fraction of inclusion is positively associated with the Weighted Fraction of Trans Mediation (WFTM), the score used to rank cis genes within each SCNA (Figure 3.4B). In particular, while Rank-1 cis genes with lower fraction of inclusion across bootstrapped resamples tend to have lower WFTM, Rank-1 cis genes with higher fractions of inclusion show more variability in WFTM but generally tend to have higher WFTM. The fraction of inclusion is negatively associated with the entropy of WFTM of cis genes in SCNAs of interest (Figure 3.4C). This is an indication that SCNAs with a single dominant cis gene

mediating the majority of trans genes (lower entropy) tend to yield more reproducible Rank-1 cis genes across bootstrapped resamples than SCNAs with multiple cis genes with similar trans mediation (higher entropy).

Sensitivity and Specificity of mediation testing

Evaluation of the Sobel test of mediation was carried out using simulated data of true positives and true negative examples of mediation. Test performance, as measured using AUC, sensitivity and specificity, was recorded for varying combinations of correlation between the independent variable and mediator ("correlationXY") and sample size ("N") (Figure 3.5). High specificity (true negative rate) was consistently achieved for all input parameter ranges. Sensitivity (true positive rate) drops under conditions of low correlation and low sample size. Nevertheless, the conditions for lower sensitivity is not characteristic of real datasets that were used to test and validate iEDGE. Specifically, high correlation between SCNA and cis genes and between SCNA and trans genes is expected given that only cis and trans genes that were significantly expressed with respect to the SCNA were considered prior to mediation testing. Additionally, sufficient sample size is achieved in most of TCGA datasets tested (Table S9), with the only exception being the TCGA Adrenocortical carcinoma dataset (n = 77) in which mediation results should be interpreted with caution.

Graphical Portal of iEDGE Results Enable Targeted Queries

In order to enable fast interactive browsing of iEDGE precomputed runs on massive datasets such as the TCGA pancancer set, we developed a web portal (http://montilab.bu.edu/iEDGE/) to allow users to query iEDGE results selectively by

cancer types, genes, and SCNAs. This portal displays graphical and tabular results for each step of the iEDGE pipeline, including differential expression of cis and trans genes, mediation analysis for driver prediction, and pathway enrichment analysis.

An overview of an example walkthrough of a targeted query is illustrated in Figure S3.2. Here, the TCGA breast cancer (BRCA) report is selected and the gene query is *ERBB2* (*HER2*). The table of differential expression results is available for the cytoband *17q12* in which *ERBB2* resides in. Additionally, results of the mediation testing and driver gene prediction is available in a bipartite graph format. In this case, the graph indicates that *ERBB2* is the top mediating cis gene and predicted driver of the SCNA.

3.4 Discussion

Methods developed for the integrative analysis of (epi-)DNA regulators and gene expression data often focus only on the genes harbored by the alteration regions, while not considering the downstream (trans-)effects, which may limit a method's ability to detect cancer driver genes. We present a computational framework for the integrative analysis of (epi-)DNA and gene expression data for large-scale datasets, iEDGE, that is able to thoroughly catalogue the cis and trans effects of epi-(DNA) alterations, and to predict the most likely cis-driver genes based on the extent of their mediation of downstream trans-genes.

The first step of the iEDGE pipeline uses differential expression analysis to determine the cis and trans genes that are associated with the presence/absence of a particular epi-DNA alteration across samples. By measuring the alterations' association with trans genes we capture meaningful biological mechanisms representing downstream

effects that are generally missed by tools that only consider genes within the alterations (e.g., GISTIC2.0). Trans genes are of potential high relevance considering that (epi-)DNA regulators such as SCNAs harbor many upstream genes in signaling pathways, e.g., transcription factors, wherein the set of target genes effected can shed light on processes or pathways associated with disease progression.

The second step of the pipeline, the mediation analysis, ranks the set of cis genes by the extent of their mediation of trans genes. We showed that mediation analysis captures important cancer driver genes in our study of the TCGA breast cancer dataset. We then expanded these results by performing a pan-cancer analysis across 19 cancer types in the TCGA, further highlighting our tool's ability to identify known, as well as potentially novel drivers.

We conducted extensive in silico validation of predicted cis driver genes, by first testing for their enrichment with known drivers from multiple cancer driver databases. We then characterized predicted drivers by testing for their "essentiality" against genetic screens and cellular model data included in the DepMap, to explore SCNA-associated gene dependencies. Both analyses showed that our list of predicted genes is significantly enriched for cancer genes of functional relevance, either as cancer drivers, potential cancer therapeutic targets, or markers of disease progression. Further experimental studies are needed to validate and characterize predicted drivers.

Similar approaches have recognized the importance of integrating gene expression and the coordinated expression of affected gene modules for identifying drivers. One notable method is CONEXIC (Akavia et al., 2010), which identified

functional gene modules for each candidate regulator in the form of copy number alteration (CNA). The conceptual steps of iEDGE are similar to the "Single Modulator Step" outlined in Akavia et al, in which, first, cis genes are defined in their "candidate driver gene" selection process, and trans genes are defined in their "target gene modules" selection step, and second, a scoring function is used to find the single candidate driver gene that best associates with the target gene expression module. The implementation details of the second step for the scoring of driver genes is different compared to iEDGE. iEDGE considers each 3-layer relationship between SCNA status, cis, and trans gene expression to calculate a mediation effect and to rank cis genes, whereas the "Single Modulator Step" of CONEXIC computes the best candidate driver using a Normal Gamma scoring function to measure each target gene's fit with each candidate driver's gene expression. In addition to the "Single Modulator Step", CONEXIC uses an iterative procedure modeled after the module network method (Lee et al. 2006; Segal et al. 2003) to improve the scores of each module and their regulatory programs. This is not a feature of iEDGE but can be ran as an independent post-processing step to further refine the list of candidate genes. On the other hand, while CONEXIC is not available for public use, iEDGE is available as an open-source R package to enable analysis of custom datasets, and a graphical web portal is available for exploration and querying of precomputed runs on the TCGA datasets.

Other model-based methods incorporate models of causal relationships between multiple levels of genomic data through the scores of conditional dependencies, measured using partial correlation coefficients for normal continuous features (Amgalan and Lee

2015), or conditional mutual inclusive information for binary or mixed binary and continuous features (Kim et al. 2016; Zhang et al. 2014). These approaches are similar to the mediation testing step of iEDGE, but they are more conservative models that detect full mediation, e.g., significant hits are instances in which the conditional independence given the mediator is zero, a condition that is rarely satisfied by genomic data, whereas mediation tests are also able to capture partial mediation in which the association between the independent and response variable is significantly reduced in size when the mediator is introduced but may still be different from zero.

One assumption used in the iEDGE analysis presented in this study is that the number of trans genes that a cis gene mediates is a proxy measure of a gene's importance (i.e., of its likelihood of being a cancer driver). This a simple and intuitive heuristic to estimate the extent of transcriptional impact from each (epi-)DNA regulator, albeit it may not be an appropriate assumption for specific use cases. For more targeted analysis, one may be only interested in predicting the cis gene that mediates gene targets in a particular pathway. Our tool is customizable in that the user can specify the set of trans genes to consider for mediation based on their membership in a pathway of interest or an experimentally derived gene signature.

We use SCNA as an example of epi-DNA events to demonstrate a convenient use case for this tool as SCNAs can be used to easily define genomic boundaries for distinguishing cis and trans acting genes. However, this tool can also be applicable to the integrative analysis of other genomic/epi-genomic data types such as DNA methylation, DNA mutations, and microRNA regulatory networks. Similarly, gene expression dataset can

use a variety of quantitative gene-centric measures such as RNA-seq, microarray, or proteomic assays.

Figure 3.1

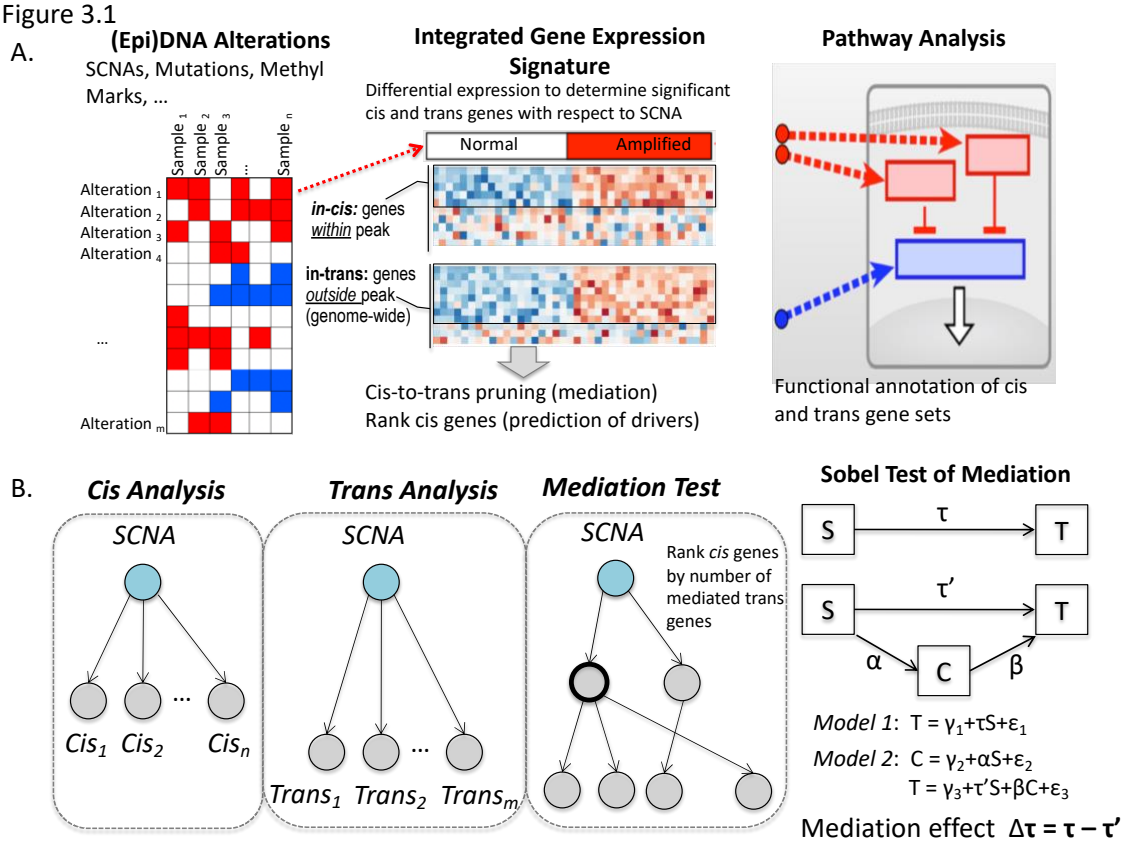


***Figure 3.1* Overview of iEDGE workflow**

A. The main workflow of iEDGE starts with identifying the cis and trans gene expression signatures of epi-DNA alterations, in this case Somatic Copy Number Alterations (SCNA) through differential expression analysis. Next, we perform pathway enrichment analysis to identify pathways or genesets associated with each SCNA.
B. The Cis-to-Trans gene mediation analysis using the Sobel test is an optional module which identifies putative driver cis genes of each epi-DNA alteration. Here, given the alteration of interest, the list of differentially expressed cis genes and trans genes, we use the mediation test to determine the cis gene which mediates the most trans gene expression.

Figure 3.2

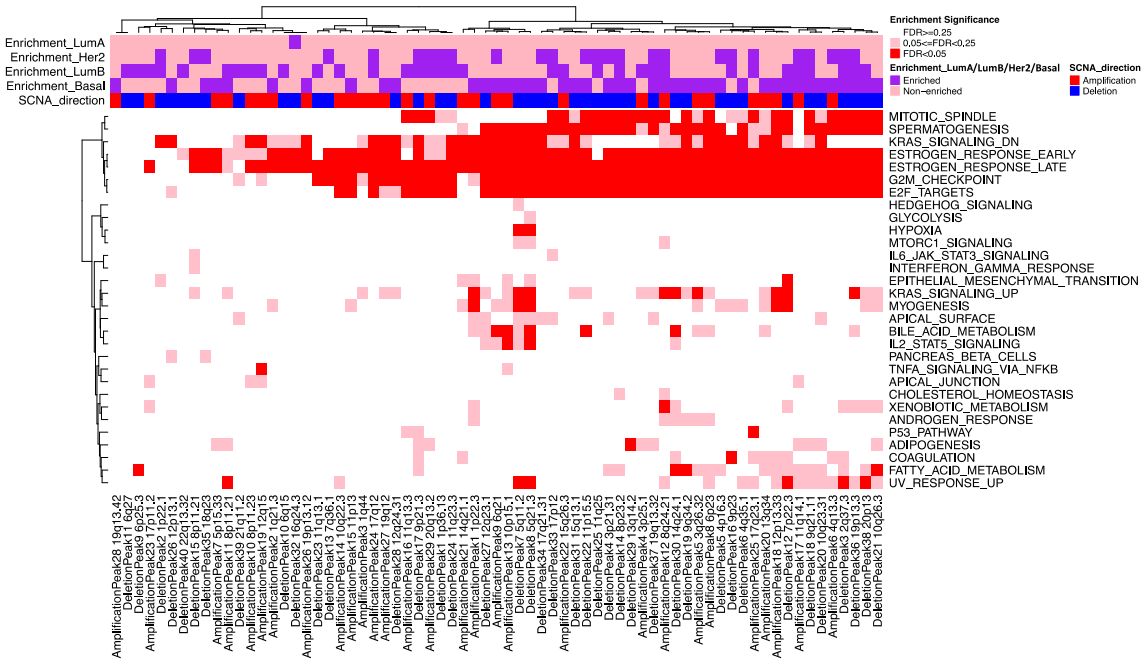


Figure 3.2 **Pathway enrichments of cis and trans gene signatures of TCGA breast cancer somatic copy number alterations**

Heatmap of gene signature in enrichments for the union of cis and trans differentially expressed gene sets with respect to each SCNA for the TCGA breast cancer dataset. Column color labels indicate the enrichment of each SCNA for a particular breast cancer subtype determined by Fisher's exact test of counts for within-subtype vs. outside-subtype counts of SCNA occurrence.

Figure 3.3

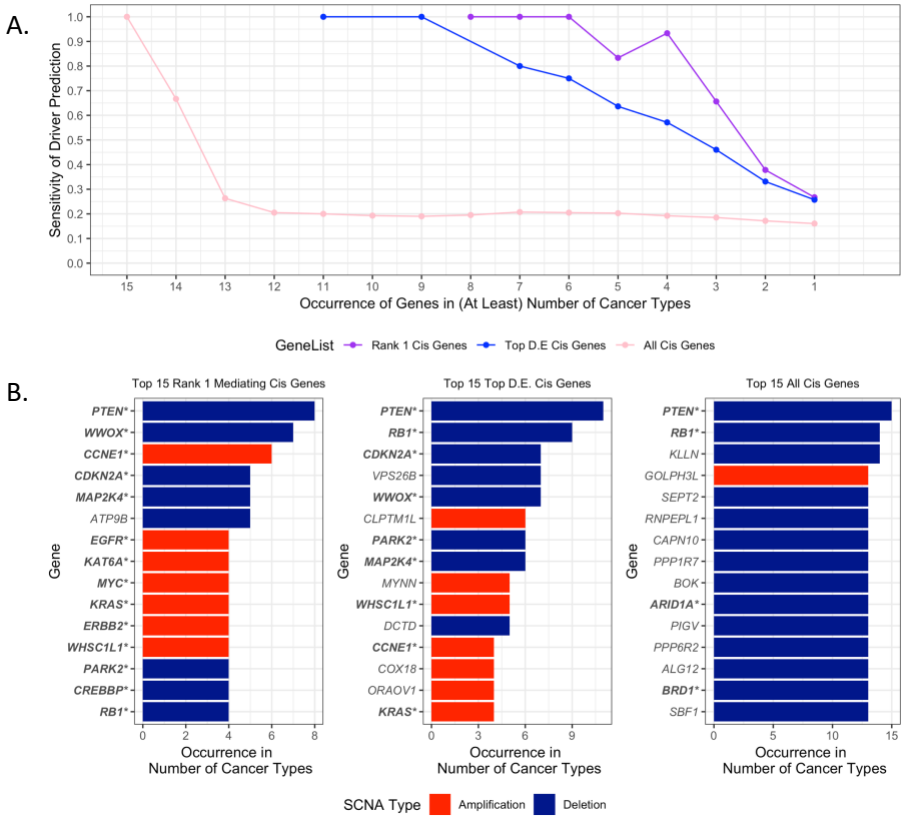


Figure 3.3 **Sensitivity of cancer driver predictions across multiple cancer types**

A. Sensitivity of Driver Predictions vs. Occurrence of Genes in Number of Cancer Types
B. Barplot of top 15 genes ranked by occurrence in number of cancer types. From the left: Rank 1 Mediating Cis Genes, middle: Top differentially expressed (D.E.) Cis Genes, right: all cis genes differentially expressed in alteration. Known cancer drivers are marked with (*) in bold font.

Figure 3.4

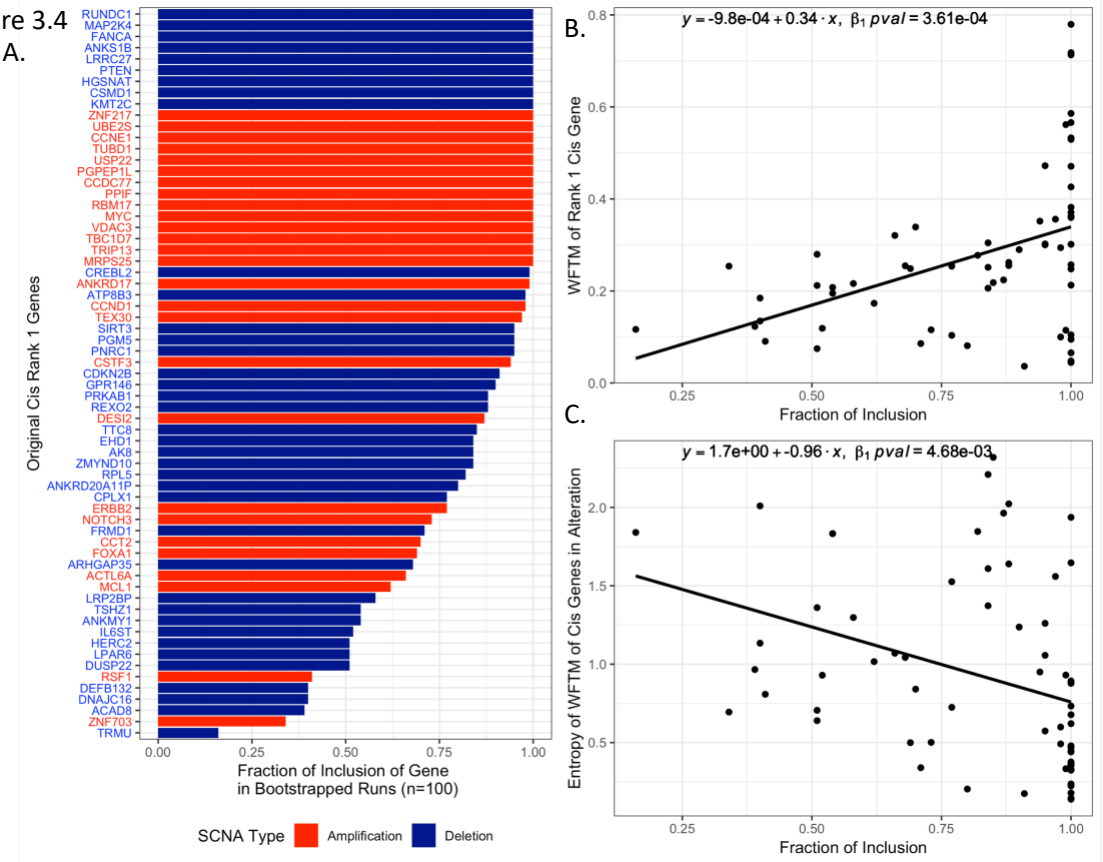


Figure 3.4 **Reproducibility of Rank 1 cis genes among bootstrapped resamples**

A. Inclusion of Cis Rank 1 Genes in Bootstrapped Resamples
B. Fraction of inclusion vs. Weighted Fraction of Trans Mediation (WFTM)
C. Fraction of inclusion vs. entropy of WFTM of cis genes in alteration

Figure 3.5

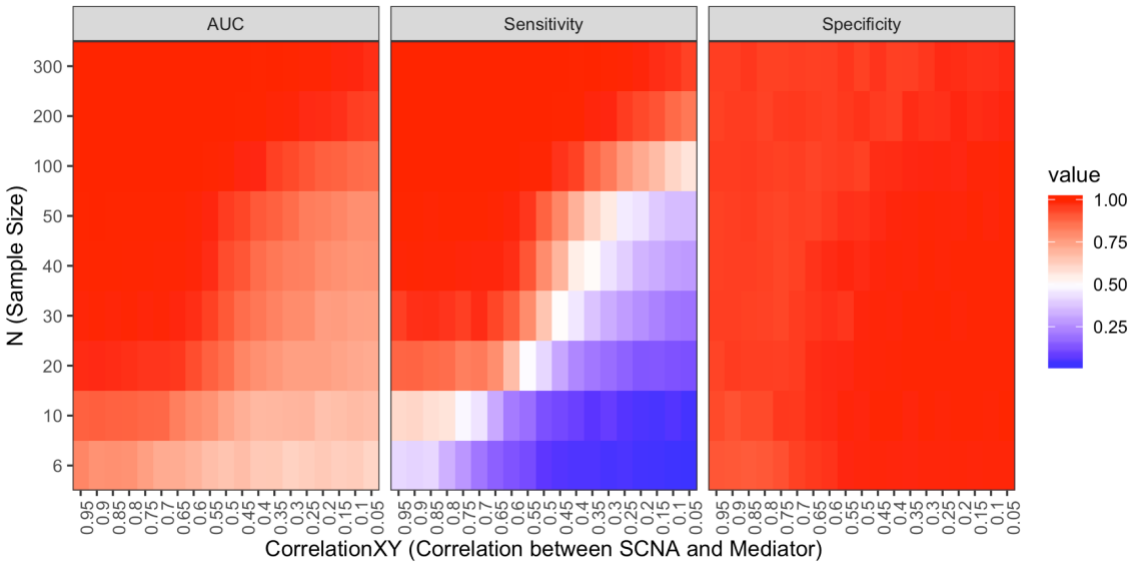


Figure 3.5 **Mediation test performance on simulated data**

AUC (Area Under the ROC Curve), sensitivity (true positive rate), specificity (true negative rate) for various values of correlationXY (correlation between SCNA and mediator) and N (sample size).

