EMERGING BIOINFORMATIC TOOLS IN TOXICOGENOMICS

EDITED BY: Danyel Jennen and Paul Jennings
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88963-521-4
DOI 10.3389/978-2-88963-521-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

EMERGING BIOINFORMATIC TOOLS IN TOXICOGENOMICS

Topic Editors:
Danyel Jennen, Maastricht University, Netherlands
Paul Jennings, Vrije Universiteit Amsterdam, Netherlands

Toxicogenomics was established as a merger of toxicology with genomics approaches and methodologies more than 15 years ago, and considered of major value for studying toxic mechanisms-of-action in greater depth and for classification of toxic agents for predicting adverse human health risks. While the original focus was on technological validation of in particular microarray-based whole genome expression analysis (transcriptomics), mainly through cross-comparing different platforms for data generation (MAQC-I), it was soon appreciated that actually the wide variety of data analysis approaches represents the major source of inter-study variation. This led to early attempts towards harmonizing data analysis protocols focusing on microarray-based models for predicting toxicological and clinical end-points and on different methods for GWAS data (MAQC-II). Simultaneously, further technological developments, geared by increasing insights into the complexity of cellular regulation, enabled analyzing molecular perturbations across multiple genomics scales (epigenomics and microRNAs, metabolomics). While these were initially still based on microarray technology, this is currently being phased out and replaced by a variety of next generation sequencing-based methods enabling exploration of genomic responses to toxicants at even greater depth (SEQC-I). This raises the demand for reliable and robust data analysis approaches, ranging from harmonized bioinformatics concepts for preprocessing raw data to non-supervised and supervised methods for capturing and integrating the dynamic perturbations of cell function across dose and time, and thus retrieving mechanistic insights across multiple regulation scales.

Traditional toxicology focused on dose-dependently determining apical endpoints of toxicity. With the advent of toxicogenomics, efforts towards better understanding underlying molecular mechanisms has led to the development of the concept of Adverse Outcome Pathways, which are basically presented as a structural network of linearly related gene-gene interactions regulating key events for inducing apical toxic endpoints of interest. Impulse challenges from exposure of biological systems to toxic agents will however induce a cascade-type of events, presenting both adverse and adaptive processes, thus requiring bioinformatics approaches and methods for complex dynamic data, generated not only across dose, but clearly also across time. Currently, time-resolved toxicogenomics data sets are increasingly being assembled in the course of large-scaled research projects, for instance devoted towards developing toxicogenomics-based predictive assays for evaluating chemical safety which are no longer animal-based.

Citation: Jennen, D., Jennings, P., eds. (2020). Emerging Bioinformatic Tools in Toxicogenomics. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-521-4

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 September 2018
doi: 10.3389/fgene.2018.00396

Analysis of Time-Series Gene Expression Data to Explore Mechanisms of Chemical-Induced Hepatic Steatosis Toxicity

Alejandro Aguayo-Orozco[1], Frederic Yves Bois[2], Søren Brunak[1] and Olivier Taboureau[1,3]**

[1] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, [2] Institut National de l'Environnement Industriel et des Risques (INERIS), Unité Modèles pour l'Ecotoxicologie et la Toxicologie (METO), Verneuil en Halatte, France, [3] UMRS 973 INSERM, Université Paris Diderot, Université Sorbonne Paris Cité, Paris, France

OPEN ACCESS

Edited by:
Paul Jennings, VU University Amsterdam, Netherlands

Reviewed by:
Joshua Harrill, Environmental Protection Agency (EPA), United States
Scott Auerbach, National Institute of Environmental Health Sciences (NIEHS), United States

***Correspondence:**
Alejandro Aguayo-Orozco
alejandro.orozco@cpr.ku.dk;
Olivier Taboureau
olivier.taboureau@cpr.ku.dk;
olivier.taboureau@univ-paris-diderot.fr

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Genetics

Received: *05 May 2018*
Accepted: *30 August 2018*
Published: *18 September 2018*

Citation:
Aguayo-Orozco A, Bois FY, Brunak S and Taboureau O (2018) Analysis of Time-Series Gene Expression Data to Explore Mechanisms of Chemical-Induced Hepatic Steatosis Toxicity. Front. Genet. 9:396. doi: 10.3389/fgene.2018.00396

Non-alcoholic fatty liver disease (NAFLD) represents a wide spectrum of disease, ranging from simple fatty liver through steatosis with inflammation and necrosis to cirrhosis. One of the most challenging problems in biomedical research and within the chemical industry is to understand the underlying mechanisms of complex disease, and complex adverse outcome pathways (AOPs). Based on a set of 28 steatotic chemicals with gene expression data measured on primary hepatocytes at three times (2, 8, and 24 h) and three doses (low, medium, and high), we identified genes and pathways, defined as molecular initiating events (MIEs) and key events (KEs) of steatosis using a combination of a time series and pathway analyses. Among the genes deregulated by these compounds, the study highlighted OSBPL9, ALDH7A1, MYADM, SLC51B, PRDX6, GPAT3, TMEM135, DLGDA5, BCO2, APO10LA, TSPAN6, NEURL1B, and DUSP1. Furthermore, pathway analysis indicated deregulation of pathways related to lipid accumulation, such as fat digestion and absorption, linoleic and linolenic acid metabolism, calcium signaling pathway, fatty acid metabolism, peroxisome, retinol metabolism, and steroid metabolic pathways in a time dependent manner. Such transcription profile analysis can help in the understanding of the steatosis evolution over time generated by chemical exposure.