Table 3.1 Differential Expression of Rank-1 Cis Genes in Somatic Copy Number Alterations (SCNA) in TCGA Breast Cancer

Alteration ID	Cis Gene	Rank	Trans Mediated	Total Trans Genes	FTM (Fraction of Trans Mediated)	WFTM (Weighted Fraction of Trans Mediated)
AmplificationPeak2	*MCL1*	1	63.23	365	0.35	0.17
AmplificationPeak3	*DESI2*	1	44.80	200	0.26	0.22
AmplificationPeak4	*MRPS25*	1	126.20	491	0.52	0.26
AmplificationPeak5	*ACTL6A*	1	503.21	1569	0.39	0.32

AmplificationPeak6	*ANKRD17*	1	69.13	603	0.31	0.11
AmplificationPeak7	*TRIP13*	1	146.48	204	0.84	0.72
AmplificationPeak8	*TBC1D7*	1	708.94	1332	0.69	0.53
AmplificationPeak10	*ZNF703*	1	65.78	259	0.41	0.25
AmplificationPeak11	*VDAC3*	1	186.36	318	0.80	0.59
AmplificationPeak12	*MYC*	1	163.50	1721	0.60	0.10
AmplificationPeak13	*RBM17*	1	1320.23	2332	0.66	0.57
AmplificationPeak14	*PPIF*	1	168.14	464	0.74	0.36
AmplificationPeak15	*CSTF3*	1	130.17	370	0.55	0.35
AmplificationPeak16	*CCND1*	1	65.00	221	0.45	0.29
AmplificationPeak17	*RSF1*	1	29.16	322	0.17	0.09
AmplificationPeak18	*CCDC77*	1	1077.01	1510	0.94	0.71
AmplificationPeak19	*CCT2*	1	153.60	453	0.49	0.34
AmplificationPeak20	*TEX30*	1	526.90	1480	0.40	0.36
AmplificationPeak21	*FOXA1*	1	109.67	441	0.67	0.25
AmplificationPeak22	*PGPEP1L*	1	327.68	769	0.95	0.43
AmplificationPeak23	*USP22*	1	5.26	118	0.11	0.04
AmplificationPeak24	*ERBB2*	1	61.96	598	0.16	0.10
AmplificationPeak25	*TUBD1*	1	146.81	690	0.57	0.21
AmplificationPeak26	*NOTCH3*	1	74.81	647	0.44	0.12
AmplificationPeak27	*CCNE1*	1	951.15	1220	0.95	0.78
AmplificationPeak28	*UBE2S*	1	51.36	97	0.79	0.53
AmplificationPeak29	*ZNF217*	1	38.31	584	0.28	0.07
DeletionPeak1	*DNAJC16*	1	57.11	424	0.18	0.13
DeletionPeak2	*RPL5*	1	63.84	230	0.35	0.28
DeletionPeak3	*ANKMY1*	1	170.45	820	0.26	0.21
DeletionPeak4	*ZMYND10*	1	298.70	1188	0.29	0.25
DeletionPeak5	*CPLX1*	1	342.50	1349	0.40	0.25
DeletionPeak6	*LRP2BP*	1	298.96	1383	0.43	0.22
DeletionPeak7	*IL6ST*	1	386.42	3251	0.32	0.12
DeletionPeak9	*DUSP22*	1	23.95	113	0.35	0.21
DeletionPeak10	*PNRC1*	1	151.14	320	0.52	0.47
DeletionPeak11	*FRMD1*	1	8.40	98	0.37	0.09
DeletionPeak12	*GPR146*	1	385.03	1328	0.44	0.29
DeletionPeak13	*KMT2C*	1	42.38	405	0.36	0.10
DeletionPeak14	*CSMD1*	1	29.16	609	0.56	0.05

DeletionPeak15	*HGSNAT*	1	93.70	378	0.55	0.25
DeletionPeak17	*CDKN2B*	1	19.64	539	0.20	0.04
DeletionPeak18	*PGM5*	1	164.75	545	0.77	0.30
DeletionPeak19	*AK8*	1	400.58	1315	0.40	0.30
DeletionPeak20	*PTEN*	1	268.89	892	0.70	0.30
DeletionPeak21	*LRRC27*	1	631.45	1185	0.68	0.53
DeletionPeak22	*SIRT3*	1	326.73	1088	0.36	0.30
DeletionPeak23	*EHD1*	1	32.56	158	0.26	0.21
DeletionPeak24	*REXO2*	1	76.59	300	0.29	0.26
DeletionPeak25	*ACAD8*	1	37.50	305	0.22	0.12
DeletionPeak26	*CREBL2*	1	132.00	235	0.66	0.56
DeletionPeak27	*ANKS1B*	1	635.40	1766	0.95	0.36
DeletionPeak28	*PRKAB1*	1	410.45	1566	0.34	0.26
DeletionPeak29	*LPAR6*	1	141.24	505	0.60	0.28
DeletionPeak30	*TTC8*	1	366.13	1679	0.24	0.22
DeletionPeak31	*HERC2*	1	100.02	1339	0.17	0.07
DeletionPeak32	*FANCA*	1	436.34	1177	0.46	0.37
DeletionPeak33	*MAP2K4*	1	256.71	545	0.80	0.47
DeletionPeak34	*RUNDC1*	1	472.54	1237	0.39	0.38
DeletionPeak35	*TSHZ1*	1	81.32	417	0.22	0.20
DeletionPeak36	*ATP8B3*	1	137.85	1379	0.53	0.10
DeletionPeak37	*ARHGAP 35*	1	272.28	1069	0.36	0.25
DeletionPeak38	*DEFB132*	1	256.26	1388	0.48	0.18
DeletionPeak39	*ANKRD20 A11P*	1	10.46	129	0.51	0.08
DeletionPeak40	*TRMU*	1	10.14	87	0.17	0.12

Table 3.2 Enrichment of Rank 1 Cis Genes in Cancer Driver Databases in TCGA breast cancer

Test_Type	Database	Count_Rank1_Driver	Count_Rank1_Unknown	Count_NonRank1_Driver	Count_NonRank1_Unknown	Count_Total	Fisher_P.Value_OneSided	Rank1_DriverGene_List
Oncogene_in_cis_Amplification	TUSON_OG	2	25	2	240	269	5.14E-02	*MYC, ERBB2*
Oncogene_in_cis_Amplification	OMIM_OG	1	26	1	241	269	1.91E-01	*MCL1*
Oncogene_in_cis_Amplification	UNIPROT_OG	9	18	23	219	269	1.61E-03	*MCL1, ACTL6A, ZNF703, MYC, CCND1, FOXA1, ERBB2, CCNE1, ZNF217*
Oncogene_in_cis_Amplification	COSMIC_OG	5	22	5	237	269	1.25E-03	*MYC, CCND1, FOXA1, ERBB2, CCNE1*
Oncogene_in_cis_Amplification	ANY_OG	9	18	27	215	269	4.10E-03	*MCL1, ACTL6A, ZNF703, MYC, CCND1, FOXA1, ERBB2, CCNE1, ZNF217*

TumorSuppressor_in_cis_Deletions	TUSON_TN	5	33	27	996	1061	4.47E-03	*RPL5, KMT2C, PTEN, MAP2K4, ARHGAP35*
TumorSuppressor_in_cis_Deletions	OMIM_TN	1	37	30	993	1061	6.82E-01	*RPL5*
TumorSuppressor_in_cis_Deletions	UNIPROT_TN	7	31	70	953	1061	1.62E-02	*ZMYND10, CSMD1, CDKN2B, PTEN, CREBL2, MAP2K4, ARHGAP35*
TumorSuppressor_in_cis_Deletions	COSMIC_TN	5	33	22	1001	1061	2.04E-03	*RPL5, KMT2C, PTEN, FANCA, MAP2K4*
TumorSuppressor_in_cis_Deletions	ANY_TN	10	28	121	902	1061	1.31E-02	*RPL5, ZMYND10, KMT2C, CSMD1, CDKN2B, PTEN, CREBL2*

								, FANCA, MAP2K4, ARHGAP35

Table 3.3 Enrichment of known cancer drivers among rank 1 cis genes in TCGA pancancer analysis (19 cancer types)

Cancer Type	ANY_OG	ANY_TN	ANY_COMBINED	ANY_OG_RANK1_CIS_IN_AMP	ANY_TN_RANK1_CIS_IN_DEL
ACC	6.77E-01	1.14E-02	2.97E-02	*FAM32A*	*PARK2, MKRN1, GABARAPL2*
BLCA	7.58E-02	6.63E-02	4.35E-03	*TRIT1, FGFR3, MYB, EGFR, KAT6A, MYC, GATA3, KRAS, ERBB2, CCNE1*	*ARID1A, FHIT, WTAP, CDKN2A, PTEN, ZFP36L1, CREBBP*
BRCA	6.18E-02	2.16E-02	5.76E-04	*MCL1, ACTL6A, ZNF703, MYC, CCND1, FOXA1, ERBB2, CCNE1, ZNF217*	*RPL5, ZMYND10, KMT2C, CSMD1, CDKN2B, PTEN, CREBL2, FANCA, MAP2K4, ARHGAP35*
CESC	2.83E-01	1.14E-02	6.07E-03	*EGFR, FADD, YAP1, ERBB2, FZR1*	*FHIT, EEF1E1, PTEN, WWOX, MAP2K4, ELAC1, STK11, BTG3, KDM6A*
COAD	4.99E-01	1.40E-01	9.54E-02	*WHSC1L1, KLF5, GRB7, PDRG1*	*REG4, MTAP, B2M, SMAD3, MAP2K4, SMAD4, BTG3*
ESCA	2.83E-01	3.04E-01	1.13E-01	*TRIT1, RIT1, EGFR, KAT6A, CCNE1*	*CDKN2A, LARP4B, ELAC1, BTG3*
GBM	6.18E-02	1.14E-02	7.84E-05	*MDM4, PDGFRA, MYC, CDK4, MDM2, AKT1*	*DFFB, PARK7, DMRTA2, MED23, QKI, MTAP, EXT2, RB1*
HNSC	1.55E-01	1.15E-01	1.13E-02	*NFE2L2, EGFR, WHSC1L1, SNAI2, PDCD1LG2, NFIB, IGF1R, CBFA2T2*	*VAMP3, NSD1, TFAP2A, CDKN2A, NOTCH1, LARP4B, PTEN, ZNF750*

KIRC	1.00E+00	8.48E-01	9.52E-01		*SKI*
KIRP	2.58E-01	1.14E-02	4.29E-03	*SQSTM1, TES*	*CDHR2, MTAP, TBRG1, NDRG2, CDK10, NF2*
LIHC	4.99E-01	1.14E-02	9.03E-03	*TERT, NOL7, ST7*	*TPRG1L, NRG1, PTEN, DLG2, SDHD, RB1, WWOX, NCOR1*
LUAD	7.58E-02	8.28E-02	6.07E-03	*RLF, EIF5A2, TERT, MET, KRAS, CDK4, MDM2, CCNE1*	*RAP1A, XPC, SESN1, PARK2, CDKN2A, LARP4B, MOAP1, WWOX, SMARCA4*
LUSC	2.39E-01	6.36E-03	9.61E-04	*BCL11A, NFE2L2, SOX2, WHSC1L1, ARHGEF39, URI1, CRKL*	*FOXP1, PARK2, PTPRD, CDKN2A, ZMYND11, PTEN, RB1, B2M, CREBBP, WWOX, NF1, RNF126*
OV	1.43E-01	2.12E-05	5.12E-06	*PPIE, WHSC1, RNF144B, KAT6A, KRAS, SIVA1, BRD4, CCNE1*	*ARID1A, ING5, UHRF2, ZMYND11, PTEN, BTRC, EI24, RB1, UBE3A, CREBBP, WWOX, MAP2K4, NF1*
PAAD	4.99E-01	3.74E-04	5.76E-04	*ACTL6A, FAM60A, MIEN1*	*RAD17, SESN1, CREBL2, CHD8, NDNL2, MAP2K4, TRIM37, ELAC1, CERK*
PRAD	4.94E-01	1.14E-02	1.59E-02	*SET, GRAP, MALT1, AURKA*	*JAK1, PRDM5, PIK3R1, ZNF292, PTEN, CDKN1B, SETD8, ZFHX3, CIC*
READ	1.82E-01	1.55E-01	2.66E-02	*YY1AP1, WHSC1L1, PAN3, PRR14, ERBB2*	*SDHB, YAP1, WWOX, SMAD4*
THCA	1.00E+00	5.38E-01	7.13E-01		*CCDC6, INTS6*
UCEC	7.11E-02	3.74E-04	5.12E-06	*SETDB1, TACC3, NEDD9, FGFR1, KAT6A, MYC, KRAS,*	*PRDM2, SFN, FHIT, PARK2, FAM120A, DLG2, TIRAP,*

				ERBB3, GRB7, CCNE1, CDC25B	*CREBBP, ZFHX3, WWOX, NF1, ZNRF3*

Table 3.4 Enrichment of Amplification-driven gene dependencies among rank 1 cis genes in amplifications in TCGA pancancer analysis (19 cancer types)

Cancer Type	P.Value (One-sided Fisher's Exact Test)	FDR (One-sided Fisher's Exact Test)	List of Genes Rank 1 with Amplification-Driven Gene Dependency
ACC	1	1	
BLCA	2.6436E-07	5.0228E-06	*EGFR, ERBB2, GATA3, KRAS, MYB, KAT6A, CCNE1*
BRCA	3.5317E-05	0.00033552	*ERBB2, ANKRD17, FOXA1, MCL1, CCND1, CCNE1*
CESC	0.00994043	0.02698116	*EGFR, ERBB2, BCL2L1*
COAD	0.56624663	0.89655717	*KLF5*
ESCA	0.0012639	0.00480281	*EGFR, ANKRD17, KAT6A, CCNE1*
GBM	0.03484641	0.07356464	*CDK4, MDM2, MDM4*
HNSC	0.00931701	0.02698116	*EGFR, IGF1R, NFIB, PPFIA1*
KIRC	1	1	
KIRP	1	1	
LIHC	1	1	
LUAD	0.00046296	0.00293207	*CDK4, KRAS, MDM2, MET, CCNE1*
LUSC	0.53780376	0.89655717	*PPFIA1*
OV	0.0127933	0.03038408	*KRAS, KAT6A, CCNE1*
PAAD	1	1	
PRAD	1	1	
READ	0.22838937	0.4339398	*ERBB2*
THCA	1	1	
UCEC	0.00098788	0.00469242	*KRAS, BCL2L1, KAT6A, IRS2, CCNE1*

Chapter 4: Conclusions and Future Directions

4.1 Summary of Thesis Aims

In this dissertation, I presented two aims that, through distinct data analysis techniques and methodological developments, addressed knowledge gaps in our understanding of the cancer genome landscape. The first aim, using gene expression datasets of chemical perturbations, is targeted toward the goal of cancer prevention. The second, using human patient cancer genomics data, focuses on the discovery of cancer driver genes and potential therapeutic agents. These two aims can be considered two complementary contributions towards the construction of the overarching "carcinogenome". While these projects utilize datasets of distinct origins, the first using perturbational gene expression profiles and, the second using multi-level genomic profiling of real human tumors, they are both necessary for the overall understanding of the cancer genome landscape from the perspective of the ultimate goals of cancer prevention and therapy.

From the point of view of cancer prevention, identification of environment agents with carcinogenic effects and understanding the mechanism of action of such agents is critical. While a handful carcinogenic chemicals have been successfully identified, less than 2% of total chemicals in consumer or industrial use have been rigorously tested for carcinogenicity. In this thesis, I addressed the knowledge gap of chemical carcinogenicity through a in-vitro gene expression profiling effort that profiles a large sample size of chemicals relevant for liver carcinogenicity (Chapter 2.1). Secondly, I contributed to methodological developments for analyzing large-scale gene expression profiles of

chemical perturbations through the work in network-based analysis of perturbational profiles (Chapter 2.2).

From the perspective of cancer therapy, I developed the tool iEDGE, an integrative approach that predicts cancer driver genes associated with (epi-)genetic marks such as somatic copy number alterations (SCNAs) (Chapter 2). This tool has proven to be instrumental for identifying known and novel cancer drivers and therapeutic markers across multiple types of cancer.

4.2 Contributions

4.2.1: Building liver carcinogenicity and genotoxicity models from in-vitro high-throughput transcriptomic assays (Chapter 2.1)

- I was a key contributor for the experimental design in all phases of the project, including chemical selection and procurement.
- I developed the computational pipeline for data analysis in the project. Computational modules include, but are not limited to, differential expression analysis, pathway enrichment analysis, scripts for comparison of our generated data to external knowledge bases such as Toxcast, Drugmatrix, and Cmap data.
- I developed the web portal, implemented using R Shiny, for user friendly query of results from the study (carcinogenome.org). Additionally, I set up and maintained the server for the web portal.
- Manuscript under review (Li et al. 2018). BioRxiv preprint available: https://www.biorxiv.org/content/early/2018/05/16/323964.

4.2.2 Network-based analysis of transcriptional profiles from chemical perturbations (Chapter 2.2)

- I provided scripts for interpretation of results and generated key figures for the manuscript.
- I aided in discussions during manuscript preparation, particularly in the interpretation of results.
- Manuscript published: Mulas F, Li A, Sherr DH, Monti S. 2017. Network-based analysis of transcriptional profiles from chemical perturbations experiments. BMC Bioinformatics 18: 130.

4.2.3 Towards cancer therapy - Molecular characterization of the cancer genome and epi-genome using integrative analysis (Chapter 3)

- I developed the R package for iEDGE (https://github.com/montilab/iEDGE).
- I tested and evaluated iEDGE on several TCGA datasets.
- I developed a web portal for querying iEDGE results on TCGA datasets (http://montilab.bu.edu/iEDGE/).
- Manuscript in preparation (Li et al. 2018).
- Previous published work based on preliminary versions of iEDGE:
    - Chapuy B, Stewart C, Dunford AJ, Kim J, Kamburov A, Redd RA, Lawrence MS, Roemer MGM, Li AJ, Ziepert M, et al. 2018. Molecular

subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat Med* **24**: 679–690.

4.3 Accomplishments and Future Directions

4.3.1: Building liver carcinogenicity and genotoxicity models from in-vitro high-throughput transcriptomic assays (Chapter 2.1)

In this study, I detailed our study aimed at short-term in-vitro high-throughput gene expression profiling of more than 300 liver carcinogens and non-carcinogens. We had to overcome many challenges due to the ambitious nature of the study which relies on the hypothesis that in-vitro short-term assays can predict and characterize in-vivo long-term carcinogenicity response.

A major accomplishment from this study is that we were able to achieve accurate classifiers with 72.2% AUC for prediction of carcinogenicity and 82.3% for prediction of genotoxicity. Additionally, we captured relevant MoAs of carcinogenicity such as upregulation of immune response, cell death, DNA repair and transcriptional regulation and downregulation of metabolism related pathways, and cell-cell organization and communication.

We showed that Transcriptional Activity Score (TAS) of the gene expression profiles, while not predictive of carcinogenicity, has a high impact on classifier performance. It is also important for capturing relevant MoAs of carcinogenicity as shown by the pathway enrichment analysis. After restricting data analysis to profiles of high TAS, we found significant improvement of classifier performance as well as increased enrichment of known MoAs relevant to carcinogenesis. We have also verified

the relevance of our findings through comparison with external resources such as Drugmatrix, Tox21, and CMap.

An important factor for achieving high TAS is dose. Although we used standardized doses across chemicals for this experiment, we offered recommendations for dose schemes for future experiments based on observations from this study. For instance, due to several chemicals not receiving adequate TAS, we recommended MTT assays for dose determination moving forward. We also showed in-vitro to in-vivo dose extrapolation procedures can be useful for dose determination for in-vitro assays in case where in-vivo dosing and carcinogenicity data is available.

This experiment is the largest study of high-throughput transcriptional profiling of liver carcinogens to our knowledge. The results from this study was instrumental for assessing the viability of in-vitro short-term screening for long-term carcinogenicity prediction. Its success set the stage for in-vitro screening of other types of carcinogens such as breast carcinogens. For example, we initiated the in-vitro screening of mammary gland carcinogens using MCF10A and p53 deficient MCF10A. More generally, the experimental and analysis pipeline used in this project, paired with our suggestions for dosing improvements, paves the way for screening of large chemical panels for other organ, disease, and adverse outcome contexts.

4.3.2 Network-based analysis of transcriptional profiles from chemical perturbations (Chapter 2.2)

In this study, we presented a computational pipeline for transcriptional network inference and comparison for analysis of chemical perturbations from high-throughput

transcriptional profiling experiments. This pipeline was applied to the analysis of gene expression profiles from rat-based chemical exposure experiments, with a focus on grouping and characterization of chemicals with known labels of carcinogenicity and genotoxicity.

We showed that groups of chemicals with similar pharmacological functions and carcinogenicity/genotoxicity labels can be identified using our pipeline. Importantly, gene modules with altered connectivity were enriched for pathways related to the chemicals' known mechanisms of action. These findings demonstrate the advantage of this network-based approach, compared to traditional differential expression analysis, in characterizing modes of actions of chemical perturbations from transcriptional profiles.

4.3.3 Towards cancer therapy - Molecular characterization of the cancer genome and epi-genome using integrative analysis (Chapter 3)

In this chapter, I presented computational tool for integrative analysis of (epi-) DNA and gene expression for large-scale genomic datasets, iEDGE. I applied this tool towards characterization of cis and trans gene expression effects of somatic copy number alterations (SCNAs) in several cancer genomic datasets and predicted cis gene drivers of each SCNA based on the extent of their mediation of downstream trans genes.

We showed that inclusion of trans genes in the integrative analysis led to more meaningful pathway enrichments that highlighted processes in subtype-specific breast cancers. Through iEDGE analysis of 19 cancer types, we showed that the mediation testing step leads to greater sensitivity in capturing known cancer driver genes compared to analysis based on differential expression of cis genes alone. In addition to testing for

enrichment of iEDGE identified drivers for known cancer drivers, we characterized predicted drivers by testing for enrichment of amplification driven gene dependencies using data from DepMap which combined gene essentiality data from genetic screens with genomic characterization of cell lines. Both analyses showed that our list of predicted genes is significantly enriched for cancer genes of functional relevance either as cancer drivers or potential prognostic markers or therapeutic targets.

While we focused on application of iEDGE for SCNA and gene expression data, this tool is applicable to a broader set of (epi-) genetic alterations including DNA methylation, DNA mutation, and microRNA regulatory networks. In general, any type of (epi-) genetic data is applicable as long as cis and trans effects can be defined. More studies are needed to evaluate iEDGE in these other contexts.

Since we used iEDGE for unbiased exploratory analysis of large-scale cancer genomic datasets, we adopted a simple heuristic for ranking of cis genes for driver prioritization, which considers only the weighted fraction of trans genes mediated. For more specific use cases, more sophisticated procedures for cis gene prioritization can be used. For instance, one may consider using weights to specify a set of a-priori defined trans genes that should be prioritized in the mediation analysis from gene sets derived from external studies.

.

APPENDIX

Table S1 List of chemicals used in HEPG2 in vitro gene expression profiling.

BUID	Chemical Name	CAS	Carcinogenicity	Genotoxicity	TAS (mean)
BUID_001193	o-Phenylenediamine.2HCl	615-28-1	+	+	0.609
BUID_001308	p-Rosaniline.HCl	569-61-9	+	+	0.55
BUID_000888	2-Methyl-1-nitroanthraquinone	129-15-7	+	+	0.534
BUID_000070	o-Aminoazotoluene	97-56-3	+	+	0.52
BUID_001483	Trp-P-1 acetate	75104-43-7	+	+	0.48
BUID_000879	3'-Methyl-4-dimethylaminoazobenzene	55-80-1	+	+	0.478
BUID_000939	Michler's ketone	90-94-8	+	+	0.433
BUID_001188	1-Phenylazo-2-naphthol	842-07-9	+	+	0.429
BUID_000117	Auramine-O	2465-27-2	+	+	0.408
BUID_000248	Captafol	(2425-06-1)	+	-	0.383
BUID_001148	Oxymetholone	434-07-1	+	-	0.362
BUID_001379	2,3,7,8-Tetrachlorodibenzo-p-dioxin	1746-01-6	+	-	0.278
BUID_000808	Kepone	143-50-0	+	-	0.266
BUID_000605	Ethinyl estradiol	57-63-6	+	-	0.255
BUID_001139	N-(9-Oxo-2-fluorenyl)acetamide	3096-50-2	+	N/A	0.254
BUID_001160	2,3,4,5,6-Pentachlorophenol, technical grade	87-86-5	+	-	0.243
BUID_001077	p-Nitrosodiphenylamine	156-10-5	+	+	0.242
BUID_001366	Tamoxifen citrate	54965-24-1	+	N/A	0.24
BUID_000284	3-Chloro-4-(dichloromethyl)-5-hydroxy-2(5H)-furanone	77439-76-0	+	N/A	0.232
BUID_000058	1-Amino-2-methylanthraquinone	82-28-0	+	+	0.225
BUID_000906	4,4'-Methylene-bis(2-chloroaniline)	101-14-4	+	+	0.216
BUID_000283	4-Chloro-4'-aminodiphenylether	101-79-1	+	N/A	0.206
BUID_000307	Chlorobenzene	108-90-7	+	-	0.2
BUID_001075	N-Nitrosodimethylamine	62-75-9	+	+	0.191
BUID_000363	p-Cresidine	120-71-8	+	+	0.19
BUID_000485	Diethylstilbestrol	56-53-1	+	-	0.188
BUID_000206	Budesonide	51333-22-3	+	N/A	0.186
BUID_000055	3-Amino-9-ethylcarbazole.HCl	6109-97-3	+	+	0.18
BUID_000513	N,N-Dimethyl-4-aminoazobenzene	60-11-7	+	+	0.177
BUID_001121	Norlestrin	8015-12-1	+	N/A	0.166
BUID_000390	DDT	50-29-3	+	-	0.163
BUID_000780	IQ.HCl		+	+	0.162
BUID_000553	2,6-Dinitrotoluene	606-20-2	+	N/A	0.151
BUID_000069	2-Aminoanthraquinone	117-79-3	+	+	0.15
BUID_000431	1,2-Dibromoethane	106-93-4	+	+	0.149
BUID_000395	Dehydroepiandrosterone acetate	853-23-6	+	N/A	0.148
BUID_000507	3,3'-Dimethoxybenzidine.2HCl	20325-40-0	+	+	0.144
BUID_001017	2-Nitrofluorene	607-57-8	+	+	0.144
BUID_001435	Triamcinolone acetonide	76-25-5	+	N/A	0.143
BUID_001479	Tris-(1,3-dichloro-2-propyl)phosphate	13674-87-8	+	+	0.141
BUID_000957	Nafenopin	3771-19-5	+	-	0.14
BUID_000791	Isoniazid	54-85-3	+	+	0.139
BUID_000874	Methyl carbamate	598-55-0	+	-	0.138
BUID_001284	C.I. acid red 114	6459-94-5	+	+	0.137
BUID_001376	3,3',4,4'-Tetraaminobiphenyl.4HCl	7411-49-6	+	N/A	0.136
BUID_000276	Chlorendic acid	115-28-6	+	-	0.135

BUID_000713	Hexachlorobenzene	118-74-1	+	-	0.134
BUID_000914	Methyleugenol	93-15-2	+	-	0.133
BUID_000187	C.I. direct black 38	1937-37-7	+	+	0.132
BUID_000053	1-Amino-2,4-dibromoanthraquinone	81-49-2	+	+	0.13
BUID_001315	Safrole	94-59-7	+	-	0.127
BUID_001463	2,4,5-Trimethylaniline	137-17-7	+	+	0.127
BUID_001276	Pyrilamine maleate	59-33-6	+	N/A	0.126
BUID_001010	Nitrobenzene	98-95-3	+	-	0.122
BUID_001518	C.I. disperse yellow 3	2832-40-8	+	+	0.122
BUID_000098	Aramite	140-57-8	+	N/A	0.12
BUID_001085	Nitrosoheptamethyleneimine	20917-49-1	+	+	0.12
BUID_001373	Telone II	542-75-6	+	+	0.12
BUID_001220	Piperonyl butoxide	51-03-6	+	-	0.118
BUID_001425	Toluene diisocyanate	26471-62-5	+	+	0.118
BUID_000744	Hydrazobenzene	122-66-7	+	+	0.117
BUID_001241	Prednisolone	50-24-8	+	-	0.117
BUID_000613	Ethyl alcohol	64-17-5	+	-	0.116
BUID_000913	Methylethylketoxime	96-29-7	+	N/A	0.115
BUID_000947	Monocrotaline	315-22-0	+	-	0.115
BUID_000019	2-Acetylaminofluorene	53-96-3	+	+	0.114
BUID_001145	4,4'-Oxydianiline	101-80-4	+	+	0.114
BUID_000534	3,3'-Dimethylbenzidine.2HCl	612-82-8	+	+	0.113
BUID_000636	di(2-Ethylhexyl)phthalate	117-81-7	+	-	0.112
BUID_000908	4,4'-Methylene-bis(2-methylaniline)	838-88-0	+	+	0.112
BUID_001175	Phenobarbital, sodium	57-30-7	+	-	0.112
BUID_000007	Acetaminophen	103-90-2	+	-	0.11
BUID_000101	Aroclor 1016	12674-11-2	+	N/A	0.11
BUID_000361	Coumarin	91-64-5	+	+	0.11
BUID_000418	2,4-Diaminotoluene	95-80-7	+	+	0.11
BUID_000365	Cupferron	135-20-6	+	+	0.109
BUID_001465	2,4,6-Trimethylaniline.HCl		+	N/A	0.109
BUID_001103	N-Nitrosomorpholine	59-89-2	+	+	0.108
BUID_000578	Doxylamine succinate	562-10-7	+	-	0.107
BUID_001404	Thioacetamide	62-55-5	+	-	0.107
BUID_000925	4-(Methylnitrosamino)-1-(3-pyridyl)-1-(butanone)	64091-91-4	+	N/A	0.106
BUID_001110	N-Nitrosopyrrolidine	930-55-2	+	+	0.106
BUID_000756	1-Hydroxyanthraquinone	129-43-1	+	N/A	0.105
BUID_000558	1,4-Dioxane	123-91-1	+	-	0.104
BUID_000715	a-1,2,3,4,5,6-Hexachlorocyclohexane	319-84-6	+	N/A	0.103
BUID_001073	N-Nitrosodiethanolamine	1116-54-7	+	+	0.103
BUID_001081	N-Nitrosoephedrine	17608-59-2	+	N/A	0.103
BUID_000840	MeIQx	77500-04-0	+	+	0.102
BUID_000299	[4-Chloro-6-(2,3-xylidino)-2-pyri-midinylthio]acetic acid	50892-23-4	+	N/A	0.1
BUID_000987	Nitrite, sodium	7632-00-0	+	+	0.1
BUID_001072	Nitrosodibutylamine	924-16-3	+	+	0.1
BUID_001108	N-Nitrosopiperidine	100-75-4	+	+	0.1
BUID_001498	Vinyl acetate	108-05-4	+	-	0.1
BUID_000343	Ciprofibrate	52214-84-3	+	N/A	0.098
BUID_000315	Chloroform	67-66-3	+	-	0.096
BUID_000498	Diisononyl phthalate	68515-48-0	+	-	0.096
BUID_000104	Aroclor 1260	11096-82-5	+	N/A	0.093
BUID_001078	N-Nitrosodipropylamine	621-64-7	+	+	0.093
BUID_001114	o-Nitrotoluene	88-72-2	+	-	0.09
BUID_000676	Furfural	98-01-1	+	+	0.089
BUID_000377	Cyclopentanone oxime	1192-28-5	+	N/A	0.088
BUID_001302	Retrorsine	480-54-6	+	+	0.086
BUID_001507	N-Vinylpyrrolidone-2	88-12-0	+	-	0.085
BUID_000858	Methapyrilene.HCl	135-23-9	+	-	0.082

BUID_000630	Ethylene thiourea	96-45-7	+	+	0.073
BUID_000658	N-(2-Fluorenyl)-2,2,2-trifluoroacetamide	363-17-7	+	N/A	0.072
BUID_000675	Furan	110-00-9	+	-	0.07
BUID_000394	Dehydroepiandrosterone	53-43-0	+	N/A	0.069
BUID_001113	o-Nitrosotoluene	611-23-4	+	N/A	0.068
BUID_000656	Fluconazole	86386-73-4	+	N/A	0.066
BUID_001074	N-Nitrosodiethylamine	55-18-5	+	+	0.064
BUID_001082	Nitrosoethylmethylamine	10595-95-6	+	+	0.064
BUID_000202	Bromodichloromethane	75-27-4	+	+	0.063
BUID_001408	4,4'-Thiodianiline	139-65-1	+	+	0.061
BUID_000682	Gemfibrozil	25812-30-0	+	N/A	0.06
BUID_000912	4,4'-Methylenedianiline.2HCl	13552-44-8	+	+	0.06
BUID_000937	Metronidazole	443-48-1	+	+	0.06
BUID_000607	Ethionine	13073-35-3	+	-	0.059
BUID_000320	3-(p-Chlorophenyl)-1,1-dimethylurea	150-68-5	+	-	0.057
BUID_000256	Carbon tetrachloride	56-23-5	+	-	0.055
BUID_000609	4-Ethoxy-phenylurea	150-69-6	+	-	0.052
BUID_001185	2-Phenyl-1,3-propanediol dicarbamate	25451-15-4	+	N/A	0.049
BUID_000444	Dichloroacetic acid	79-43-6	+	+	0.047
BUID_000006	Acetamide	60-35-5	+	-	0.046
BUID_000349	Clofibrate	637-07-0	+	N/A	0.046
BUID_000141	Benzidine	92-87-5	+	+	0.044
BUID_000737	Hydrazine sulfate	10034-93-2	+	+	0.042
BUID_001167	Petasitenine	60102-37-6	+	N/A	0.026
BUID_001309	Rotenone	83-79-4	-	-	0.672
BUID_001303	Rhodamine 6G	989-38-8	-	-	0.665
BUID_000721	Hexachlorophene	70-30-4	-	-	0.584
BUID_000862	Methotrexate	59-05-2	-	-	0.494
BUID_001277	Pyrimethamine	58-14-0	-	-	0.458
BUID_000719	Hexachlorocyclopentadiene	77-47-4	-	-	0.407
BUID_001378	2,2',5,5'-Tetrachlorobenzidine	15721-02-5	-	+	0.394
BUID_001386	Tetraethylthiuram disulfide	97-77-8	-	-	0.332
BUID_001456	Tricresyl phosphate	1330-78-5	-	-	0.321
BUID_000380	Cyclosporin A	59865-13-3	-	-	0.31
BUID_000764	8-Hydroxyquinoline	148-24-3	-	+	0.306
BUID_001396	Tetramethylthiuram disulfide	137-26-8	-	+	0.284
BUID_000421	2,5-Diaminotoluene sulfate	6369-59-1	-	+	0.274
BUID_001299	Retinoic acid	302-79-4	-	N/A	0.273
BUID_001338	Sorbic acid	110-44-1	-	-	0.273
BUID_001259	Propyl gallate	121-79-9	-	-	0.26
BUID_000589	Endosulfan	115-29-7	-	-	0.246
BUID_000293	2-Chloro-p-phenylenediamine sulfate	61702-44-1	-	+	0.242
BUID_000963	N-(1-Naphthyl)ethylenediamine.2HCl	1465-25-4	-	+	0.234
BUID_001405	4,4'-Thiobis(6-tert-butyl-m-cresol)	96-69-5	-	-	0.234
BUID_001173	Phenformin.HCl	834-28-6	-	-	0.226
BUID_001394	Tetrakis(hydroxymethyl)phosphonium chloride	124-64-1	-	-	0.204
BUID_001385	Tetracycline.HCl	64-75-5	-	-	0.198
BUID_001489	Turmeric oleoresin (79%-85% curcumin)	8024-37-1	-	-	0.198
BUID_000185	Bisphenol A	80-05-7	-	-	0.193
BUID_000818	Lithocholic acid	434-13-9	-	-	0.192
BUID_000238	Caffeine	58-08-2	-	-	0.187
BUID_000435	Dibutyltin diacetate	1067-33-0	-	-	0.179
BUID_000692	Glutaraldehyde	111-30-8	-	+	0.179
BUID_000144	Benzoate, sodium	532-32-1	-	N/A	0.176

BUID_000302	2-Chloroacetophenone	532-27-4	-	-	0.174
BUID_001004	4-Nitro-o-phenylenediamine	99-56-9	-	+	0.173
BUID_000446	1,2-Dichlorobenzene	95-50-1	-	-	0.172
BUID_000587	Emodin	518-82-1	-	+	0.172
BUID_000904	alpha-Methyldopa sesquihydrate	41372-08-1	-	-	0.166
BUID_001400	Theophylline	58-55-9	-	-	0.166
BUID_000563	Diphenyl-p-phenylenediamine	74-31-7	-	+	0.165
BUID_000733	4-Hexylresorcinol	136-77-6	-	-	0.165
BUID_000644	Etodolac	41340-25-4	-	N/A	0.163
BUID_000111	Aspirin	50-78-2	-	-	0.162
BUID_000556	2,4-Dinitrotoluene	121-14-2	-	+	0.162
BUID_000094	p-Anisidine.HCl	20265-97-8	-	+	0.161
BUID_000408	Diallyl phthalate	131-17-9	-	-	0.161
BUID_000590	Endrin	72-20-8	-	-	0.161
BUID_000455	1,1-Dichloroethane	75-34-3	-	-	0.159
BUID_000618	p,p'-Ethyl-DDD	72-56-0	-	+	0.158
BUID_001183	Phenyl-b-naphthylamine	135-88-6	-	-	0.152
BUID_000896	Methyl parathion	298-00-0	-	+	0.149
BUID_000192	FD & C blue no. 1	3844-45-9	-	N/A	0.148
BUID_000295	3-Chloro-p-toluidine	95-74-9	-	-	0.148
BUID_000397	Deltamethrin	52918-63-5	-	-	0.148
BUID_000375	Cyclohexylamine.HCl	4998-76-9	-	-	0.147
BUID_001293	FD & C red no. 3	16423-68-0	-	-	0.146
BUID_000109	l-Ascorbic acid	50-81-7	-	-	0.144
BUID_000649	Fenthion	55-38-9	-	-	0.144
BUID_001182	1-Phenyl-3-methyl-5-pyrazolone	89-25-8	-	-	0.144
BUID_000258	Carbromal	77-65-6	-	-	0.143
BUID_001250	Promethazine.HCl	58-33-3	-	-	0.142
BUID_000196	HC blue no. 2	33229-34-4	-	+	0.14
BUID_000844	dl-Menthol	15356-70-4	-	-	0.14
BUID_000867	Methoxychlor	72-43-5	-	-	0.139
BUID_000304	p-Chloroaniline	106-47-8	-	+	0.138
BUID_000312	(2-Chloroethyl)trimethylammonium chloride	999-81-5	-	-	0.138
BUID_000571	2,5-Dithiobiurea	142-46-1	-	-	0.138
BUID_001184	N-Phenyl-p-phenylenediamine.HCl	2198-59-6	-	-	0.138
BUID_000125	Azinphosmethyl	86-50-0	-	+	0.137
BUID_000501	Dimethoate	60-51-5	-	+	0.136
BUID_000974	Nickel (II) sulfate hexahydrate	10101-97-0	-	-	0.135
BUID_000151	1H-Benzotriazole	95-14-7	-	+	0.133
BUID_001154	Penicillin VK	132-98-9	-	-	0.133
BUID_000829	Malaoxon	1634-78-2	-	-	0.13
BUID_000836	d-Mannitol	69-65-8	-	-	0.13
BUID_001395	Tetrakis(hydroxymethyl)phosphonium sulfate	55566-30-8	-	-	0.128
BUID_000426	Dibenzo-p-dioxin	262-12-4	-	-	0.126
BUID_000876	2-Methyl-4-chlorophenoxyacetic acid	94-74-6	-	N/A	0.126
BUID_001028	3-Nitropropionic acid	504-88-1	-	+	0.126
BUID_001007	p-Nitroaniline	100-01-6	-	+	0.125
BUID_001523	FD & C yellow no. 5	1934-21-0	-	-	0.125
BUID_000562	Diphenhydramine.HCl	147-24-0	-	-	0.124
BUID_000777	Iodoform	75-47-8	-	+	0.124
BUID_000332	Chlorpheniramine maleate	113-92-8	-	-	0.123
BUID_000024	Acrolein	107-02-8	-	+	0.122
BUID_000559	Dioxathion	78-34-2	-	+	0.121
BUID_000423	Diazinon	333-41-5	-	-	0.12
BUID_000747	Hydrochlorothiazide	58-93-5	-	-	0.12
BUID_000008	Acetohexamide	968-81-0	-	-	0.119
BUID_001152	Parathion	56-38-2	-	-	0.119
BUID_000460	2,4-Dichlorophenoxyacetic acid	94-75-7	-	-	0.118

BUID_001422	Tolazamide	1156-19-0	-	-	0.118
BUID_001486	l-Tryptophan	73-22-3	-	-	0.118
BUID_001503	Vinyl toluene	25013-15-4	-	-	0.115
BUID_000033	Adipamide	628-94-4	-	-	0.114
BUID_000505	2,4-Dimethoxyaniline.HCl	54150-69-5	-	+	0.114
BUID_000454	p,p'-Dichlorodiphenyl sulfone	80-07-9	-	-	0.113
BUID_000538	Dimethylformamide	68-12-2	-	-	0.112
BUID_001213	Phthalic anhydride	85-44-9	-	-	0.112
BUID_001285	C.I. food red 3	3567-69-9	-	+	0.112
BUID_001449	Trichlorofluoromethane	75-69-4	-	-	0.112
BUID_000303	4'-(Chloroacetyl)-acetanilide	140-49-8	-	+	0.111
BUID_001024	1-Nitronaphthalene	86-57-7	-	+	0.111
BUID_001187	1-Phenyl-2-thiourea	103-85-5	-	-	0.11
BUID_001524	FD & C yellow no. 6	2783-94-0	-	-	0.11
BUID_000333	Chlorpropamide	94-20-2	-	-	0.109
BUID_000798	Isopropanol	67-63-0	-	-	0.107
BUID_001423	Tolbutamide	64-77-7	-	-	0.107
BUID_000725	Hexamethylenetetramine	100-97-0	-	+	0.105
BUID_000336	Choline chloride	67-48-1	-	-	0.104
BUID_000360	Coumaphos	56-72-4	-	-	0.104
BUID_000648	Fenaminosulf, formulated	140-56-7	-	+	0.104
BUID_001335	Sodium diethyldithiocarbamate trihydrate	148-18-5	-	-	0.104
BUID_000230	g-Butyrolactone	96-48-0	-	-	0.103
BUID_000477	N,N-Diethyl-m-toluamide	134-62-3	-	-	0.102
BUID_000830	Malathion	121-75-5	-	-	0.102
BUID_000657	Fluometuron	2164-17-2	-	-	0.098
BUID_000520	Dimethyl terephthalate	120-61-6	-	-	0.097
BUID_000787	Isobutyraldehyde	78-84-2	-	-	0.097
BUID_000340	Cimetidine	51481-61-9	-	N/A	0.095
BUID_000646	Eugenol	97-53-0	-	-	0.095
BUID_000373	Cyclohexanone	108-94-1	-	-	0.094
BUID_001149	Oxytetracycline.HCl	2058-46-0	-	-	0.093
BUID_001319	Scopolamine hydrobromide trihydrate	6533-68-2	-	-	0.093
BUID_001370	Tegafur	37076-68-9	-	N/A	0.092
BUID_000946	Monochloroacetic acid	79-11-8	-	-	0.091
BUID_001009	4-Nitroanthranilic acid	619-17-0	-	+	0.091
BUID_001212	Phthalamide	88-96-0	-	-	0.091
BUID_000855	Methacrylonitrile	126-98-7	-	-	0.09
BUID_001274	Pyrazinamide	98-96-4	-	-	0.089
BUID_001453	2,4,5-Trichlorophenoxyacetic acid	93-76-5	-	-	0.089
BUID_001493	Urea	57-13-6	-	-	0.088
BUID_000155	Benzyl alcohol	100-51-6	-	-	0.086
BUID_001176	Phenol	108-95-2	-	-	0.086
BUID_001446	1,1,1-Trichloroethane, technical grade	71-55-6	-	+	0.084
BUID_000040	Aldicarb	116-06-3	-	-	0.083
BUID_000193	FD & C blue no. 2	860-22-0	-	-	0.082
BUID_001515	Xylene mixture (60% m-xylene, 9% o-xylene, 14% p-xylene, 17% ethylbenzene)	1330-20-7	-	-	0.082
BUID_000457	2,4-Dichlorophenol	120-83-2	-	-	0.081
BUID_000568	Dipropylene glycol	25265-71-8	-	-	0.08
BUID_001192	m-Phenylenediamine.2HCl	541-69-5	-	+	0.079
BUID_001349	Succinic anhydride	108-30-5	-	-	0.076
BUID_000134	Barium chloride dihydrate	10326-27-9	-	-	0.074
BUID_000211	n-Butyl chloride	109-69-3	-	-	0.071
BUID_000344	Citral	5392-40-5	-	-	0.071
BUID_000010	Acetonitrile	75-05-8	-	-	0.07
BUID_000084	Ampicillin trihydrate	7177-48-2	-	-	0.07
BUID_001195	Phenylephrine.HCl	61-76-7	-	-	0.07
BUID_000831	Maleic hydrazide	123-33-1	-	-	0.068
BUID_000885	Methyl methacrylate	80-62-6	-	-	0.068

BUID_001281	Quinapril.HCl	82586-55-8	-	N/A	0.068
BUID_001298	Resorcinol	108-46-3	-	-	0.068
BUID_000246	Caprolactam	105-60-2	-	-	0.066
BUID_001132	C.I. acid orange 10	1936-15-8	-	-	0.065
BUID_001002	3-Nitro-4-hydroxyphenylarsonic acid	121-19-7	-	-	0.063
BUID_000800	Isopropyl-N-(3-chlorophenyl)carbamate	101-21-3	-	N/A	0.062
BUID_001137	Oxamyl	23135-22-0	-	N/A	0.059
BUID_001356	Sulfisoxazole	127-69-5	-	-	0.058
BUID_001526	HC yellow 4	59820-43-8	-	+	0.055
BUID_000259	b-Carotene	7235-40-7	-	+	0.054
BUID_001415	Tin (II) chloride	7772-99-8	-	-	0.054
BUID_001401	Thiabendazole	148-79-8	-	+	0.052
BUID_000704	FD & C green no. 3	2353-45-9	-	-	0.05
BUID_000148	Benzoin	119-53-9	-	-	0.048
BUID_001475	Triprolidine.HCl monohydrate	6138-79-0	-	-	0.048
BUID_001296	HC red no. 3	2871-01-4	-	+	0.047
BUID_000341	trans-Cinnamaldehyde	14371-10-9	-	-	0.045
BUID_000684	Geranyl acetate	mixture (105-87-3)	-	-	0.045
BUID_000503	Dimethoxane	828-00-2	-	+	0.042
BUID_000317	2-(Chloromethyl)pyridine.HCl	6959-47-3	-	+	0.038
BUID_001214	Picloram, technical grade	1918-02-1	-	-	0.028
BUID_001711	Suberoylanilide hydroxamic acid	149647-78-9	N/A	N/A	0.605
BUID_000143	Benzo(a)pyrene	50-32-8	N/A	+	0.46
BUID_002874	Indoxyl Sulfate	2642-37-7	N/A	N/A	0.341
BUID_002897	sulforaphane	4478-93-7	N/A	N/A	0.304
BUID_000533	7,12-Dimethylbenz(a)anthracene	57-97-6	N/A	+	0.302
BUID_001648	Triclosan	3380-34-5	N/A	N/A	0.214
BUID_001967	Clotrimazole	23593-75-1	N/A	N/A	0.211
BUID_002901	2-(1'H-indole-3'-carbonyl)-thiazole-4-carboxylic acid methyl ester	448906-42-1	N/A	N/A	0.207
BUID_002354	Quinoxyfen	124495-18-7	N/A	N/A	0.201
BUID_000225	tert-Butylhydroquinone	1948-33-0	N/A	-	0.198
BUID_002870	1-Methyl-N-[2-methyl-4-[2-(2-methylphenyl)diazenyl]phen yl-1H-pyrazole-5-carboxamide	301326-22-7	N/A	N/A	0.176
BUID_001838	3,3',5,5'-Tetrabromobisphenol A	79-94-7	N/A	N/A	0.173
BUID_002612	Spironolactone	52-01-7	N/A	N/A	0.168
BUID_002485	Prallethrin	23031-36-9	N/A	N/A	0.161
BUID_002896	nifedipine	21829-25-4	N/A	N/A	0.16
BUID_002873	Indole-3-carbinol (I3C)	700-06-1	N/A	N/A	0.147
BUID_002881	rosiglitazone	122320-73-4	N/A	N/A	0.146
BUID_002875	4-Hydroxy-2-quinolinecarboxylic acid	492-27-3	N/A	N/A	0.135
BUID_002883	Tris(2-chloroethyl)phosphate	13674-84-5	N/A	N/A	0.133
BUID_002886	Acid Orange 156	68555-86-2	N/A	N/A	0.124
BUID_002882	triphenyl phosphine oxide	791-28-6	N/A	N/A	0.118
BUID_002895	Pregnenolone 16alpha-carbonitrile	1434-54-4	N/A	N/A	0.118
BUID_002889	pregnenolone	145-13-1	N/A	N/A	0.117
BUID_002877	4,8-Dihydroxy-2-quinolinecarboxylic acid	59-00-7	N/A	N/A	0.101
BUID_002879	indole-3-aldehyde	487-89-8	N/A	N/A	0.099
BUID_002871	2-Amino-3-oxo-3H-phenoxazine-1,9-dicarboxylic acid	606-59-7	N/A	N/A	0.092
BUID_001476	Tris(2-chloroethyl)phosphate	115-96-8	N/A	-	0.088
BUID_002309	Triphenyl phosphate	115-86-6	N/A	N/A	0.085
BUID_002884	2-ethyl-2-hexenal	645-62-5	N/A	N/A	0.069

BUID_002904	Glycel	70901-12-1	N/A	N/A	0.058
BUID_000210	Butyl benzyl phthalate	85-68-7	N/A	-	0.057
BUID_002900	L-kynurenine	2922-83-0	N/A	N/A	0.057
BUID_001439	Tributyl phosphate	126-73-8	N/A	-	0.051
BUID_002586	Mono(2-ethylhexyl) phthalate	4376-20-9	N/A	N/A	0.034

Table S2 Ranked list of differentially enriched pathways (c2 reactome) between carcinogens vs. non-carcinogens across multiple TAS subsets

Signature	Rank	Pathway ID	P.Value TAS>0	P.Value TAS>0.2	P.Value TAS>0.3	P.Value TAS>0.4	P.Value Combined	FDR Combined
CARC_UP	1	REACTOME_EXTRINSIC_PATHWAY_FOR_APOPTOSIS	1.24E-01	3.53E-04	1.94E-02	2.38E-04	4.29E-07	1.45E-04
CARC_UP	2	REACTOME_TRANSCRIPTION	1.17E-01	4.09E-04	2.30E-03	8.71E-04	2.24E-07	1.45E-04
CARC_UP	3	REACTOME_RNA_POL_III_TRANSCRIPTION	3.86E-02	7.84E-03	1.46E-02	1.54E-04	1.23E-06	2.77E-04
CARC_UP	4	REACTOME_NEF_MEDIATED_DOWNREGULATION_OF_MHC_CLASS_I_COMPLEX_CELL_SURFACE_EXPRESSION	5.16E-04	9.06E-03	3.93E-03	1.11E-01	3.17E-06	5.34E-04
CARC_UP	5	REACTOME_TRAF3_DEPENDENT_IRF_ACTIVATION_PATHWAY	8.13E-03	2.45E-02	2.59E-02	7.75E-04	5.64E-06	6.34E-04
CARC_UP	6	REACTOME_NFKB_ACTIVATION_THROUGH_FADD_RIP1_PATHWAY_MEDIATED_BY_CASPASE_8_AND10	1.73E-02	4.69E-03	2.10E-02	2.06E-03	5.05E-06	6.34E-04
CARC_UP	7	REACTOME_APOPTOSIS_INDUCED_DNA_FRAGMENTATION	1.91E-01	1.55E-03	5.93E-03	3.61E-03	8.35E-06	8.04E-04
CARC_UP	8	REACTOME_METABOLISM_OF_NON_CODING_RNA	4.99E-01	3.12E-02	1.56E-02	1.38E-04	3.41E-05	2.87E-03
CARC_UP	9	REACTOME_ABORTIVE_ELONGATION_OF_HIV1_TRANSCRIPT_IN_THE_ABSENCE_OF_TAT	4.17E-01	1.01E-01	6.45E-03	1.53E-04	4.09E-05	3.06E-03
CARC_UP	10	REACTOME_HS_GAG_DEGRADATION	7.20E-03	9.13E-03	2.48E-02	4.07E-02	6.03E-05	3.42E-03
CARC_UP	11	REACTOME_MRNA_DECAY_BY_5_TO_3_EXORIBONUCLEASE	4.63E-01	8.63E-02	2.49E-02	6.76E-05	6.09E-05	3.42E-03
CARC_UP	12	REACTOME_RNA_POL_III_TRANSCRIPTION_INITIATION_FROM_TYPE_3_PROMOTER	2.56E-01	1.44E-02	1.75E-02	8.57E-04	5.16E-05	3.42E-03
CARC_UP	13	REACTOME_ACTIVATED_AMPK_STIMULATES_FATTY_ACID_OXIDATION_IN_MUSCLE	2.11E-02	5.72E-02	2.79E-02	2.47E-03	7.25E-05	3.76E-03
CARC_UP	14	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	8.11E-05	1.56E-01	1.48E-01	1.08E-01	1.51E-04	6.78E-03
CARC_UP	15	REACTOME_DEADENYLATION_DEPENDENT_MRNA_DECAY	4.91E-01	7.36E-03	2.68E-02	2.07E-03	1.50E-04	6.78E-03
CARC_UP	16	REACTOME_RESOLUTION_OF_AP_SITES_VIA_THE_MULTIPLE_NUCLEOTIDE_PATCH_REPLACEMENT_PATHWAY	4.81E-01	3.32E-01	1.08E-02	1.42E-04	1.77E-04	7.14E-03
CARC_UP	17	REACTOME_MRNA_3_END_PROCESSING	2.58E-01	1.42E-01	6.71E-03	1.02E-03	1.80E-04	7.14E-03
CARC_UP	18	REACTOME_ELONGATION_ARREST_AND_RECOVERY	3.30E-01	9.85E-02	2.02E-02	5.71E-04	2.49E-04	9.32E-03
CARC_UP	19	REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM	4.82E-01	5.49E-02	1.75E-02	8.80E-04	2.67E-04	9.46E-03
CARC_UP	20	REACTOME_RNA_POL_I_RNA_POL_III_AND_MITOCHONDRIAL_TRANSCRIPTION	3.77E-01	7.44E-04	4.54E-02	5.26E-02	3.98E-04	1.34E-02
CARC_UP	21	REACTOME_NUCLEOTIDE_EXCISION_REPAIR	4.82E-01	2.52E-01	1.19E-02	6.33E-04	5.10E-04	1.45E-02
CARC_UP	22	REACTOME_FORMATION_OF_INCISION_COMPLEX_IN_GG_NER	4.18E-01	2.44E-02	1.48E-02	6.11E-03	5.15E-04	1.45E-02
CARC_UP	23	REACTOME_RNA_POL_II_PRE_TRANSCRIPTION_EVENTS	3.64E-01	6.20E-02	1.98E-02	1.81E-03	4.64E-04	1.45E-02
CARC_UP	24	REACTOME_CLEAVAGE_OF_GROWING_TRANSCRIPT_IN_THE_TERMINATION_REGION_	3.24E-01	1.47E-01	9.18E-03	2.08E-03	5.10E-04	1.45E-02
CARC_UP	25	REACTOME_UNWINDING_OF_DNA	3.35E-01	8.75E-02	1.46E-02	2.42E-03	5.64E-04	1.46E-02
CARC_UP	26	REACTOME_RNA_POL_III_TRANSCRIPTION_INITIATION_FROM_TYPE_2_PROMOTER	1.06E-01	2.75E-02	6.47E-02	5.26E-03	5.48E-04	1.46E-02
CARC_UP	27	REACTOME_RNA_POL_II_TRANSCRIPTION	3.97E-01	1.11E-01	1.88E-02	1.66E-03	7.09E-04	1.77E-02
CARC_UP	28	REACTOME_FORMATION_OF_THE_HIV1_EARLY_ELONGATION_COMPLEX	4.16E-01	7.00E-02	2.20E-02	2.38E-03	7.69E-04	1.85E-02
CARC_UP	29	REACTOME_GLOBAL_GENOMIC_NER_GG_NER	4.88E-01	2.21E-01	1.54E-02	1.39E-03	1.07E-03	2.48E-02

CARC_UP	30	REACTOME_BASE_EXCISION_REPAIR	4.94E-01	4.80E-01	1.82E-02	6.20E-04	1.20E-03	2.70E-02
CARC_UP	31	REACTOME_INTERFERON_GAMMA_SIGNALING	1.93E-04	2.19E-01	2.61E-01	2.88E-01	1.37E-03	2.98E-02
CARC_UP	32	REACTOME_CREB_PHOSPHORYLATION_THROUGH_THE_ACTIVATION_OF_RAS	1.74E-02	1.57E-02	7.01E-02	1.87E-01	1.51E-03	2.99E-02
CARC_UP	33	REACTOME_VIRAL_MESSENGER_RNA_SYNTHESIS	4.25E-01	1.10E-02	1.50E-01	5.12E-03	1.50E-03	2.99E-02
CARC_UP	34	REACTOME_AKT_PHOSPHORYLATES_TARGETS_IN_THE_CYTOSOL	1.47E-01	2.41E-02	2.99E-02	3.38E-02	1.51E-03	2.99E-02
CARC_UP	35	REACTOME_NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE	2.03E-02	1.62E-01	7.06E-02	1.69E-02	1.62E-03	3.06E-02
CARC_UP	36	REACTOME_MITOTIC_G1_G1_S_PHASES	4.57E-01	3.12E-02	4.99E-02	5.58E-03	1.63E-03	3.06E-02
CARC_UP	37	REACTOME_POST_NMDA_RECEPTOR_ACTIVATION_EVENTS	7.25E-03	9.12E-03	2.05E-01	3.41E-01	1.84E-03	3.17E-02
CARC_UP	38	REACTOME_FORMATION_OF_RNA_POL_II_ELONGATION_COMPLEX_	3.44E-01	8.21E-02	3.75E-02	4.08E-03	1.75E-03	3.17E-02
CARC_UP	39	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT	5.56E-02	4.35E-03	2.14E-01	8.69E-02	1.80E-03	3.17E-02
CARC_UP	40	REACTOME_SLBP_DEPENDENT_PROCESSING_OF_REPLICATION_DEPENDENT_HISTONE_PRE_MRNAS	3.61E-01	8.28E-02	1.70E-02	9.69E-03	1.93E-03	3.26E-02
CARC_DN	1	REACTOME_GLUCURONIDATION	2.16E-03	4.37E-05	1.10E-04	1.92E-05	1.71E-12	1.15E-09
CARC_DN	2	REACTOME_INTEGRIN_CELL_SURFACE_INTERACTIONS	2.64E-01	4.75E-04	4.30E-06	6.09E-06	2.23E-11	7.51E-09
CARC_DN	3	REACTOME_COSTIMULATION_BY_THE_CD28_FAMILY	1.03E-01	1.05E-04	4.90E-06	3.38E-04	1.04E-10	2.33E-08
CARC_DN	4	REACTOME_TIGHT_JUNCTION_INTERACTIONS	1.99E-01	1.89E-03	1.65E-05	5.45E-06	1.85E-10	3.12E-08
CARC_DN	5	REACTOME_AXON_GUIDANCE	4.35E-01	5.86E-03	1.57E-05	1.36E-06	2.85E-10	3.85E-08
CARC_DN	6	REACTOME_GLUTAMATE_NEUROTRANSMITTER_RELEASE_CYCLE	1.52E-01	8.81E-03	2.93E-05	3.50E-06	6.61E-10	7.42E-08
CARC_DN	7	REACTOME_TRANSPORT_TO_THE_GOLGI_AND_SUBSEQUENT_MODIFICATION	5.43E-04	1.34E-04	1.18E-03	3.19E-03	1.22E-09	9.94E-08
CARC_DN	8	REACTOME_CTLA4_INHIBITORY_SIGNALING	5.93E-02	1.69E-03	3.56E-05	7.12E-05	1.15E-09	9.94E-08
CARC_DN	9	REACTOME_TRAFFICKING_OF_AMPA_RECEPTORS	2.98E-01	8.39E-04	4.77E-04	2.51E-06	1.33E-09	9.94E-08
CARC_DN	10	REACTOME_COMMON_PATHWAY	9.99E-03	5.36E-05	1.73E-03	4.80E-04	1.90E-09	1.28E-07
CARC_DN	11	REACTOME_GLYCOPROTEIN_HORMONES	9.54E-03	9.93E-03	6.88E-04	1.54E-05	3.93E-09	2.17E-07
CARC_DN	12	REACTOME_CELL_CELL_JUNCTION_ORGANIZATION	4.76E-01	2.14E-03	5.39E-05	2.13E-05	4.51E-09	2.17E-07
CARC_DN	13	REACTOME_ELEVATION_OF_CYTOSOLIC_CA2_LEVELS	1.30E-01	1.21E-02	1.36E-04	5.26E-06	4.37E-09	2.17E-07
CARC_DN	14	REACTOME_SEMAPHORIN_INTERACTIONS	1.31E-01	8.61E-03	1.74E-05	4.62E-05	3.59E-09	2.17E-07
CARC_DN	15	REACTOME_CELL_CELL_COMMUNICATION	4.74E-01	2.03E-02	2.76E-05	6.93E-06	6.77E-09	3.04E-07
CARC_DN	16	REACTOME_CD28_DEPENDENT_VAV1_PATHWAY	8.92E-02	9.38E-04	1.29E-04	3.28E-04	1.22E-08	5.13E-07
CARC_DN	17	REACTOME_CD28_CO_STIMULATION	1.99E-01	1.18E-03	8.47E-05	2.40E-04	1.59E-08	6.30E-07
CARC_DN	18	REACTOME_REGULATION_OF_INSULIN_SECRETION_BY_GLUCAGON_LIKE_PEPTIDE1	4.99E-01	3.01E-02	1.13E-04	3.50E-06	1.92E-08	7.20E-07
CARC_DN	19	REACTOME_GOLGI_ASSOCIATED_VESICLE_BIOGENESIS	6.04E-02	6.74E-03	6.96E-04	3.71E-05	3.20E-08	1.13E-06
CARC_DN	20	REACTOME_L1CAM_INTERACTIONS	3.04E-01	1.01E-01	5.30E-04	1.13E-06	5.25E-08	1.77E-06
CARC_DN	21	REACTOME_PLATELET_ADHESION_TO_EXPOSED_COLLAGEN	3.79E-01	2.65E-03	2.44E-04	8.71E-05	5.98E-08	1.92E-06
CARC_DN	22	REACTOME_PROSTACYCLIN_SIGNALLING_THROUGH_PROSTACYCLIN_RECEPTOR	3.73E-01	2.83E-03	2.44E-04	1.19E-04	8.20E-08	2.51E-06
CARC_DN	23	REACTOME_CELL_JUNCTION_ORGANIZATION	4.68E-01	1.14E-02	1.35E-04	5.61E-05	1.05E-07	3.08E-06

CARC_DN	24	REACTOME_PD1_SIGNALING	2.09E-01	1.73E-04	1.73E-04	7.38E-03	1.18E-07	3.31E-06
CARC_DN	25	REACTOME_PEPTIDE_HORMONE_BIOSYNTHESIS	5.09E-02	2.70E-02	1.27E-03	3.43E-05	1.48E-07	3.89E-06
CARC_DN	26	REACTOME_ACETYLCHOLINE_NEUROTRANSMITTER_RELEASE_CYCLE	4.17E-01	2.24E-03	5.11E-04	1.27E-04	1.50E-07	3.89E-06
CARC_DN	27	REACTOME_DEVELOPMENTAL_BIOLOGY	4.44E-01	5.13E-02	1.77E-04	1.70E-05	1.68E-07	4.07E-06
CARC_DN	28	REACTOME_PHOSPHORYLATION_OF_CD3_AND_TCR_ZETA_CHAINS	1.11E-01	3.37E-04	3.18E-04	5.82E-03	1.69E-07	4.07E-06
CARC_DN	29	REACTOME_ACTIVATION_OF_RAC	1.63E-01	1.70E-02	9.39E-04	2.80E-05	1.75E-07	4.08E-06
CARC_DN	30	REACTOME_FORMATION_OF_FIBRIN_CLOT_CLOTTING_CASCADE	9.32E-02	4.32E-04	1.19E-03	1.60E-03	1.84E-07	4.13E-06
CARC_DN	31	REACTOME_REGULATION_OF_INSULIN_LIKE_GROWTH_FACTOR_IGF_ACTIVITY_BY_INSULIN_LIKE_GROWTH_FACTOR_BINDING_PROTEINS_IGFBPS	1.80E-02	3.84E-03	1.28E-03	1.06E-03	2.20E-07	4.78E-06
CARC_DN	32	REACTOME_ACTIVATION_OF_KAINATE_RECEPTORS_UPON_GLUTAMATE_BINDING	3.85E-01	6.11E-03	4.12E-04	1.54E-04	3.30E-07	6.95E-06
CARC_DN	33	REACTOME_NETRIN1_SIGNALING	2.66E-01	1.29E-02	6.80E-04	8.95E-05	4.42E-07	9.03E-06
CARC_DN	34	REACTOME_AMINO_ACID_SYNTHESIS_AND_INTERCONVERSION_TRANSAMINATION	2.13E-02	4.88E-05	7.34E-03	3.08E-02	4.90E-07	9.72E-06
CARC_DN	35	REACTOME_N_GLYCAN_ANTENNAE_ELONGATION	1.96E-02	5.07E-03	3.04E-03	1.33E-03	7.82E-07	1.46E-05
CARC_DN	36	REACTOME_THROMBIN_SIGNALLING_THROUGH_PROTEINASE_ACTIVATED_RECEPTORS_PARS	4.67E-01	1.15E-02	6.69E-04	1.11E-04	7.77E-07	1.46E-05
CARC_DN	37	REACTOME_ION_TRANSPORT_BY_P_TYPE_ATPASES	1.08E-01	5.48E-02	2.28E-03	4.29E-05	1.08E-06	1.96E-05
CARC_DN	38	REACTOME_N_GLYCAN_ANTENNAE_ELONGATION_IN_THE_MEDIAL_TRANS_GOLGI	1.53E-02	1.10E-02	1.73E-03	2.08E-03	1.12E-06	1.98E-05
CARC_DN	39	REACTOME_PLATELET_HOMEOSTASIS	4.62E-01	2.97E-02	5.63E-04	8.36E-05	1.18E-06	2.04E-05
CARC_DN	40	REACTOME_SIGNALING_BY_PDGF	3.58E-01	2.68E-03	2.87E-04	3.53E-03	1.68E-06	2.83E-05

Table S3 Ranked list of differentially enriched pathways (c2 reactome) between genotoxicants vs. non-genotoxicants across multiple TAS subsets

Signature	Rank	Pathway ID	P.Value TAS>0	P.Value TAS>0.2	P.Value TAS>0.3	P.Value TAS>0.4	P.Value Combined	FDR Combined
GTX_UP	1	REACTOME_FORMATION_OF_INCISION_COMPLEX_IN_GG_NER	2.54E-03	3.90E-06	1.01E-05	1.46E-05	1.82E-14	1.23E-11
GTX_UP	2	REACTOME_AKT_PHOSPHORYLATES_TARGETS_IN_THE_CYTOSOL	2.01E-02	3.17E-05	1.05E-05	1.27E-05	7.80E-13	2.63E-10
GTX_UP	3	REACTOME_EXTRINSIC_PATHWAY_FOR_APOPTOSIS	1.27E-03	4.38E-05	1.13E-04	2.35E-05	1.29E-12	2.89E-10
GTX_UP	4	REACTOME_PRE_NOTCH_TRANSCRIPTION_AND_TRANSLATION	6.16E-02	1.55E-05	5.72E-05	1.11E-05	4.74E-12	7.99E-10
GTX_UP	5	REACTOME_GLOBAL_GENOMIC_NER_GG_NER	2.63E-02	1.83E-05	1.48E-04	2.11E-05	1.08E-11	1.46E-09
GTX_UP	6	REACTOME_NEF_MEDIATED_DOWNREGULATION_OF_MHC_CLASS_I_COMPLEX_CELL_SURFACE_EXPRESSION	7.12E-04	3.39E-04	2.21E-04	2.14E-03	5.57E-10	6.25E-08
GTX_UP	7	REACTOME_NUCLEOTIDE_EXCISION_REPAIR	6.59E-02	7.06E-05	6.48E-04	8.19E-05	1.12E-09	1.08E-07
GTX_UP	8	REACTOME_ACTIVATED_AMPK_STIMULATES_FATTY_ACID_OXIDATION_IN_MUSCLE	8.02E-02	6.93E-04	2.34E-04	2.18E-04	9.95E-09	8.38E-07
GTX_UP	9	REACTOME_RNA_POL_I_TRANSCRIPTION_INITIATION	6.99E-03	2.11E-03	1.69E-03	2.59E-03	1.59E-07	1.19E-05
GTX_UP	10	REACTOME_INTRINSIC_PATHWAY_FOR_APOPTOSIS	5.34E-02	3.29E-04	8.36E-03	5.59E-04	1.96E-07	1.32E-05
GTX_UP	11	REACTOME_MRNA_DECAY_BY_5_TO_3_EXORIBONUCLEASE	2.21E-01	7.81E-03	3.81E-03	2.05E-05	3.02E-07	1.85E-05
GTX_UP	12	REACTOME_APOPTOSIS	1.97E-02	4.76E-04	7.18E-03	2.24E-03	3.34E-07	1.88E-05
GTX_UP	13	REACTOME_FORMATION_OF_TRANSCRIPTION_COUPLED_NER_TC_NER_REPAIR_COMPLEX	7.09E-02	3.31E-04	3.18E-03	2.76E-03	4.38E-07	2.15E-05
GTX_UP	14	REACTOME_P53_DEPENDENT_G1_DNA_DAMAGE_RESPONSE	1.24E-01	7.02E-04	2.59E-03	9.34E-04	4.46E-07	2.15E-05
GTX_UP	15	REACTOME_ELONGATION_ARREST_AND_RECOVERY	1.42E-01	2.71E-03	4.08E-03	1.58E-04	5.12E-07	2.30E-05
GTX_UP	16	REACTOME_FORMATION_OF_RNA_POL_II_ELONGATION_COMPLEX_	8.21E-02	1.89E-03	3.51E-03	5.13E-04	5.68E-07	2.39E-05
GTX_UP	17	REACTOME_PROCESSING_OF_INTRONLESS_PRE_MRNAS	3.48E-01	4.63E-03	3.14E-03	8.11E-05	7.96E-07	3.16E-05
GTX_UP	18	REACTOME_ACTIVATION_OF_BH3_ONLY_PROTEINS	1.95E-01	1.62E-03	6.37E-03	3.98E-04	1.42E-06	5.31E-05
GTX_UP	19	REACTOME_DNA_REPAIR	7.99E-02	8.62E-04	1.44E-02	1.06E-03	1.79E-06	6.36E-05
GTX_UP	20	REACTOME_RNA_POL_II_PRE_TRANSCRIPTION_EVENTS	2.10E-01	1.30E-03	4.29E-03	1.44E-03	2.70E-06	9.09E-05
GTX_UP	21	REACTOME_FORMATION_OF_THE_HIV1_EARLY_ELONGATION_COMPLEX	1.15E-01	1.74E-03	5.97E-03	1.55E-03	2.92E-06	9.39E-05
GTX_UP	22	REACTOME_RNA_POL_I_TRANSCRIPTION_TERMINATION	1.22E-02	3.79E-03	3.71E-03	1.35E-02	3.56E-06	1.09E-04
GTX_UP	23	REACTOME_NOTCH_HLH_TRANSCRIPTION_PATHWAY	1.63E-01	8.84E-03	9.80E-04	2.08E-03	4.33E-06	1.27E-04
GTX_UP	24	REACTOME_RNA_POL_III_TRANSCRIPTION	2.65E-01	2.02E-03	9.07E-03	6.43E-04	4.58E-06	1.29E-04
GTX_UP	25	REACTOME_TRANSCRIPTION	1.87E-01	3.54E-03	4.81E-03	1.17E-03	5.28E-06	1.42E-04
GTX_UP	26	REACTOME_SCFSKP2_MEDIATED_DEGRADATION_OF_P27_P21	2.16E-01	2.02E-03	6.57E-03	1.47E-03	5.89E-06	1.53E-04
GTX_UP	27	REACTOME_ABORTIVE_ELONGATION_OF_HIV1_TRANSCRIPT_IN_THE_ABSENCE_OF_TAT	2.22E-01	4.24E-03	9.06E-03	5.39E-04	6.37E-06	1.59E-04
GTX_UP	28	REACTOME_RESOLUTION_OF_AP_SITES_VIA_THE_MULTIPLE_NUCLEOTIDE_PATCH_REPLACEMENT_PATHWAY	1.72E-01	1.45E-02	2.14E-02	9.35E-05	6.84E-06	1.65E-04
GTX_UP	29	REACTOME_PURINE_CATABOLISM	5.06E-03	7.70E-03	1.89E-02	7.58E-03	7.49E-06	1.71E-04

GTX_UP	30	REACTOME_HS_GAG_DEGRADATION	1.56E-01	3.46E-03	4.70E-03	2.24E-03	7.63E-06	1.71E-04
GTX_UP	31	REACTOME_TRANSCRIPTION_COUPLED_NER_TC_NER	2.28E-01	1.37E-03	1.24E-02	2.54E-03	1.22E-05	2.66E-04
GTX_UP	32	REACTOME_ER_PHAGOSOME_PATHWAY	1.22E-01	5.51E-03	1.39E-02	1.31E-03	1.46E-05	3.07E-04
GTX_UP	33	REACTOME_BASE_EXCISION_REPAIR	1.54E-01	1.48E-02	2.63E-02	3.07E-04	2.07E-05	4.23E-04
GTX_UP	34	REACTOME_RECRUITMENT_OF_MITOTIC_CENTROSOME_PROTEINS_AND_COMPLEXES	3.75E-01	1.33E-02	3.28E-03	2.17E-03	3.59E-05	7.11E-04
GTX_UP	35	REACTOME_CELL_CYCLE_CHECKPOINTS	3.04E-02	5.52E-03	3.99E-02	6.06E-03	3.99E-05	7.68E-04
GTX_UP	36	REACTOME_TRAF3_DEPENDENT_IRF_ACTIVATION_PATHWAY	4.40E-03	2.27E-02	2.78E-02	1.74E-02	4.62E-05	8.22E-04
GTX_UP	37	REACTOME_DOWNSTREAM_SIGNALING_EVENTS_OF_B_CELL_RECEPTOR_BCR	3.42E-02	8.38E-03	1.67E-02	1.02E-02	4.64E-05	8.22E-04
GTX_UP	38	REACTOME_SYNTHESIS_OF_DNA	3.35E-02	6.00E-03	4.09E-02	5.81E-03	4.58E-05	8.22E-04
GTX_UP	39	REACTOME_RNA_POL_III_TRANSCRIPTION_INITIATION_FROM_TYPE_3_PROMOTER	2.04E-02	6.66E-03	4.54E-02	8.65E-03	5.02E-05	8.67E-04
GTX_UP	40	REACTOME_LOSS_OF_NLP_FROM_MITOTIC_CENTROSOMES	4.19E-01	1.16E-02	5.27E-03	2.52E-03	5.89E-05	9.68E-04
GTX_DN	1	REACTOME_PYRUVATE_METABOLISM	3.13E-03	3.00E-06	2.36E-06	2.97E-05	8.63E-15	5.82E-12
GTX_DN	2	REACTOME_ACTIVATION_OF_KAINATE_RECEPTORS_UPON_GLUTAMATE_BINDING	1.36E-01	1.68E-04	3.61E-06	2.45E-06	1.73E-12	3.88E-10
GTX_DN	3	REACTOME_ION_CHANNEL_TRANSPORT	2.97E-01	5.64E-05	9.30E-06	9.83E-07	1.34E-12	3.88E-10
GTX_DN	4	REACTOME_IONOTROPIC_ACTIVITY_OF_KAINATE_RECEPTORS	3.96E-02	4.74E-04	6.12E-05	5.38E-07	4.83E-12	6.51E-10
GTX_DN	5	REACTOME_CREB_PHOSPHORYLATION_THROUGH_THE_ACTIVATION_OF_CAMKII	8.22E-02	3.66E-05	3.40E-05	5.13E-06	4.16E-12	6.51E-10
GTX_DN	6	REACTOME_TRANSPORT_TO_THE_GOLGI_AND_SUBSEQUENT_MODIFICATION	3.75E-05	1.33E-05	1.65E-04	1.51E-02	9.16E-12	1.03E-09
GTX_DN	7	REACTOME_PLATELET_CALCIUM_HOMEOSTASIS	4.09E-01	2.47E-04	3.44E-05	5.23E-07	1.29E-11	1.24E-09
GTX_DN	8	REACTOME_INTEGRIN_CELL_SURFACE_INTERACTIONS	1.56E-05	4.58E-05	2.61E-03	3.84E-03	4.51E-11	3.80E-09
GTX_DN	9	REACTOME_THROMBIN_SIGNALLING_THROUGH_PROTEINASE_ACTIVATED_RECEPTORS_PARS	1.03E-01	5.44E-05	4.38E-05	4.94E-05	7.32E-11	5.48E-09
GTX_DN	10	REACTOME_ION_TRANSPORT_BY_P_TYPE_ATPASES	4.97E-01	2.83E-04	1.58E-05	6.87E-06	9.01E-11	6.07E-09
GTX_DN	11	REACTOME_GLYCOPROTEIN_HORMONES	1.33E-01	1.28E-03	4.58E-05	3.23E-06	1.42E-10	8.73E-09
GTX_DN	12	REACTOME_NETRIN1_SIGNALING	1.75E-01	2.59E-05	9.66E-05	7.00E-05	1.70E-10	9.52E-09
GTX_DN	13	REACTOME_PLATELET_ADHESION_TO_EXPOSED_COLLAGEN	7.65E-02	2.07E-03	1.35E-04	7.99E-06	8.01E-10	4.15E-08
GTX_DN	14	REACTOME_PTM_GAMMA_CARBOXYLATION_HYPUSINE_FORMATION_AND_ARYLSULFATASE_ACTIVATION	1.26E-03	1.82E-02	3.80E-04	4.03E-05	1.54E-09	7.40E-08
GTX_DN	15	REACTOME_ACETYLCHOLINE_NEUROTRANSMITTER_RELEASE_CYCLE	3.00E-01	1.96E-04	1.63E-04	5.71E-05	2.28E-09	1.02E-07
GTX_DN	16	REACTOME_TIGHT_JUNCTION_INTERACTIONS	2.08E-01	1.07E-03	6.30E-04	4.96E-06	2.83E-09	1.19E-07
GTX_DN	17	REACTOME_PLATELET_HOMEOSTASIS	4.93E-01	2.55E-04	9.18E-05	8.45E-05	3.84E-09	1.52E-07
GTX_DN	18	REACTOME_GLUCURONIDATION	4.17E-02	2.67E-03	1.16E-04	8.73E-05	4.37E-09	1.57E-07
GTX_DN	19	REACTOME_INTERACTION_BETWEEN_L1_AND_ANKYRINS	3.38E-01	1.22E-03	1.08E-03	2.55E-06	4.42E-09	1.57E-07
GTX_DN	20	REACTOME_COMMON_PATHWAY	4.66E-04	4.06E-03	5.00E-04	1.50E-03	5.36E-09	1.79E-07
GTX_DN	21	REACTOME_PEPTIDE_HORMONE_BIOSYNTHESIS	1.78E-01	3.73E-03	1.19E-04	1.88E-05	5.59E-09	1.79E-07
GTX_DN	22	REACTOME_REGULATION_OF_PYRUVATE_DEHYDROGENASE_PDH_COMPLEX	1.09E-02	2.32E-04	2.71E-04	2.62E-03	6.63E-09	2.03E-07
GTX_DN	23	REACTOME_ELEVATION_OF_CYTOSOLIC_CA2_LEVELS	3.00E-01	7.06E-04	2.20E-04	6.25E-05	1.02E-08	2.98E-07

GTX_DN	24	REACTOME_ETHANOL_OXIDATION	5.74E-03	4.12E-03	8.88E-04	2.40E-04	1.66E-08	4.67E-07
GTX_DN	25	REACTOME_CELL_CELL_JUNCTION_ORGANIZATION	1.47E-01	4.08E-04	1.14E-03	9.12E-05	2.01E-08	5.43E-07
GTX_DN	26	REACTOME_N_GLYCAN_ANTENNAE_ELONGATION	2.36E-02	2.22E-04	1.27E-03	2.37E-03	4.58E-08	1.19E-06
GTX_DN	27	REACTOME_INTEGRATION_OF_PROVIRUS	2.05E-02	3.87E-02	2.38E-04	1.11E-04	5.86E-08	1.46E-06
GTX_DN	28	REACTOME_GOLGI_ASSOCIATED_VESICLE_BIOGENESIS	1.64E-01	2.15E-03	4.20E-04	1.50E-04	6.18E-08	1.49E-06
GTX_DN	29	REACTOME_UNBLOCKING_OF_NMDA_RECEPTOR_GLUTAMATE_BINDING_AND_ACTIVATION	4.28E-01	1.61E-03	1.23E-03	4.17E-05	9.33E-08	2.17E-06
GTX_DN	30	REACTOME_ACTIVATION_OF_RAC	7.94E-02	9.71E-04	1.18E-03	4.80E-04	1.12E-07	2.51E-06
GTX_DN	31	REACTOME_REGULATION_OF_INSULIN_SECRETION_BY_ACETYLCHOLINE	5.28E-02	7.85E-04	1.70E-03	1.26E-03	2.10E-07	4.56E-06
GTX_DN	32	REACTOME_REGULATION_OF_INSULIN_SECRETION	3.07E-01	1.00E-03	4.65E-04	7.55E-04	2.48E-07	5.07E-06
GTX_DN	33	REACTOME_SYNTHESIS_OF_PIPS_AT_THE_LATE_ENDOSOME_MEMBRANE	1.71E-01	8.88E-03	4.78E-04	1.47E-04	2.46E-07	5.07E-06
GTX_DN	34	REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES	3.46E-01	2.22E-03	1.59E-03	1.03E-04	2.83E-07	5.61E-06
GTX_DN	35	REACTOME_RAS_ACTIVATION_UOPN_CA2_INFUX_THROUGH_NMDA_RECEPTOR	1.58E-01	9.74E-04	3.53E-03	2.56E-04	3.09E-07	5.95E-06
GTX_DN	36	REACTOME_PROLACTIN_RECEPTOR_SIGNALING	7.93E-02	1.77E-03	6.03E-04	1.71E-03	3.22E-07	6.04E-06
GTX_DN	37	REACTOME_COPI_MEDIATED_TRANSPORT	7.78E-02	2.92E-04	1.47E-03	4.60E-03	3.38E-07	6.15E-06
GTX_DN	38	REACTOME_REGULATION_OF_INSULIN_LIKE_GROWTH_FACTOR_IGF_ACTIVITY_BY_INSULIN_LIKE_GROWTH_FACTOR_BINDING_PROTEINS_IGFBPS	6.41E-03	6.90E-03	1.64E-03	3.83E-03	5.68E-07	1.01E-05
GTX_DN	39	REACTOME_CD28_DEPENDENT_PI3K_AKT_SIGNALING	2.61E-02	2.11E-04	1.67E-02	3.51E-03	6.45E-07	1.08E-05
GTX_DN	40	REACTOME_OPSINS	3.00E-01	2.01E-03	6.62E-03	8.20E-05	6.55E-07	1.08E-05

Table S4 GSEA analysis of enrichment of Drugmatrix signatures in L1000 profiles

Drugmatrix_sample_type	signature_name	comparison	L1000_subset	direction_of_enrichment	direction_match	ES (Enrichment Score)	NES (Normalized Enrichment Score)	NOM.p.val	FDR.q.val
CELL_LOWDOSE	UP_CARC_CELL_LOWDOSE	carc	tas_0.4	pos	MATCH	2.97E-01	1.57E+00	3.59E-04	1.41E-03
CELL_LOWDOSE	DN_CARC_CELL_LOWDOSE	carc	tas_0.4	neg	MATCH	-2.20E-01	-1.08E+00	2.57E-01	3.62E-01
CELL_LOWDOSE	UP_CARC_CELL_LOWDOSE	carc	tas_0.3	pos	MATCH	3.02E-01	1.50E+00	6.53E-04	4.88E-03
CELL_LOWDOSE	DN_CARC_CELL_LOWDOSE	carc	tas_0.3	neg	MATCH	-2.35E-01	-1.15E+00	1.16E-01	1.88E-01
CELL_LOWDOSE	UP_CARC_CELL_LOWDOSE	carc	tas_0.2	pos	MATCH	2.99E-01	1.51E+00	0.00E+00	5.17E-03
CELL_LOWDOSE	DN_CARC_CELL_LOWDOSE	carc	tas_0.2	neg	MATCH	-2.68E-01	-1.29E+00	2.03E-02	3.58E-02
CELL_LOWDOSE	DN_CARC_CELL_LOWDOSE	carc	tas_0.0	neg	MATCH	-2.28E-01	-1.34E+00	6.67E-03	1.51E-02
CELL_LOWDOSE	UP_CARC_CELL_LOWDOSE	carc	tas_0.0	neg	NONMATCH	-2.08E-01	-1.23E+00	1.17E-02	4.31E-02
CELL_LOWDOSE	UP_GTX_CELL_LOWDOSE	gtx	tas_0.4	pos	MATCH	1.77E-01	1.28E+00	0.00E+00	2.99E-02
CELL_LOWDOSE	DN_GTX_CELL_LOWDOSE	gtx	tas_0.4	neg	MATCH	-2.43E-01	-9.34E-01	6.42E-01	6.64E-01
CELL_LOWDOSE	UP_GTX_CELL_LOWDOSE	gtx	tas_0.3	pos	MATCH	2.18E-01	1.42E+00	0.00E+00	3.84E-03
CELL_LOWDOSE	DN_GTX_CELL_LOWDOSE	gtx	tas_0.3	neg	MATCH	-3.19E-01	-1.30E+00	6.64E-02	6.86E-02
CELL_LOWDOSE	UP_GTX_CELL_LOWDOSE	gtx	tas_0.2	pos	MATCH	1.84E-01	1.19E+00	5.88E-02	7.84E-02
CELL_LOWDOSE	DN_GTX_CELL_LOWDOSE	gtx	tas_0.2	neg	MATCH	-2.67E-01	-1.08E+00	3.35E-01	3.21E-01
CELL_LOWDOSE	UP_GTX_CELL_LOWDOSE	gtx	tas_0.0	pos	MATCH	2.42E-01	1.37E+00	0.00E+00	1.11E-02
CELL_LOWDOSE	DN_GTX_CELL_LOWDOSE	gtx	tas_0.0	neg	MATCH	-2.05E-01	-8.44E-01	7.77E-01	8.02E-01
CELL	DN_CARC_CELL	carc	tas_0.4	neg	MATCH	-2.00E-01	-8.92E-01	7.09E-01	7.47E-01
CELL	UP_CARC_CELL	carc	tas_0.4	neg	NONMATCH	-3.59E-01	-1.63E+00	7.39E-04	1.33E-03
CELL	DN_CARC_CELL	carc	tas_0.3	neg	MATCH	-2.30E-01	-1.03E+00	3.89E-01	3.78E-01
CELL	UP_CARC_CELL	carc	tas_0.3	neg	NONMATCH	-3.58E-01	-1.63E+00	7.53E-04	7.50E-04
CELL	DN_CARC_CELL	carc	tas_0.2	pos	NONMATCH	2.36E-01	1.07E+00	2.80E-01	2.58E-01
CELL	UP_CARC_CELL	carc	tas_0.2	neg	NONMATCH	-4.03E-01	-1.81E+00	1.67E-04	8.77E-05
CELL	DN_CARC_CELL	carc	tas_0.0	pos	NONMATCH	2.53E-01	1.10E+00	2.93E-01	2.68E-01
CELL	UP_CARC_CELL	carc	tas_0.0	neg	NONMATCH	-2.85E-01	-1.56E+00	0.00E+00	3.49E-03
CELL	UP_GTX_CELL	gtx	tas_0.4	pos	MATCH	1.97E-01	1.29E+00	2.22E-02	5.97E-02
CELL	DN_GTX_CELL	gtx	tas_0.4	neg	MATCH	-3.66E-01	-1.48E+00	1.00E-04	4.06E-03
CELL	UP_GTX_CELL	gtx	tas_0.3	pos	MATCH	2.19E-01	1.30E+00	3.01E-02	3.33E-02
CELL	DN_GTX_CELL	gtx	tas_0.3	neg	MATCH	-4.30E-01	-1.85E+00	0.00E+00	0.00E+00
CELL	UP_GTX_CELL	gtx	tas_0.2	pos	MATCH	2.18E-01	1.29E+00	2.04E-02	5.78E-02
CELL	DN_GTX_CELL	gtx	tas_0.2	neg	MATCH	-3.79E-01	-1.63E+00	0.00E+00	7.10E-04
CELL	UP_GTX_CELL	gtx	tas_0.0	pos	MATCH	2.62E-01	1.39E+00	7.24E-03	1.26E-02

CELL	DN_GTX_CELL	gtx	tas_0.0	neg	MATCH	-3.10E-01	-1.39E+00	1.04E-02	2.98E-02
LIVER	UP_CARC_LIVER	carc	tas_0.4	pos	MATCH	2.34E-01	1.23E+00	4.59E-02	6.00E-02
LIVER	DN_CARC_LIVER	carc	tas_0.4	neg	MATCH	-3.17E-01	-1.48E+00	3.05E-03	6.62E-03
LIVER	UP_CARC_LIVER	carc	tas_0.3	pos	MATCH	2.44E-01	1.19E+00	8.08E-02	9.39E-02
LIVER	DN_CARC_LIVER	carc	tas_0.3	neg	MATCH	-3.68E-01	-1.72E+00	0.00E+00	4.17E-04
LIVER	UP_CARC_LIVER	carc	tas_0.2	pos	MATCH	2.39E-01	1.19E+00	7.24E-02	1.28E-01
LIVER	DN_CARC_LIVER	carc	tas_0.2	neg	MATCH	-4.14E-01	-1.91E+00	0.00E+00	0.00E+00
LIVER	DN_CARC_LIVER	carc	tas_0.0	neg	MATCH	-2.51E-01	-1.41E+00	1.03E-03	1.21E-02
LIVER	UP_CARC_LIVER	carc	tas_0.0	neg	NONMATCH	-2.38E-01	-1.39E+00	5.49E-03	1.08E-02
LIVER	DN_GTX_LIVER	gtx	tas_0.4	pos	NONMATCH	1.85E-01	1.20E+00	0.00E+00	2.99E-02
LIVER	UP_GTX_LIVER	gtx	tas_0.4	neg	NONMATCH	-2.60E-01	-1.04E+00	4.12E-01	6.13E-01
LIVER	DN_GTX_LIVER	gtx	tas_0.3	pos	NONMATCH	1.94E-01	1.16E+00	1.33E-01	1.11E-01
LIVER	UP_GTX_LIVER	gtx	tas_0.3	neg	NONMATCH	-3.03E-01	-1.28E+00	4.78E-02	5.27E-02
LIVER	DN_GTX_LIVER	gtx	tas_0.2	pos	NONMATCH	2.06E-01	1.23E+00	5.71E-02	7.10E-02
LIVER	UP_GTX_LIVER	gtx	tas_0.2	neg	NONMATCH	-2.99E-01	-1.26E+00	6.30E-02	1.04E-01
LIVER	DN_GTX_LIVER	gtx	tas_0.0	pos	NONMATCH	2.71E-01	1.42E+00	6.47E-03	1.30E-02
LIVER	UP_GTX_LIVER	gtx	tas_0.0	neg	NONMATCH	-3.95E-01	-1.72E+00	1.10E-04	3.90E-04

Table S5 Genesets of literature referenced AhR targets

Geneset Name	AhR geneset 1	AhR geneset 2	AhR geneset 3	AhR geneset 4
Description	omeprazole (AhR agonist) responsive gene signature from cryopreserved human hepatocytes	Cross-referenced AhR regulated genes (table 1)	Predicted functional partners of AhR in human from stringDB	Microarray identified TCDD (AhR agonist) responsive genes
Reference	Moscovitz et al., 2018	Beischlag et al., 2008	stringDB (Szklarczyk et al., 2017) with data referencing multiple sources	Lo and Matthews, 2012
Reference URL	http://jpet.aspetjournals.org/content/365/2/262/	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2583464/	https://string-db.org/cgi/network.pl?taskId=z6NWkfGxpKH3	https://www.ncbi.nlm.nih.gov/pubmed/22903824
Genes (overlap with gene symbols in L1000 gene expression)	CYP1A1	CYP1A1	HSP90AA1	CYP1A1
	CYP3A4	CYP1A2	ARNT	CYP1A2
	SLC10A1	CYP1B1	AIP	CYP1B1
	SLCO1B1	ALDH3A1	CYP1A1	ALDH1A3
		EREG	CYP1B1	TIPARP
		NQO1	CYP1A2	SLC7A5
		ALAS1	ESR1	SECTM1
		PTGS2	MAF	DLX2
		CDKN1B	NFE2L2	LRRC15
			RB1	RND1
				VIPR1
				EDC3
				ST3GAL1
				LMCD1
				PYGL
				RUNX2
				LMO2
				NEDD9
				VTCN1
				KRT20
				DISC1
				ALS2CL
				PRPS1
				MAPRE2
				LRRC23
				DDIT4
				SLC2A11

				VDR
				RUNX1
				PCBP3
				SLC3A2
				STC2
				CRISPLD2
				TPCN1
				BCL3
				DNMBP
				TXNRD1
				LAMA3
				TFAP2A
				SFT2D2
				FAM65A
				IER3
				MTMR6
				ATXN1
				CACNA1D
				ELF4
				TRIM13
				RRP12
				FLVCR2
				MYBBP1A
				ESPN
				NOP16
				SAT1
				BATF
				FOSL2
				RIN1
				ALDH1B1
				NAT10
				SLC27A2
				NFE2L2
				NIN
				PSPC1
				DDX21
				USP3

				NOP2
				UBE2G2
				DKC1
				IGF1R
				TRAFD1
				OSBPL2
				FAM32A
				XPO6
				ZNF259
				DDX24
				MICAL2
				NPY1R
				PSG5
				RET
				LRRFIP2
				TRIM36

Table S6 Comparison of genes and gene sets identified as differentially connected in the network-based approach (A) and by the standard differential expression analysis (B).

	A	B	A and B	only in A	only in B
genes	2293	2540	1224	1069	1316
gene sets	27	29	21	6	8

Table S7 Aggregate Network-related modules with connectivity specifically altered by compound groups.

Specificity score is computed only for HF module. GOC = Gain of connectivity, LOC = Loss of connectivity; HF= High Frequency, LF= Low Frequency

MAIN FUNCTION	ID	MODULE	SPECIFICITY SCORE	GAIN/LOSS	TYPE	ENRICHED HALLMARKS
SOLVENTS	G1	navajowhite2	0.25	GOC	HF	TNFA_SIGNALING_VIA_NFKB,

						IL6_JAK_STAT3_SIGNALING, INTERFERON_GAMMA_RESPONSE
		cyan	0.28	GOC	HF	EPITHELIAL_MESENCHYMAL_TRANSITION
ANTI-FUNGALS	G2	grey60	0.27	GOC	HF	ANGIOGENESIS
		orangered4	0.3	GOC	HF	CHOLESTEROL_HOMEOSTASIS
		skyblue3	0.26	GOC	HF	MYC_TARGETS_V1
STATINS	G3	thistle1	0.3	GOC	HF	FATTY_ACID_METABOLISM, CHOLESTEROL_HOMEOSTASIS
		honeydew1	0.28	GOC	HF	CHOLESTEROL_HOMEOSTASIS, MTORC1_SIGNALING
		coral1	NA	GOC	LF	TNFA_SIGNALING_VIA_NFKB, INFLAMMATORY_RESPONSE, UV_RESPONSE_UP
ESTROGENS	G4	palevioletred3	0.24	GOC	HF	INTERFERON_ALPHA_RESPONSE
		mediumorchid	NA	GOC	LF	MYC_TARGETS_V2
FIBRATES	G5	coral1	NA	GOC	LF	TNFA_SIGNALING_VIA_NFKB, HYPOXIA
STEROIDS	G6	lightcyan	0.26	GOC	HF	FATTY_ACID_METABOLISM
CHEMOTHERAPEUTICS	G9	palevioletred2	NA	GOC	LF	P53_PATHWAY
		thistle3	NA	GOC	LF	INTERFERON_GAMMA_RESPONSE, INTERFERON_ALPHA_RESPONSE
ALKYLATING-CANCER	G10	lightpink4	NA	GOC	LF	G2M_CHECKPOINT
n.c (anti-cancer, estrogens)	G11	coral1	NA	GOC	LF	TNFA_SIGNALING_VIA_NFKB
		tomato	NA	GOC	LF	MYC_TARGETS_V2
ANTI-INFLAMM/ FUNGAL	G12	coral1	NA	GOC	LF	APOPTOSIS
		lightslateblue	NA	GOC	LF	HYPOXIA
		lavenderblush2	NA	GOC	LF	MYC_TARGETS_V2, MYC_TARGETS_V1
		salmon2	NA	GOC	LF	HEME_METABOLISM

ANTISEPTICS / ESTROGENS	G13	skyblue3	0.26	GOC	HF	XENOBIOTIC_METABOLISM,
						MTORC1_SIGNALING
		coral1	NA	GOC	LF	TNFA_SIGNALING_VIA_NFKB,
						HYPOXIA,
						TGF_BETA_SIGNALING,
						P53_PATHWAY,
						APOPTOSIS

Table S8 Amplification-driven gene dependencies among cis genes in amplifications in TCGA Breast Cancer

Gene Symbol	Group	Coefficient (Slope)	Pvalue	FDR	Significance (FDR < 0.05)
ERBB2	Cis Rank 1	-0.1318508	3.4332E-42	5.6758E-38	TRUE
CCNE1	Cis Rank 1	-0.2001617	6.0405E-21	3.3287E-17	TRUE
CCND1	Cis Rank 1	-0.1272435	5.7819E-09	7.9655E-06	TRUE
FOXA1	Cis Rank 1	-0.1102331	1.1134E-07	0.00010226	TRUE
ANKRD17	Cis Rank 1	-0.1526054	2.7519E-06	0.00129983	TRUE
IRS2	Cis NonRank 1	-0.0527969	4.4622E-06	0.00204913	TRUE
MCL1	Cis Rank 1	-0.1394048	8.2763E-06	0.00333715	TRUE
IGF1R	Cis NonRank 1	-0.1016525	5.2538E-05	0.01400912	TRUE
LIG4	Cis NonRank 1	-0.038646	0.00015365	0.03097666	TRUE
KCTD21	Cis NonRank 1	-0.0398466	0.00050753	0.07425207	FALSE
ZNF784	Cis NonRank 1	-0.0560269	0.00146073	0.12609062	FALSE
TFDP1	Cis NonRank 1	-0.0592732	0.00537496	0.23634839	FALSE
ZNF703	Cis Rank 1	-0.0282214	0.00597179	0.24932848	FALSE
ZNF217	Cis Rank 1	-0.0409088	0.00780709	0.27931467	FALSE
OR2W3	Cis NonRank 1	-0.0429025	0.00834605	0.28566641	FALSE
ZNF580	Cis NonRank 1	-0.0564098	0.00914388	0.29492223	FALSE
KCNMB3	Cis NonRank 1	-0.0464382	0.01100085	0.31560188	FALSE
SDCCAG8	Cis NonRank 1	-0.059072	0.01302775	0.3407829	FALSE
AAMDC	Cis NonRank 1	-0.0371954	0.01310872	0.3418193	FALSE
TGDS	Cis NonRank 1	-0.0272693	0.01514159	0.36611654	FALSE
AURKC	Cis NonRank 1	-0.0337778	0.02033904	0.41563045	FALSE
ADPRHL1	Cis NonRank 1	-0.0347564	0.02246353	0.4343475	FALSE
ZNF460	Cis NonRank 1	-0.0495407	0.0273906	0.47662306	FALSE
SLC15A1	Cis NonRank 1	-0.0295258	0.02886242	0.48689136	FALSE

USP13	Cis NonRank 1	-0.0243926	0.0339077	0.52535421	FALSE
PPP6R1	Cis NonRank 1	-0.0492691	0.03430773	0.52613678	FALSE
ZNF274	Cis NonRank 1	-0.0130762	0.03443639	0.52614633	FALSE
ZMIZ1	Cis NonRank 1	-0.0360358	0.03463215	0.52768551	FALSE
COX18	Cis NonRank 1	-0.0529388	0.03755672	0.54511644	FALSE
UGGT2	Cis NonRank 1	-0.0254059	0.0465598	0.58668186	FALSE
ZNF581	Cis NonRank 1	-0.0356601	0.05952823	0.63864307	FALSE
RAP2A	Cis NonRank 1	-0.0179813	0.06904373	0.67660397	FALSE
PIK3CA	Cis NonRank 1	-0.0382066	0.07308259	0.68804182	FALSE
GOLPH3L	Cis NonRank 1	-0.0411693	0.07306642	0.68804182	FALSE
ZFP28	Cis NonRank 1	-0.0302816	0.08118507	0.70932517	FALSE
GRK1	Cis NonRank 1	-0.019318	0.08227314	0.71189798	FALSE
ZNF547	Cis NonRank 1	-0.0365368	0.08970384	0.7275414	FALSE
ZIC5	Cis NonRank 1	-0.0087923	0.09056707	0.72823675	FALSE
MYO7A	Cis NonRank 1	-0.019245	0.09573144	0.74142596	FALSE
ERLIN2	Cis NonRank 1	-0.0177344	0.10002168	0.75059392	FALSE
GPR180	Cis NonRank 1	-0.020376	0.10523636	0.76066616	FALSE
DOCK9	Cis NonRank 1	-0.0142928	0.10594672	0.7627615	FALSE
EFNB2	Cis NonRank 1	-0.0128848	0.10886666	0.77012566	FALSE
ERCC5	Cis NonRank 1	-0.0161536	0.11989628	0.78553788	FALSE
CLPTM1L	Cis NonRank 1	-0.0228713	0.12886525	0.80060138	FALSE
GAB2	Cis NonRank 1	-0.0137559	0.12947951	0.80321025	FALSE
LAMP1	Cis NonRank 1	-0.0168387	0.13092574	0.80568805	FALSE
ZNF550	Cis NonRank 1	-0.0300106	0.13525624	0.80803726	FALSE
SIRT5	Cis NonRank 1	-0.017817	0.13769317	0.81210972	FALSE
KIAA1549L	Cis NonRank 1	-0.0173794	0.13922227	0.81646775	FALSE
FRS2	Cis NonRank 1	-0.0191833	0.14175891	0.82201276	FALSE
RSF1	Cis Rank 1	-0.0140401	0.14580254	0.82774988	FALSE
NOTCH3	Cis Rank 1	-0.0310624	0.14787267	0.83073603	FALSE
TBL1XR1	Cis NonRank 1	-0.0073624	0.15076845	0.8323311	FALSE
ZNF579	Cis NonRank 1	-0.0226136	0.17174845	0.8657518	FALSE
CEP72	Cis NonRank 1	-0.0143598	0.17808883	0.87429438	FALSE
CASP14	Cis NonRank 1	-0.0165124	0.17935747	0.87544663	FALSE
EPN1	Cis NonRank 1	-0.0191911	0.18880276	0.88773813	FALSE
BRD9	Cis NonRank 1	-0.0126168	0.19704566	0.90237085	FALSE
VDAC3	Cis Rank 1	-0.0073304	0.21867352	0.9293645	FALSE
TCAP	Cis NonRank 1	-0.0057259	0.23178031	0.94193514	FALSE

ZNF551	Cis NonRank 1	-0.0122978	0.24341286	0.95655563	FALSE
PPP1R12C	Cis NonRank 1	-0.0130338	0.2462885	0.95792972	FALSE
KCTD14	Cis NonRank 1	-0.0093918	0.24753845	0.95792972	FALSE
ZNF814	Cis NonRank 1	-0.0109311	0.24757471	0.95792972	FALSE
OXGR1	Cis NonRank 1	-0.0092899	0.25523995	0.96360588	FALSE
PAK1	Cis NonRank 1	-0.0079182	0.25611266	0.96469101	FALSE
TRIM58	Cis NonRank 1	-0.0150117	0.26071762	0.9691522	FALSE
ZNF124	Cis NonRank 1	-0.0161519	0.2631421	0.9727034	FALSE
PPIF	Cis Rank 1	-0.015168	0.26886982	0.9755213	FALSE
SFTPA2	Cis NonRank 1	-0.0093825	0.26782012	0.9755213	FALSE
FBXO18	Cis NonRank 1	-0.0131004	0.28189291	0.98546279	FALSE
USP22	Cis Rank 1	-0.0032396	0.28351024	0.98735859	FALSE
PDCD6	Cis NonRank 1	-0.001922	0.4418571	1	FALSE
RANBP9	Cis NonRank 1	0.00725409	0.67576174	1	FALSE
HIPK3	Cis NonRank 1	0.00387192	0.57073055	1	FALSE
MBNL2	Cis NonRank 1	0.00619651	0.62847808	1	FALSE
ZNF256	Cis NonRank 1	0.00366045	0.56749171	1	FALSE
ABCC4	Cis NonRank 1	-4.381E-05	0.49854696	1	FALSE
NET1	Cis NonRank 1	-0.0017059	0.45180462	1	FALSE
TUBGCP3	Cis NonRank 1	0.36036427	1	1	FALSE
ZBTB18	Cis NonRank 1	0.00331216	0.58383905	1	FALSE
CCT2	Cis Rank 1	0.20523307	1	1	FALSE
SLC12A7	Cis NonRank 1	0.02398174	0.97398575	1	FALSE
POP4	Cis NonRank 1	0.01143103	0.69054661	1	FALSE
STARD3	Cis NonRank 1	0.02507874	0.99530024	1	FALSE
ILVBL	Cis NonRank 1	0.00652016	0.59546773	1	FALSE
CHML	Cis NonRank 1	-0.0086256	0.37707321	1	FALSE
RDH13	Cis NonRank 1	-0.0021493	0.44929504	1	FALSE
U2AF2	Cis NonRank 1	0.34284534	1	1	FALSE
COX20	Cis NonRank 1	-0.0007345	0.48711615	1	FALSE
CLNS1A	Cis NonRank 1	0.11039367	0.99998894	1	FALSE
TEX29	Cis NonRank 1	0.00916023	0.66953932	1	FALSE
ZNF543	Cis NonRank 1	0.01926515	0.817738	1	FALSE
ZNF787	Cis NonRank 1	0.03438221	0.91090574	1	FALSE
COL4A2	Cis NonRank 1	0.00374998	0.80869596	1	FALSE
FBXL14	Cis NonRank 1	0.00399172	0.59095499	1	FALSE
MIPOL1	Cis NonRank 1	0.02211109	0.87408151	1	FALSE

ZNF417	Cis NonRank 1	-0.0100134	0.31311064	1	FALSE
ZNF548	Cis NonRank 1	0.01220992	0.66355746	1	FALSE
ZNF524	Cis NonRank 1	-0.0117947	0.31393786	1	FALSE
CSTF3	Cis Rank 1	0.07041714	0.99749749	1	FALSE
ADSS	Cis NonRank 1	-0.004601	0.43117564	1	FALSE
CNST	Cis NonRank 1	0.01516517	0.81571513	1	FALSE
CLYBL	Cis NonRank 1	0.04911335	0.99869721	1	FALSE
NLRP7	Cis NonRank 1	0.04503415	0.89737595	1	FALSE
ENSA	Cis NonRank 1	0.01932006	0.86217819	1	FALSE
GDPD4	Cis NonRank 1	0.00872305	0.68971421	1	FALSE
ORAOV1	Cis NonRank 1	0.00671214	0.83792541	1	FALSE
FH	Cis NonRank 1	0.02158723	0.91443093	1	FALSE
DZIP1	Cis NonRank 1	0.01210949	0.76195028	1	FALSE
MYO16	Cis NonRank 1	0.00088472	0.52240798	1	FALSE
ERC1	Cis NonRank 1	0.00027291	0.50600706	1	FALSE
RPRD2	Cis NonRank 1	0.04653797	0.96617587	1	FALSE
ATP11A	Cis NonRank 1	0.01861572	0.91502289	1	FALSE
MCF2L	Cis NonRank 1	0.00112167	0.54951037	1	FALSE
OPN3	Cis NonRank 1	0.00258379	0.54858556	1	FALSE
ZIM2	Cis NonRank 1	0.01408318	0.71449807	1	FALSE
TMEM86B	Cis NonRank 1	0.0317536	0.8096276	1	FALSE
ZNF549	Cis NonRank 1	0.01614318	0.79541213	1	FALSE
AHCTF1	Cis NonRank 1	0.1165711	0.99993125	1	FALSE
NALCN	Cis NonRank 1	0.0361846	0.9918266	1	FALSE
GDI2	Cis NonRank 1	0.0264093	0.97024903	1	FALSE
PGBD2	Cis NonRank 1	-0.012215	0.31741354	1	FALSE
ZNF544	Cis NonRank 1	0.00014742	0.50701348	1	FALSE
UBE2S	Cis Rank 1	-0.0002232	0.4967775	1	FALSE
AQP11	Cis NonRank 1	-0.000388	0.48385538	1	FALSE
B4GALNT3	Cis NonRank 1	0.03205786	0.99741086	1	FALSE
CHAMP1	Cis NonRank 1	0.01999445	0.87090736	1	FALSE
ZIK1	Cis NonRank 1	0.00135627	0.52367868	1	FALSE
ZNF776	Cis NonRank 1	-0.0093528	0.31281467	1	FALSE
ANXA11	Cis NonRank 1	0.0183763	0.85595005	1	FALSE
HNRNPU	Cis NonRank 1	0.06397617	0.9448164	1	FALSE
UBAC2	Cis NonRank 1	0.0172858	0.94747788	1	FALSE
NLRP9	Cis NonRank 1	-0.0051715	0.40295852	1	FALSE

ZNF530	Cis NonRank 1	0.00146413	0.51582234	1	FALSE
IKBKB	Cis NonRank 1	0.00921423	0.81573412	1	FALSE
IL2RA	Cis NonRank 1	0.01098336	0.79837212	1	FALSE
IL15RA	Cis NonRank 1	-0.0094688	0.32167013	1	FALSE
ING1	Cis NonRank 1	0.00651614	0.69047774	1	FALSE
ZNF773	Cis NonRank 1	0.00178368	0.52766934	1	FALSE
IPO5	Cis NonRank 1	0.04046899	0.9675687	1	FALSE
ZNF749	Cis NonRank 1	-0.0057844	0.3874731	1	FALSE
ZNF805	Cis NonRank 1	0.02633724	0.91054938	1	FALSE
MYC	Cis Rank 1	0.00310355	0.62264384	1	FALSE
NDUFB5	Cis NonRank 1	0.07522992	0.99967947	1	FALSE
NDUFC2	Cis NonRank 1	0.00719401	0.77825697	1	FALSE
NDUFS6	Cis NonRank 1	0.03548807	0.99020773	1	FALSE
CCDC73	Cis NonRank 1	-0.0155654	0.31633398	1	FALSE
ATP4B	Cis NonRank 1	0.00891931	0.68066209	1	FALSE
PCCA	Cis NonRank 1	0.00156008	0.59766705	1	FALSE
SCCPDH	Cis NonRank 1	0.00055506	0.51082471	1	FALSE
TUBD1	Cis Rank 1	-0.0068403	0.35081598	1	FALSE
ZNF639	Cis NonRank 1	0.02179233	0.81436787	1	FALSE
TBC1D7	Cis Rank 1	-0.0004486	0.48874937	1	FALSE
NOL7	Cis NonRank 1	0.25806095	0.99999999	1	FALSE
PEG3	Cis NonRank 1	0.02845491	0.88613651	1	FALSE
CALML5	Cis NonRank 1	0.02939626	0.98692185	1	FALSE
PNMT	Cis NonRank 1	0.00473866	0.70003615	1	FALSE
POLB	Cis NonRank 1	0.01352516	0.92393409	1	FALSE
GFOD1	Cis NonRank 1	0.05242106	0.99847828	1	FALSE
ANKRD16	Cis NonRank 1	0.00452439	0.59629394	1	FALSE
BIVM	Cis NonRank 1	-0.0021457	0.45076616	1	FALSE
TMCO3	Cis NonRank 1	-0.0005413	0.48963553	1	FALSE
ARGLU1	Cis NonRank 1	0.17135131	1	1	FALSE
KIF26B	Cis NonRank 1	0.03437522	0.94586509	1	FALSE
DCUN1D2	Cis NonRank 1	0.01922781	0.85136991	1	FALSE
ZNF444	Cis NonRank 1	0.00128708	0.52400767	1	FALSE
TCP11L1	Cis NonRank 1	-0.0041673	0.46740414	1	FALSE
ANKRD10	Cis NonRank 1	0.01242958	0.7481094	1	FALSE
RAB20	Cis NonRank 1	0.00513508	0.64823167	1	FALSE
NLRP2	Cis NonRank 1	0.04543463	0.91893445	1	FALSE

ZNF692	Cis NonRank 1	0.0366747	0.99512297	1	FALSE
ZNF416	Cis NonRank 1	0.00535888	0.61706086	1	FALSE
MFN1	Cis NonRank 1	0.01586675	0.75753924	1	FALSE
DNAJC3	Cis NonRank 1	-0.0037168	0.37456116	1	FALSE
NAT14	Cis NonRank 1	0.05151886	0.88202146	1	FALSE
ZNF695	Cis NonRank 1	0.00876887	0.66611578	1	FALSE
MRPL47	Cis NonRank 1	0.03804039	0.94524677	1	FALSE
VN1R1	Cis NonRank 1	0.00329472	0.54253805	1	FALSE
ZNF304	Cis NonRank 1	0.00990978	0.63604804	1	FALSE
USP35	Cis NonRank 1	-0.0103196	0.30421358	1	FALSE
ZNF71	Cis NonRank 1	0.00090043	0.5140303	1	FALSE
RAD52	Cis NonRank 1	0.02801393	0.87351326	1	FALSE
KDM5A	Cis NonRank 1	0.01293002	0.84845963	1	FALSE
RPL28	Cis NonRank 1	0.22930621	1	1	FALSE
SDHA	Cis NonRank 1	0.02519182	0.98700674	1	FALSE
MCUR1	Cis NonRank 1	-0.0075276	0.33206972	1	FALSE
ZNF667	Cis NonRank 1	-0.0016475	0.47357935	1	FALSE
TFB2M	Cis NonRank 1	-0.0017956	0.45985776	1	FALSE
CCDC168	Cis NonRank 1	0.02531129	0.84279923	1	FALSE
MRPS25	Cis Rank 1	0.02503258	0.86001733	1	FALSE
SMYD3	Cis NonRank 1	-0.0008826	0.42640566	1	FALSE
MRPL36	Cis NonRank 1	0.00297571	0.5616329	1	FALSE
UPF3A	Cis NonRank 1	0.00760859	0.721719	1	FALSE
WNK1	Cis NonRank 1	0.02791257	0.88906637	1	FALSE
SLC6A12	Cis NonRank 1	0.01633251	0.8664876	1	FALSE
SLC9A3	Cis NonRank 1	0.01403311	0.86322223	1	FALSE
TERT	Cis NonRank 1	0.03769704	0.99431035	1	FALSE
TPP2	Cis NonRank 1	0.01766255	0.89700012	1	FALSE
SDHAP3	Cis NonRank 1	0.00023022	0.52212488	1	FALSE
ZIC2	Cis NonRank 1	0.0032614	0.71358973	1	FALSE
ZNF134	Cis NonRank 1	0.01797694	0.73679406	1	FALSE
ALG8	Cis NonRank 1	0.02118257	0.96126484	1	FALSE
PRRG4	Cis NonRank 1	0.02034037	0.88747519	1	FALSE
KDELC1	Cis NonRank 1	-0.0060086	0.3477355	1	FALSE
ZSCAN5A	Cis NonRank 1	-0.0041315	0.3214184	1	FALSE
PLEKHF1	Cis NonRank 1	0.02221683	0.95075949	1	FALSE
CARS2	Cis NonRank 1	0.07344072	0.99689601	1	FALSE

ADIPOR2	Cis NonRank 1	0.00748951	0.69238424	1	FALSE
ZNF419	Cis NonRank 1	0.01949632	0.81605689	1	FALSE
ISOC2	Cis NonRank 1	0.01045267	0.63652305	1	FALSE
GRTP1	Cis NonRank 1	-0.0012897	0.46418954	1	FALSE
QSER1	Cis NonRank 1	0.03740575	0.95679269	1	FALSE
TUBAL3	Cis NonRank 1	-0.0084994	0.35463976	1	FALSE
ZNF669	Cis NonRank 1	0.00301203	0.5500145	1	FALSE
LPCAT1	Cis NonRank 1	0.0101044	0.96678242	1	FALSE
ZNF672	Cis NonRank 1	0.04377144	0.99240955	1	FALSE
ZNF606	Cis NonRank 1	-0.0076844	0.36851278	1	FALSE
TMEM254	Cis NonRank 1	0.02167449	0.87358835	1	FALSE
TARS2	Cis NonRank 1	0.05377597	0.9837105	1	FALSE
SH3BP5L	Cis NonRank 1	-0.003006	0.44156323	1	FALSE
YEATS4	Cis NonRank 1	0.03673502	0.9980863	1	FALSE
WNT5B	Cis NonRank 1	0.00256322	0.56125752	1	FALSE
HORMAD1	Cis NonRank 1	0.01499117	0.85847843	1	FALSE
PPP1R1B	Cis NonRank 1	-0.0012583	0.44251376	1	FALSE
STK24	Cis NonRank 1	0.00103379	0.57550395	1	FALSE
EFCAB2	Cis NonRank 1	-0.0070778	0.33680328	1	FALSE
CCDC77	Cis Rank 1	0.00957246	0.66797495	1	FALSE
CUL4A	Cis NonRank 1	0.01362829	0.83588589	1	FALSE
ZNF496	Cis NonRank 1	-0.0062141	0.37298522	1	FALSE
TMTC4	Cis NonRank 1	0.03332745	0.96368663	1	FALSE
ZNF587	Cis NonRank 1	0.00317171	0.57310265	1	FALSE
FIZ1	Cis NonRank 1	-0.0106968	0.33561027	1	FALSE
ABHD13	Cis NonRank 1	0.00160335	0.55362007	1	FALSE
RBM17	Cis Rank 1	0.16532471	1	1	FALSE
KMO	Cis NonRank 1	0.00628078	0.63366937	1	FALSE
ACTL6A	Cis Rank 1	0.33688424	1	1	FALSE
URI1	Cis NonRank 1	0.22177821	1	1	FALSE
ARHGEF7	Cis NonRank 1	0.02126154	0.96254141	1	FALSE
CDC16	Cis NonRank 1	0.12764563	0.99999545	1	FALSE
ZNF628	Cis NonRank 1	0.01167726	0.66146549	1	FALSE
CLDN10	Cis NonRank 1	0.00971491	0.85467553	1	FALSE
EXO1	Cis NonRank 1	-0.0083666	0.34351086	1	FALSE
INTS4	Cis NonRank 1	0.22039359	1	1	FALSE
TEX30	Cis Rank 1	0.04007422	0.97387295	1	FALSE

TRIP13	Cis Rank 1	0.09562849	1	1	FALSE
PGAP3	Cis NonRank 1	-0.0003154	0.48669291	1	FALSE
ZNF670	Cis NonRank 1	0.06718332	0.99765413	1	FALSE
TM9SF2	Cis NonRank 1	-0.001623	0.44830685	1	FALSE
ZNF264	Cis NonRank 1	0.05007457	0.96943805	1	FALSE
CD59	Cis NonRank 1	0.00826601	0.63027026	1	FALSE
CEP170	Cis NonRank 1	0.00231814	0.54710312	1	FALSE

Table S9 Summary of TCGA datasets used in iEDGE pancancer analysis

Cancer Type Abbreviation	Cancer Name	Number of Samples (with SCNA & GEP)	Number of SCNAs (GISTIC2.0)	Number of Amplifications	Number of Deletions
ACC	Adrenocortical carcinoma	77	44	18	26
BLCA	Bladder Urothelial Carcinoma	404	73	37	36
BRCA	Breast invasive carcinoma	1075	70	28	42
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	292	63	26	37
COAD	Colon adenocarcinoma	447	66	22	44
ESCA	Esophageal carcinoma	183	82	30	52
GBM	Glioblastoma multiforme	146	68	24	44
HNSC	Head and Neck squamous cell carcinoma	514	76	28	48
KIRC	Kidney renal clear cell carcinoma	525	29	10	19
KIRP	Kidney renal papillary cell carcinoma	256	28	7	21
LIHC	Liver hepatocellular carcinoma	364	61	27	34
LUAD	Lung adenocarcinoma	512	75	29	46
LUSC	Lung squamous cell carcinoma	498	83	30	53
OV	Ovarian serous cystadenocarcinoma	300	73	33	40

PAAD	Pancreatic adenocarcinoma	177	56	23	33
PRAD	Prostate adenocarcinoma	491	63	28	35
READ	Rectum adenocarcinoma	164	58	22	36
THCA	Thyroid carcinoma	495	41	9	32
UCEC	Uterine Corpus Endometrial Carcinoma	537	101	50	51

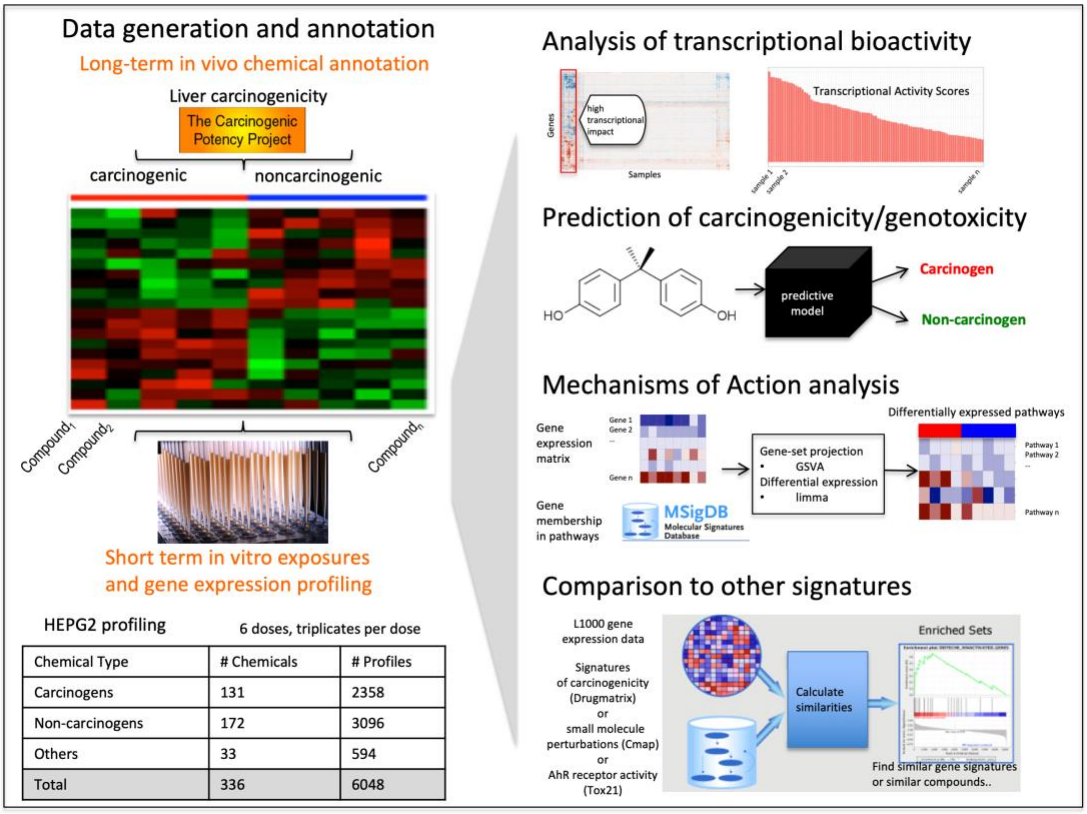


Figure S2.1.1 **Overview of Experimental Design and Analysis Aims**

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 (Cmap) and AhR Receptor activity (Tox21).

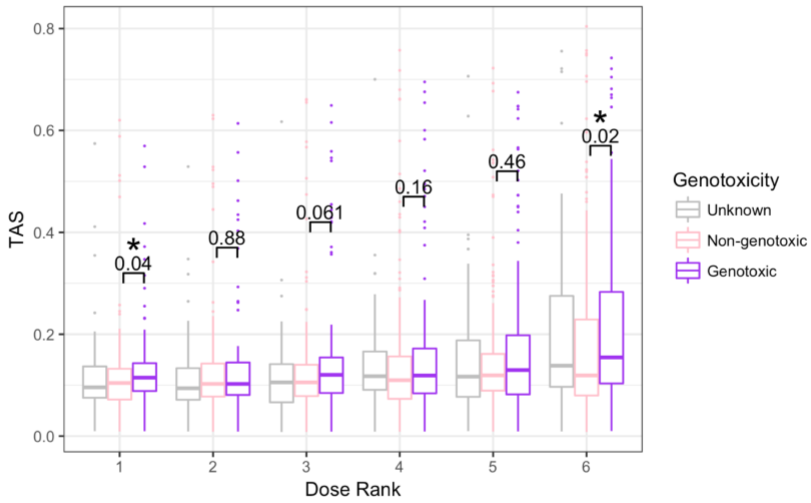


Figure S2.1.2 **Distribution of TAS grouped by chemical genotoxicity within each dose level**

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.

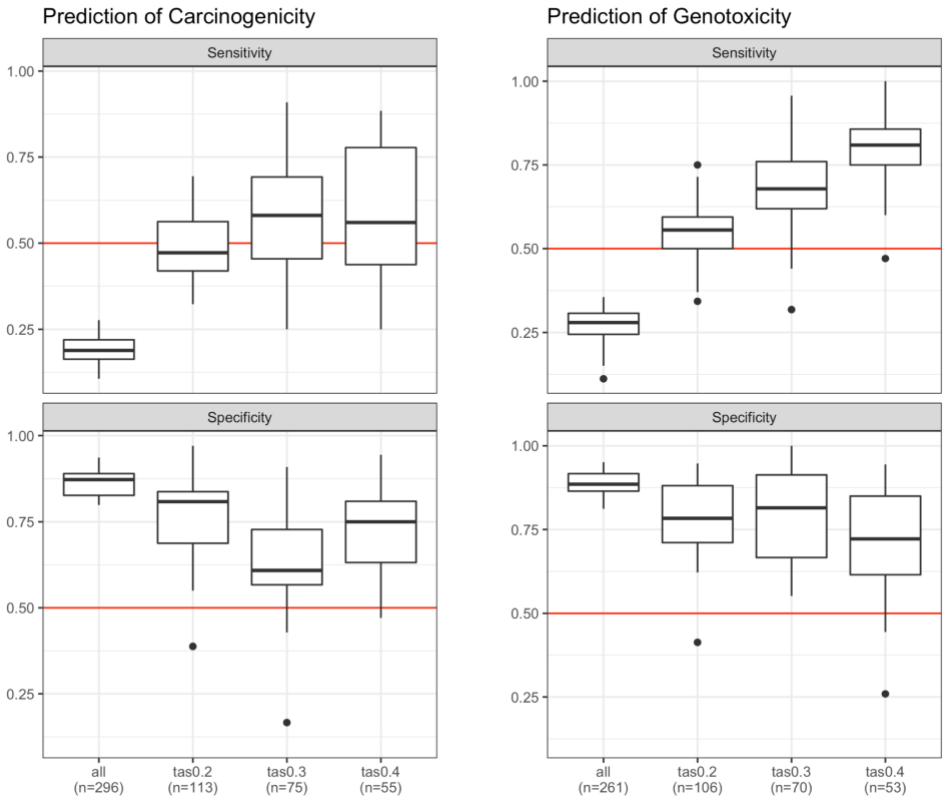


Figure S2.1.3 **Sensitivity and specificity rates of classifiers at threshold of 0.3 in predictive models of carcinogenicity and genotoxicity**

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.

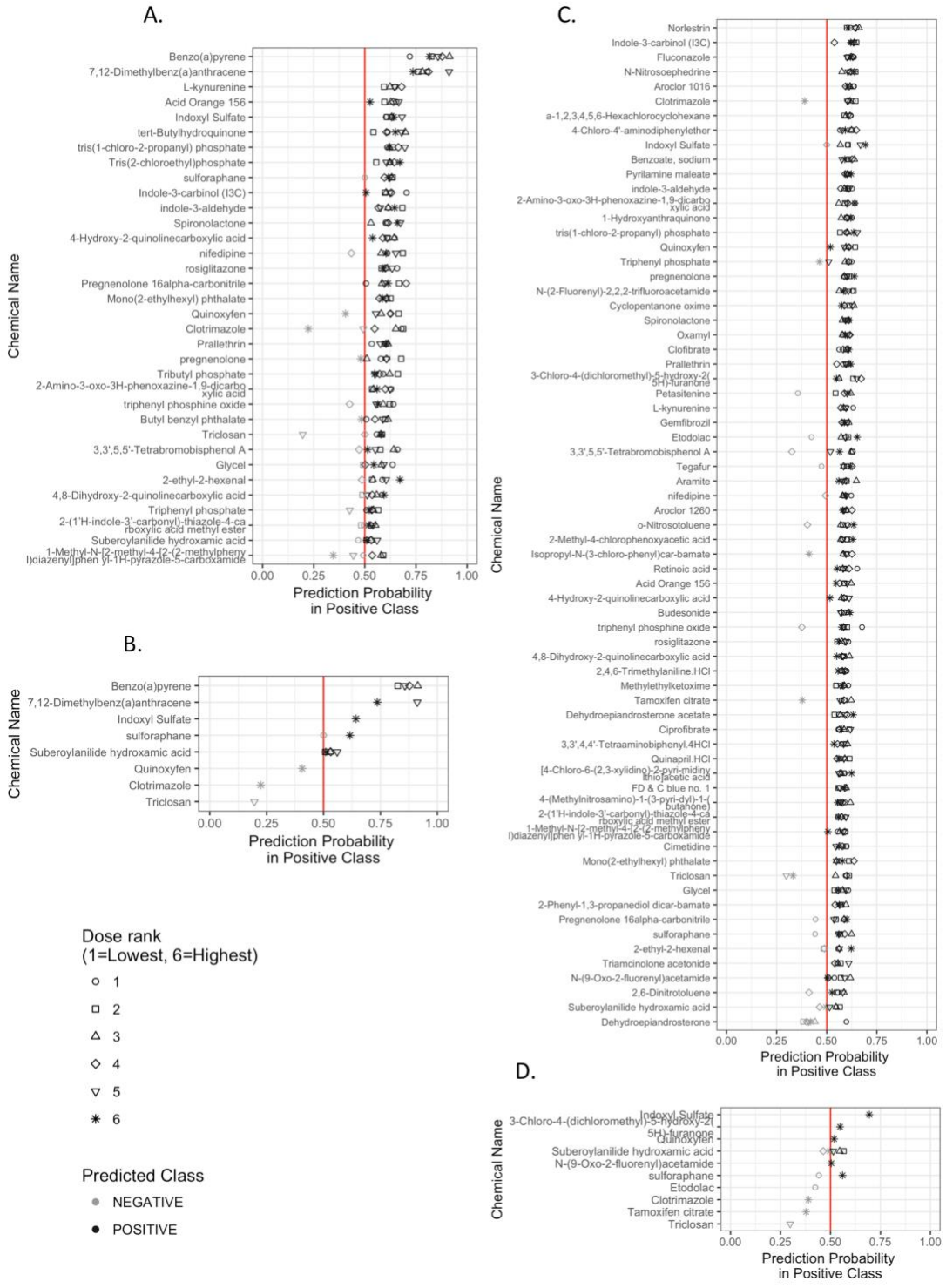


Figure S2.1.4 **Prediction probabilities on unlabeled chemicals**

A. prediction of carcinogenicity in all unlabeled profiles, **B.** prediction of carcinogenicity in unlabeled profiles with TAS > 0.4, **C.** prediction of genotoxicity in all unlabeled profiles **D.** prediction of genotoxicity in unlabeled profiles with TAS > 0.4

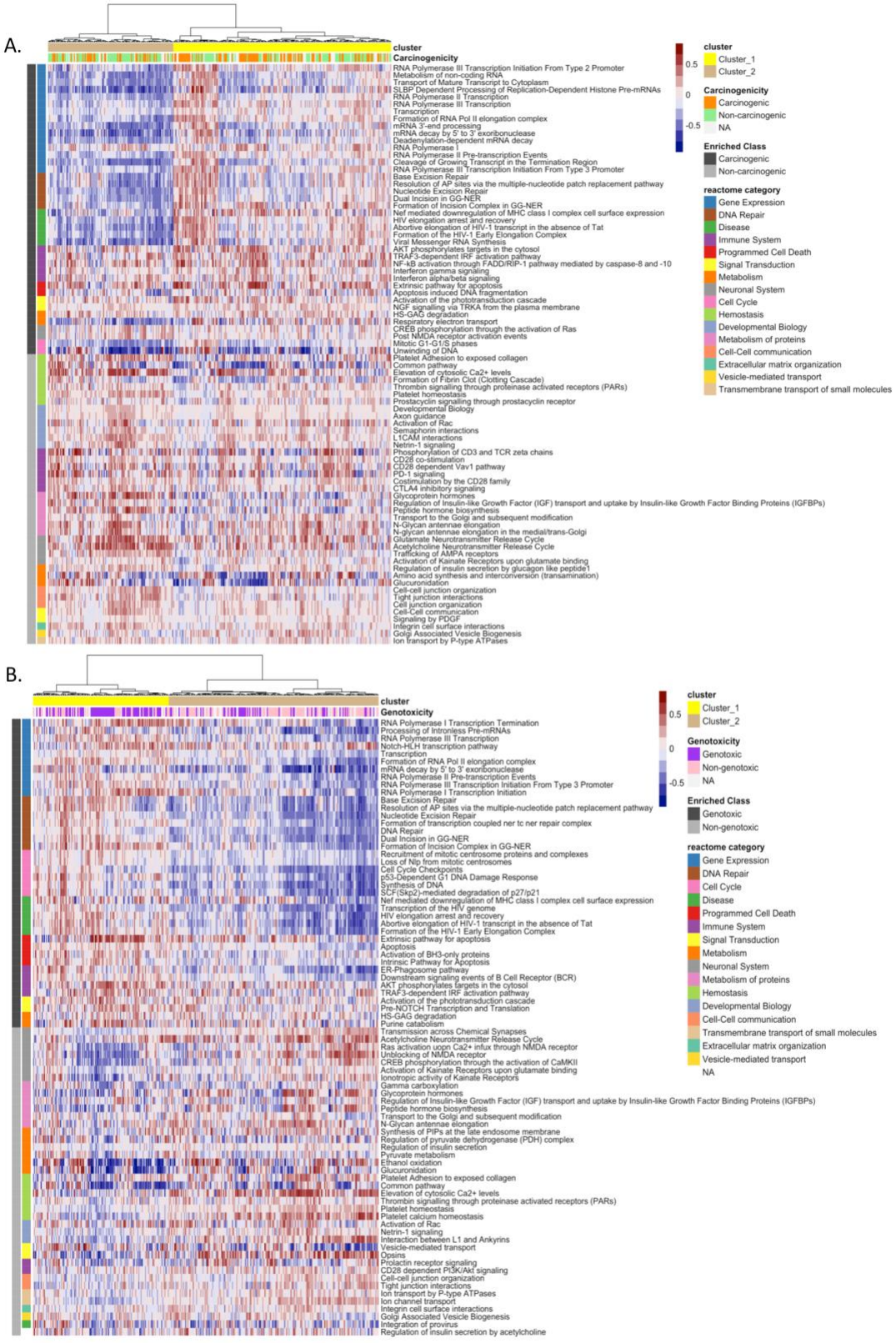

A.
cluster
Carcinogenicity
RNA Polymerase III Transcription Initiation From Type 2 Promoter
Metabolism of non-coding RNA
Transport of Mature Transcript to Cytoplasm
SLBP Dependent Processing of Replication-Dependent Histone Pre-mRNAs
RNA Polymerase II Transcription
RNA Polymerase III Transcription
Transcription
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+ influx 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 S2.1.5 **Heatmap of pathway enrichment scores (GSVA) for top 40 upregulated and downregulated differential pathways**

Differential pathways of **A.** carcinogenicity and **B.** genotoxicity 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 S3.1

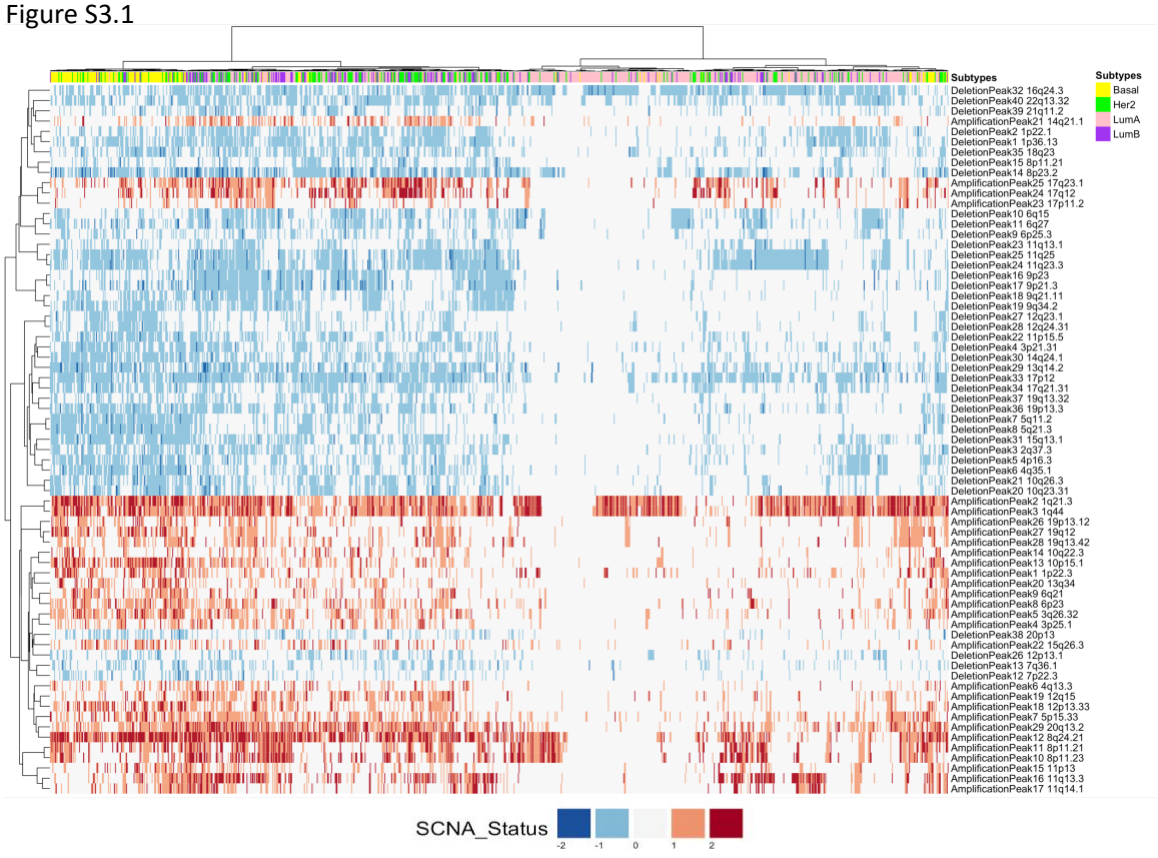


Figure S3.1 **Somatic copy number alteration status (SCNA) across subtyped TCGA breast cancer samples**

Figure S3.2

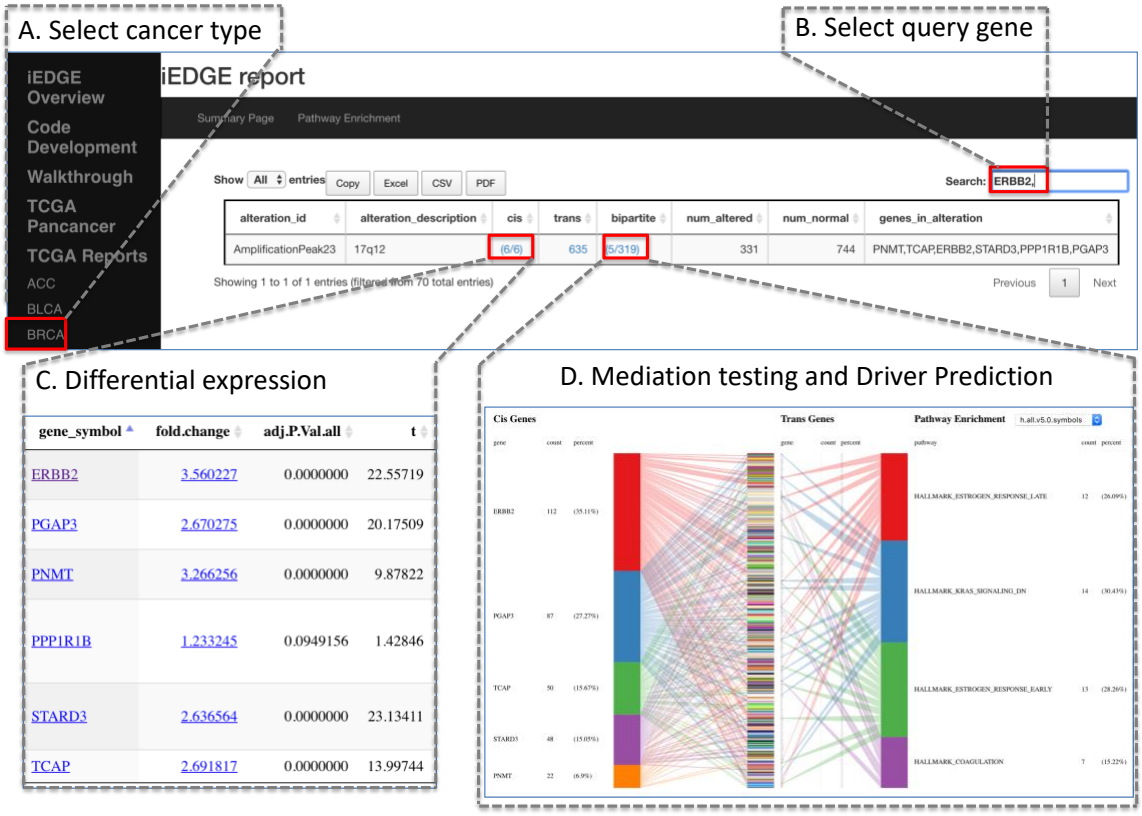


***Figure S3.2* iEDGE portal overview**

A. Selection of iEDGE report by cancer type
B. Selection of query gene
C. Differential expression table report for cis genes
D. Graphical report of mediation testing and driver prediction

BIBLIOGRAPHY

Abdo, K. M., Eustis, S. L., Haseman, J., Huff, J. E., Peters, A., & Persing, R. (1988). Toxicity and carcinogenicity of rotenone given in the feed to F344/N rats and B6C3F1 mice for up to two years. *Drug and Chemical Toxicology*, *11*(3), 225–235. https://doi.org/10.3109/01480548809017879

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., … Pe'er, D. (2010). An Integrated Approach to Uncover Drivers of Cancer. *Cell*, *143*(6), 1005–1017. https://doi.org/10.1016/j.cell.2010.11.013

American Cancer Society. (n.d.-a). Asbestos and Cancer Risk. Retrieved November 27, 2018, from https://www.cancer.org/cancer/cancer-causes/asbestos.html

American Cancer Society. (n.d.-b). Cancer Facts & Figures 2017. Retrieved December 3, 2018, from https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html

American Lung Association. (n.d.). Lung Cancer Fact Sheet | American Lung Association. Retrieved November 27, 2018, from https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html

Amgalan, B., & Lee, H. (2015). DEOD: uncovering dominant effects of cancer-driver genes based on a partial covariance selection method. *Bioinformatics*, *31*(15), 2452–2460. https://doi.org/10.1093/bioinformatics/btv175

Anand, P., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., … Aggarwal, B. B. (2008). Cancer is a Preventable Disease that Requires Major Lifestyle Changes. *Pharmaceutical Research*, *25*(9), 2097–2116. https://doi.org/10.1007/s11095-008-9661-9

Aviram, M., Rosenblat, M., Bisgaier, C. L., & Newton, R. S. (1998). Atorvastatin and gemfibrozil metabolites, but not the parent drugs, are potent antioxidants against lipoprotein oxidation. *Atherosclerosis*, *138*(2), 271–280. https://doi.org/10.1016/S0021-9150(98)00032-X

Barabási, A.-L. (2009). Scale-Free Networks: A Decade and Beyond. *Science*, *325*(5939), 412–413. https://doi.org/10.1126/science.1173299

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–607. https://doi.org/10.1038/nature11003

Bavetsias, V., & Linardopoulos, S. (2015). Aurora Kinase Inhibitors: Current Status and Outlook. *Frontiers in Oncology*, *5*, 278. https://doi.org/10.3389/fonc.2015.00278

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Broad Institute TCGA Genome Data Analysis Center. (2015). *SNP6 Copy number analysis (GISTIC2)*. Broad Institute of MIT and Harvard. https://doi.org/10.7908/C1Z0379T

Bucher, J. R., & Portier, C. (2004). Human carcinogenic risk evaluation, Part V: The national toxicology program vision for assessing the human carcinogenic hazard of chemicals. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, *82*(2), 363–366. https://doi.org/10.1093/toxsci/kfh293

Butte, A. J., & Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 418–429.

Campbell, K. J., Dhayade, S., Ferrari, N., Sims, A. H., Johnson, E., Mason, S. M., … Blyth, K. (2018). MCL-1 is a prognostic indicator and drug target in breast cancer. *Cell Death & Disease*, *9*(2), 19. https://doi.org/10.1038/s41419-017-0035-2

Cancer Cell Line Encyclopedia Consortium, & Genomics of Drug Sensitivity in Cancer Consortium. (2015). Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, *528*(7580), 84–87. https://doi.org/10.1038/nature15736

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., … Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Carter, S. L., Brechbühler, C. M., Griffin, M., & Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics (Oxford, England)*, *20*(14), 2242–2250. https://doi.org/10.1093/bioinformatics/bth234

Center for Disease Control. (2016, January 1). CDC Press Releases. Retrieved November 27, 2018, from https://www.cdc.gov/media/releases/2016/p1110-vital-signs-cancer-tobacco.html

Chapuy, B., Stewart, C., Dunford, A. J., Kim, J., Kamburov, A., Redd, R. A., … Shipp, M. A. (2018). Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nature Medicine*, *24*(5), 679–690. https://doi.org/10.1038/s41591-018-0016-8

Cosmic. (n.d.). Cancer Gene Census. Retrieved December 11, 2018, from http://cancer.sanger.ac.uk/census

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., … D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, *42*(Database issue), D472-477. https://doi.org/10.1093/nar/gkt1102

Davidson, M. D., Ware, B. R., & Khetani, S. R. (2015). Stem Cell-Derived Liver Cells for Drug Testing and Disease Modeling. *Discovery Medicine*, *19*(106), 349–358.

Davis, A. P., Grondin, C. J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., … Mattingly, C. J. (2015). The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Research*, *43*(Database issue), D914-920. https://doi.org/10.1093/nar/gku935

Davis, J. C., Furstenthal, L., Desai, A. A., Norris, T., Sutaria, S., Fleming, E., & Ma, P. (2009). The microeconomics of personalized medicine: today's challenge and tomorrow's promise. *Nature Reviews Drug Discovery*, *8*(4), 279–286. https://doi.org/10.1038/nrd2825

Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., & Elledge, S. J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, *155*(4), 948–962. https://doi.org/10.1016/j.cell.2013.10.011

Deng, M., Li, F., Ballif, B. A., Li, S., Chen, X., Guo, L., & Ye, X. (2009). Identification and Functional Analysis of a Novel Cyclin E/Cdk2 Substrate Ankrd17. *The Journal of Biological Chemistry*, *284*(12), 7875–7888. https://doi.org/10.1074/jbc.M807827200

Ding, W., Levy, D. D., Bishop, M. E., Pearce, M. G., Davis, K. J., Jeffrey, A. M., … Manjanatha, M. G. (2015). In vivo genotoxicity of estragole in male F344 rats. *Environmental and Molecular Mutagenesis*, *56*(4), 356–365. https://doi.org/10.1002/em.21918

Dong, L., Ding, H., Li, Y., Xue, D., Li, Z., Liu, Y., … Wang, P. (2019). TRIP13 is a predictor for poor prognosis and regulates cell proliferation, migration and invasion in prostate cancer. *International Journal of Biological Macromolecules*, *121*, 200–206. https://doi.org/10.1016/j.ijbiomac.2018.09.168

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics (Oxford, England)*, *21*(16), 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Eichner, J., Kossler, N., Wrzodek, C., Kalkuhl, A., Bach Toft, D., Ostenfeldt, N., … Zell, A. (2013). A toxicogenomic approach for the prediction of murine hepatocarcinogenesis using ensemble feature selection. *PloS One*, *8*(9), e73938. https://doi.org/10.1371/journal.pone.0073938

Ellinger-Ziegelbauer, H., Gmuender, H., Bandenburg, A., & Ahr, H. J. (2008). Prediction of a carcinogenic potential of rat hepatocarcinogens using toxicogenomics analysis of short-term in vivo studies. *Mutation Research*, *637*(1–2), 23–39. https://doi.org/10.1016/j.mrfmmm.2007.06.010

Fábián, Á., Vereb, G., & Szöllősi, J. (2013). The hitchhikers guide to cancer stem cell theory: markers, pathways and therapy. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology*, *83*(1), 62–71. https://doi.org/10.1002/cyto.a.22206

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., … D'Eustachio, P. (2016). The Reactome pathway Knowledgebase. *Nucleic Acids Research*, *44*(D1), D481-487. https://doi.org/10.1093/nar/gkv1351

Fielden, M. R., Brennan, R., & Gollub, J. (2007). A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by nongenotoxic chemicals. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, *99*(1), 90–100. https://doi.org/10.1093/toxsci/kfm156

Fitzpatrick, R. B. (2008). CPDB: Carcinogenic Potency Database. *Medical Reference Services Quarterly*, *27*(3), 303–311. https://doi.org/10.1080/02763860802198895

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., … Campbell, P. J. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, *45*(D1), D777–D783. https://doi.org/10.1093/nar/gkw1121

Ganter, B., Snyder, R. D., Halbert, D. N., & Lee, M. D. (2006). Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database. *Pharmacogenomics*, *7*(7), 1025–1044. https://doi.org/10.2217/14622416.7.7.1025

Ganter, B., Tugendreich, S., Pearson, C. I., Ayanoglu, E., Baumhueter, S., Bostian, K. A., … Jarnagin, K. (2005). Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *Journal of Biotechnology*, *119*(3), 219–244. https://doi.org/10.1016/j.jbiotec.2005.03.022

Gold, L. S., Manley, N. B., Slone, T. H., Rohrbach, L., & Garfinkel, G. B. (2005). Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature through 1997 and by the National Toxicology Program in 1997-1998. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, *85*(2), 747–808. https://doi.org/10.1093/toxsci/kfi161

Gupta, M., Mazumdar, U. K., Sivakumar, T., Vamsi, M. L. M., Karki, S. S., Sambathkumar, R., & Manikandan, L. (2003). Evaluation of Anti-inflammatory Activity of Chloroform Extract of Bryonia laciniosa in Experimental Animal Models. *Biological and Pharmaceutical Bulletin*, *26*(9), 1342–1344. https://doi.org/10.1248/bpb.26.1342

Gusenleitner, D., Auerbach, S. S., Melia, T., Gómez, H. F., Sherr, D. H., & Monti, S. (2014). Genomic models of short-term exposure accurately predict long-term chemical carcinogenicity and identify putative mechanisms of action. *PloS One*, *9*(7), e102579. https://doi.org/10.1371/journal.pone.0102579

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, *33*(suppl_1), D514–D517. https://doi.org/10.1093/nar/gki033

Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, *14*(1), 7. https://doi.org/10.1186/1471-2105-14-7

Haworth, C. M. A., Dale, P., & Plomin, R. (2008). A Twin Study into the Genetic and Environmental Influences on Academic Performance in Science in nine-year-old Boys and Girls. *International Journal of Science Education*, *30*(8), 1003. https://doi.org/10.1080/09500690701324190

Huff, J., Jacobson, M. F., & Davis, D. L. (2008). The Limits of Two-Year Bioassay Exposure Regimens for Identifying Chemical Carcinogens. *Environmental Health Perspectives*, *116*(11), 1439–1442. https://doi.org/10.1289/ehp.10716

Irigaray, P., Newby, J. A., Clapp, R., Hardell, L., Howard, V., Montagnier, L., … Belpomme, D. (2007). Lifestyle-related factors and environmental agents causing

cancer: an overview. *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie*, *61*(10), 640–658. https://doi.org/10.1016/j.biopha.2007.10.006

Jamdade, V. S., Mundhe, N. A., Kumar, P., Tadla, V., & Lahkar, M. (2016). Raloxifene Inhibits NF-kB Pathway and Potentiates Anti-Tumour Activity of Cisplatin with Simultaneous Reduction in its Nephrotoxictiy. *Pathology & Oncology Research*, *22*(1), 145–153. https://doi.org/10.1007/s12253-015-9988-6

Jennen, D. G. J., Magkoufopoulou, C., Ketelslegers, H. B., van Herwijnen, M. H. M., Kleinjans, J. C. S., & van Delft, J. H. M. (2010). Comparison of HepG2 and HepaRG by Whole-Genome Gene Expression Analysis for the Purpose of Chemical Hazard Identification. *Toxicological Sciences*, *115*(1), 66–79. https://doi.org/10.1093/toxsci/kfq026

Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., … Dix, D. J. (2010). In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environmental Health Perspectives*, *118*(4), 485–492. https://doi.org/10.1289/ehp.0901392

Kastan, M. B., & Bartek, J. (2004). Cell-cycle checkpoints and cancer. *Nature*, *432*, 316–323. https://doi.org/10.1038/nature03097

Kim, J. W., Botvinnik, O. B., Abudayyeh, O., Birger, C., Rosenbluh, J., Shrestha, Y., … Tamayo, P. (2016). Characterizing genomic alterations in cancer by complementary functional associations. *Nature Biotechnology*, *34*(5), 539–546. https://doi.org/10.1038/nbt.3527

Kim, R.-K., Suh, Y., Yoo, K.-C., Cui, Y.-H., Kim, H., Kim, M.-J., … Lee, S.-J. (2015). Activation of KRAS promotes the mesenchymal features of basal-type breast cancer. *Experimental & Molecular Medicine*, *47*(1), e137. https://doi.org/10.1038/emm.2014.99

Kleinstreuer, N. C., Dix, D. J., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., … Judson, R. S. (2013). In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, *131*(1), 40–55. https://doi.org/10.1093/toxsci/kfs285

Kossler, N., Matheis, K. A., Ostenfeldt, N., Bach Toft, D., Dhalluin, S., Deschl, U., & Kalkuhl, A. (2015). Identification of specific mRNA signatures as fingerprints for carcinogenesis in mice induced by genotoxic and nongenotoxic hepatocarcinogens. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, *143*(2), 277–295. https://doi.org/10.1093/toxsci/kfu248

Kriebel, D., Hoppin, P. J., Jacobs, M. M., & Clapp, R. W. (2016). Environmental and Economic Strategies for Primary Prevention of Cancer in Early Life. *Pediatrics*, *138*(Supplement 1), S56–S64. https://doi.org/10.1542/peds.2015-4268I

Kufer, T. A., Silljé, H. H. W., Körner, R., Gruss, O. J., Meraldi, P., & Nigg, E. A. (2002). Human TPX2 is required for targeting Aurora-A kinase to the spindle. *The Journal of Cell Biology*, *158*(4), 617–623. https://doi.org/10.1083/jcb.200204155

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(1), 1–26. https://doi.org/10.18637/jss.v028.i05

Lai, Y.-P., Wang, L.-B., Wang, W.-A., Lai, L.-C., Tsai, M.-H., Lu, T.-P., & Chuang, E. Y. (2017). iGC—an integrated analysis package of gene expression and copy number alteration. *BMC Bioinformatics*, *18*. https://doi.org/10.1186/s12859-016-1438-2

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England)*, *24*(5), 719–720. https://doi.org/10.1093/bioinformatics/btm563

Lee, J. H., Ilic, Z., & Sell, S. (1996). Cell kinetics of repair after allyl alcohol-induced liver necrosis in mice. *International Journal of Experimental Pathology*, *77*(2), 63–72. https://doi.org/10.1046/j.1365-2613.1996.00964.x

Lee, S.-I., Pe'er, D., Dudley, A. M., Church, G. M., & Koller, D. (2006). Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proceedings of the National Academy of Sciences*, *103*(38), 14062–14067. https://doi.org/10.1073/pnas.0601852103

Lemen, R. A., Dement, J. M., & Wagoner, J. K. (1980). Epidemiology of asbestos-related diseases. *Environmental Health Perspectives*, *34*, 1–11.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest, *2*, 5.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)*, *27*(12), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Magkoufopoulou, C., Claessen, S. M. H., Tsamou, M., Jennen, D. G. J., Kleinjans, J. C. S., & van Delft, J. H. M. (2012). A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo. *Carcinogenesis*, *33*(7), 1421–1429. https://doi.org/10.1093/carcin/bgs182

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, *7*(Suppl 1), S7. https://doi.org/10.1186/1471-2105-7-S1-S7

McFarland, J. M., Ho, Z. V., Kugener, G., Dempster, J. M., Montgomery, P. G., Bryan, J. G., … Tsherniak, A. (2018). Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nature Communications*, *9*(1), 4610. https://doi.org/10.1038/s41467-018-06916-5

Mélard, P., Idrissi, Y., Andrique, L., Poglio, S., Prochazkova-Carlotti, M., Berhouet, S., … Cappellen, D. (2016). Molecular alterations and tumor suppressive function of the DUSP22 (Dual Specificity Phosphatase 22) gene in peripheral T-cell lymphoma subtypes. *Oncotarget*, *7*(42), 68734–68748. https://doi.org/10.18632/oncotarget.11930

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, *12*(4), R41. https://doi.org/10.1186/gb-2011-12-4-r41

Meyer, K. B., & Carroll, J. S. (2012). FOXA1 and breast cancer risk. *Nature Genetics*, *44*(11), 1176–1177. https://doi.org/10.1038/ng.2449

Monti, S., Chapuy, B., Takeyama, K., Rodig, S. J., Hao, Y., Yeda, K. T., … Shipp, M. A. (2012). Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell*, *22*(3), 359–372. https://doi.org/10.1016/j.ccr.2012.07.014

National Toxicology Program. (1978). Bioassay of pyrimethamine for possible carcinogenicity. *National Cancer Institute Carcinogenesis Technical Report Series*, *77*, 1–107.

National Toxicology Program. (1989). NTP Toxicology and Carcinogenesis Studies of Rhodamine 6G (C.I. Basic Red 1) (CAS No. 989-38-8) in F344/N Rats and B6C3F1 Mice (Feed Studies). *National Toxicology Program Technical Report Series*, *364*, 1–192.

National Toxicology Program. (1994). NTP Toxicology and Carcinogenesis Studies of Hexachlorocyclopentadiene (CAS No. 77-47-4) in F344/N Rats and B6C3F1 Mice (Inhalation Studies). *National Toxicology Program Technical Report Series*, *437*, 1–308.

Nie, A. Y., McMillian, M., Parker, J. B., Leone, A., Bryant, S., Yieh, L., … Lord, P. G. (2006). Predictive toxicogenomics approaches reveal underlying molecular mechanisms of nongenotoxic carcinogenicity. *Molecular Carcinogenesis*, *45*(12), 914–933. https://doi.org/10.1002/mc.20205

Niwa, T., Aoyama, I., Takayama, F., Tsukushi, S., Miyazaki, T., Owada, A., & Shiigai, T. (1999). Urinary indoxyl sulfate is a clinical factor that affects the progression of renal failure. *Mineral and Electrolyte Metabolism*, *25*(1–2), 118–122. https://doi.org/10.1159/000057433

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., … Bernard, P. S. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, *27*(8), 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370

Pearce, R. G., Setzer, R. W., Strope, C. L., Sipes, N. S., & Wambaugh, J. F. (2017). httk: R Package for High-Throughput Toxicokinetics. *Journal of Statistical Software*, *79*(1), 1–26. https://doi.org/10.18637/jss.v079.i04

Peck, D., Crawford, E. D., Ross, K. N., Stegmaier, K., Golub, T. R., & Lamb, J. (2006). A method for high-throughput gene expression signature analysis. *Genome Biology*, *7*(7), R61. https://doi.org/10.1186/gb-2006-7-7-r61

Pommier, Y. (2013). Drugging topoisomerases: lessons and challenges. *ACS Chemical Biology*, *8*(1), 82–95. https://doi.org/10.1021/cb300648v

Pommier, Y., Barcelo, J., Rao, V. A., Sordet, O., Jobson, A. G., Thibaut, L., … Redon, C. (2006). Repair of Topoisomerase I-Mediated DNA Damage. *Progress in Nucleic Acid Research and Molecular Biology*, *81*, 179–229. https://doi.org/10.1016/S0079-6603(06)81005-6

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.2307/2284239

Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., … Thomas, R. S. (2016). ToxCast Chemical Landscape:

Paving the Road to 21st Century Toxicology. *Chemical Research in Toxicology*, *29*(8), 1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Rodrigues, T., Santos, A. C., Pigoso, A. A., Mingatto, F. E., Uyemura, S. A., & Curti, C. (2002). Thioridazine interacts with the membrane of mitochondria acquiring antioxidant activity toward apoptosis – potentially implicated mechanisms. *British Journal of Pharmacology*, *136*(1), 136–142. https://doi.org/10.1038/sj.bjp.0704672

Ryffel, B. (1992). The carcinogenicity of ciclosporin. *Toxicology*, *73*(1), 1–22.

Schmidt, C. W. (2009). TOX 21: New Dimensions of Toxicity Testing. *Environmental Health Perspectives*, *117*(8), A348–A353.

Schröder, A., Wollnik, J., Wrzodek, C., Dräger, A., Bonin, M., Burk, O., … Zell, A. (2011). Inferring statin-induced gene regulatory relationships in primary human hepatocytes. *Bioinformatics (Oxford, England)*, *27*(18), 2473–2477. https://doi.org/10.1093/bioinformatics/btr416

Schröder, M. S., Culhane, A. C., Quackenbush, J., & Haibe-Kains, B. (2011). survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, *27*(22), 3206–3208. https://doi.org/10.1093/bioinformatics/btr511

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, *34*(2), 166–176. https://doi.org/10.1038/ng1165

Sekine, Y., Ikeda, O., Hayakawa, Y., Tsuji, S., Imoto, S., Aoki, N., … Matsuda, T. (2007). DUSP22/LMW-DSP2 regulates estrogen receptor-α-mediated signaling through dephosphorylation of Ser-118. *Oncogene*, *26*(41), 6038–6049. https://doi.org/10.1038/sj.onc.1210426

Shimizu, H., Yisireyili, M., Higashiyama, Y., Nishijima, F., & Niwa, T. (2013). Indoxyl sulfate upregulates renal expression of ICAM-1 via production of ROS and activation of NF-κB and p53 in proximal tubular cells. *Life Sciences*, *92*(2), 143–148. https://doi.org/10.1016/j.lfs.2012.11.012

Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, *68*(1), 7–30. https://doi.org/10.3322/caac.21442

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 397–420). New York, NY: Springer New York. https://doi.org/10.1007/0-387-29362-0_23

Sobel, M. E. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, *13*, 290–312. https://doi.org/10.2307/270723

Soldatow, V. Y., LeCluyse, E. L., Griffith, L. G., & Rusyn, I. (2013). In vitro models for liver toxicity testing. *Toxicology Research*, *2*(1), 23–39. https://doi.org/10.1039/C2TX20051A

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., … Golub, T. R. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, *171*(6), 1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Tan, B. T., Park, C. Y., Ailles, L. E., & Weissman, I. L. (2006). The cancer stem cell hypothesis: a work in progress. *Laboratory Investigation*, *86*(12), 1203–1207. https://doi.org/10.1038/labinvest.3700488

Tawa, G. J., AbdulHameed, M. D. M., Yu, X., Kumar, K., Ippolito, D. L., Lewis, J. A., … Wallqvist, A. (2014). Characterization of Chemically Induced Liver Injuries Using Gene Co-Expression Modules. *PLOS ONE*, *9*(9), e107230. https://doi.org/10.1371/journal.pone.0107230

The Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., … Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*, 1113–1120. https://doi.org/10.1038/ng.2764

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(Database issue), D158–D169. https://doi.org/10.1093/nar/gkw1099

Tice, R. R., Austin, C. P., Kavlock, R. J., & Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: a Tox21 update. *Environmental Health Perspectives*, *121*(7), 756–765. https://doi.org/10.1289/ehp.1205784

Tomasetti, C., Li, L., & Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science (New York, N.Y.)*, *355*(6331), 1330–1334. https://doi.org/10.1126/science.aaf9011

Tomasetti, C., & Vogelstein, B. (2015a). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, *347*(6217), 78–81. https://doi.org/10.1126/science.1260825

Tomasetti, C., & Vogelstein, B. (2015b). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, *347*(6217), 78–81. https://doi.org/10.1126/science.1260825

Uehara, T., Minowa, Y., Morikawa, Y., Kondo, C., Maruyama, T., Kato, I., … Urushidani, T. (2011). Prediction model of potential hepatocarcinogenicity of rat hepatocarcinogens using a large-scale toxicogenomics database. *Toxicology and Applied Pharmacology*, *255*(3), 297–306. https://doi.org/10.1016/j.taap.2011.07.001

Underhill, G. H., & Khetani, S. R. (2018). Bioengineered Liver Models for Drug Testing and Cell Differentiation Studies. *Cellular and Molecular Gastroenterology and Hepatology*, *5*(3), 426-439.e1. https://doi.org/10.1016/j.jcmgh.2017.11.012

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, *339*(6127), 1546–1558. https://doi.org/10.1126/science.1235122

Wang, H. J., Zakhari, S., & Jung, M. K. (2010). Alcohol, inflammation, and gut-liver-brain interactions in tissue damage and disease development. *World Journal of Gastroenterology*, *16*(11), 1304–1313. https://doi.org/10.3748/wjg.v16.i11.1304

Wang, L., & Eastmond, D. A. (2002). Catalytic inhibitors of topoisomerase II are DNA-damaging agents: induction of chromosomal damage by merbarone and ICRF-187. *Environmental and Molecular Mutagenesis*, *39*(4), 348–356. https://doi.org/10.1002/em.10072

Ward, J. M. (2007). The Two-Year Rodent Carcinogenesis Bioassay — Will It Survive? *Journal of Toxicologic Pathology*, *20*(1), 13–19. https://doi.org/10.1293/tox.20.13

Waters, M. D., Jackson, M., & Lea, I. (2010). Characterizing and predicting carcinogenicity and mode of action using conventional and toxicogenomics methods. *Mutation Research/Reviews in Mutation Research*, *705*(3), 184–200. https://doi.org/10.1016/j.mrrev.2010.04.005

Xie, T., d' Ario, G., Lamb, J. R., Martin, E., Wang, K., Tejpar, S., … Hodgson, J. G. (2012). A Comprehensive Characterization of Genome-Wide Copy Number Aberrations in Colorectal Cancer Reveals Novel Oncogenes and Patterns of Alterations. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0042001

Yan, M., Wang, C., He, B., Yang, M., Tong, M., Long, Z., … Liu, Q. (2016). Aurora-A Kinase: A Potent Oncogene and Target for Cancer Therapy. *Medicinal Research Reviews*, *36*(6), 1036–1079. https://doi.org/10.1002/med.21399

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., … Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, *45*(10), 1134–1140. https://doi.org/10.1038/ng.2760

Zhai, C., Li, Y., Mascarenhas, C., Lin, Q., Li, K., Vyrides, I., … Panaretou, B. (2014). The function of ORAOV1/LTO1, a gene that is overexpressed frequently in cancer: essential roles in the function and biogenesis of the ribosome. *Oncogene*, *33*(4), 484–494. https://doi.org/10.1038/onc.2012.604

Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., … Emilsson, V. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*, *153*(3), 707–720. https://doi.org/10.1016/j.cell.2013.03.030

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*, Article17. https://doi.org/10.2202/1544-6115.1128

Zhang, L., & Gong, F. (2015). Involvement of USP24 in the DNA damage response. *Molecular & Cellular Oncology*, *3*(1). https://doi.org/10.1080/23723556.2015.1011888

Zhang, X., Zhao, J., Hao, J.-K., Zhao, X.-M., & Chen, L. (2015). Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Research*, *43*(5), e31. https://doi.org/10.1093/nar/gku1315

Zhang, Z., Jia, C., Hu, Y., Sun, L., Jiao, J., Zhao, L., … Hu, J. (2012). The estrogenic potential of salicylate esters and their possible risks in foods and cosmetics. *Toxicology Letters*, *209*(2), 146–153. https://doi.org/10.1016/j.toxlet.2011.12.004

CURRICULUM VITAE