Keywords: hepatic steatosis, gene expression, transcriptomics, time-series analysis, pathways analysis, drug induced liver injury, DILI

INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD) is diagnosed increasingly worldwide and is considered to be the most common liver disorder in the West (Rector et al., 2008). NAFLD refers to a spectrum of hepatic disorders, ranging from simple hepatic steatosis with no apparent specific symptoms to hepatocellular carcinoma (Jozefczuk et al., 2012). Hepatic steatosis is caused by abnormal accumulation of triglycerides (TG) in the liver due to chemical exposures other than excessive alcohol consumption. This accumulation of TG in vesicles impairs hepatic function and makes the liver highly susceptible to other injuries related to metabolic syndrome and systemic energy

metabolism (Marchesini et al., 2003). Simultaneously, it affects the local immune system and that may lead to more severe autoimmune diseases (Antherieu et al., 2011). From a metabolic point of view, steatosis occurs when the fatty acids (FAs) influx or synthesis in the liver exceeds the capacity to clear them. The metabolic pathways leading to the development of hepatic steatosis are multiple, including enhanced non-esterified FA release from adipose tissue (lipolysis), increased *de novo* FAs (lipogenesis) and decreased β-oxidation (Fuchs et al., 2014).

Some toxicogenomics studies have been reported on drug-induced steatosis (DIS) or steatohepatitis (DISH) (Lake et al., 2011; Starmann et al., 2012; Hebels et al., 2014; Rabinowich and Shibolet, 2015), however, the mechanisms of action leading to steatosis are not fully understood. To support toxicity evidence with mechanistic pathways and mode of action for drug safety and risk assessment, the OECD has recently developed the adverse outcome pathway (AOP) concept. The AOP concept involves all the essential steps that take place in the toxicity pathways, from the molecular initiating event (MIE) at the protein or gene level, passing through organelle effect, cellular, tissue, organ and finally population effect. One key principle is that AOPs are chemical agnostic pathways (Vinken, 2013). Steatosis is one of the AOPs highly investigated and although AOPs are chemical agnostic pathways, the activation of some specific molecules, which lead to the over or under regulation of key events (KEs) with steatosis as a final outcome has been reported on the OECD AOP website[1].

In this study, we decided to analyze transcriptomic data on a set of 28 drugs tested in primary human hepatocytes and suspected to cause steatosis. In order to obtain an overall understanding of the disease, steatosis-producing chemicals were compiled and analyzed together. An interesting feature is that compounds have been studied at different times and doses, so we were able to perform a time-series analysis at the gene level but also at the pathway level. The distinction between time points and concentrations can explain how the different KEs affect one another, which will help in explaining complex hepatotoxicity. The results of our analysis support previously reported finding and provide new hypotheses that could be investigated further.

MATERIALS AND METHODS

Chemicals

For the current analysis, 28 compounds were selected according to their ability to induce steatosis in primary human hepatocytes (PHH) and the availability of gene expression data in the TG-GATEs (Toxicogenomics Project–Genomics-Assisted Toxicity Evaluation System) database (Igarashi et al., 2015). Furthermore, seven non-steatotic compounds available in TG-GATEs were also included according to the study carried out by Sahini and Borlak (2014) and Sahini et al. (2014) as negative controls. The negative controls have been associated with other histopathological observations in rat *in vivo* such as necrosis, cellular infiltration, fibrosis and granuloma (**Supplementary Table S1**). The TG-GATEs database contains data from PHH exposed to those compounds and collected using Affymetrix HG U133 Plus 2.0 gene expression microarrays. Two replicates were tested at three dose levels (low, medium, and high) and at three time points (2, 8, and 24 h after initial dosing). For each experiment, corresponding untreated controls are also tested. The 35 chemicals used in the study are summarized in **Table 1**. The specific dose and times can be found in **Supplementary Table S2**.

Microarray Data Analysis

All data were analyzed using the robust multi-array average (RMA) methodology in the Bioconductor R package for background-adjusted, normalized, and log-transformed perfect matched values of individual probes from the Affymetrix Human Genome U133 Plus 2.0 array (Irizarry et al., 2003). 54,675 probes corresponding to 19,945 uniquely annotated Gene Symbol IDs define each microarray. There is a total of 225 experiments according to concentration, time of exposure and compound used for the treatment. These experiments where analyzed in four steps: (1) all the experiments have been normalized concertedly. Such global normalization highlights the most important genes, which are those affected by the toxicity of more than one compound, and most likely in more than one time point and/or concentration (Krug et al., 2013). When dealing with gene expression microarray data, results can be affected by small differences in any number of non-biological variables, i.e., reagents or different technicians. (2) The two replicates per compound and condition were averaged. (3) Batch effect was accounted in the design matrix, reducing the bias effect on further steps of the analysis similarly to what has been performed by Grimberg et al. (2014). Concretely, for each gene a linear model following Eq. 1 (corresponding to a *t*-test comparison between two groups) was performed (Ritchie et al., 2015). (4) Subsequently, differentially expressed genes (DEGs) were calculated by dividing the average signal obtained from the chemical exposed group by the average signal from control receiving the vehicle only. The Student *t*-test was used to calculate the *p* value which was corrected by Bonferroni multiple testing. Finally, DEGs were selected by considering the *p* values less than 0.05 and fold-changes higher than 1.5. Genes that met these criteria also in the negative control set were removed from the deregulated gene's list for steatosis, assuming that these genes were not related to steatosis.

$$Y_{ij} = \alpha_j + x_i\beta_j + \varepsilon_{ij} \quad (1)$$

Time-Series Analysis

To characterize the deregulation of genes related to steatosis over time, after drug administration, a time-series analysis was performed on the 28 compounds using the package MasigPro in R (Nueda et al., 2014). This analysis was performed for each compound individually. With MasigPro, genes with significant temporal expression changes were selected and their variance at the different concentration (low, medium, and high) were

[1]https://aopwiki.org/

TABLE 1 | Compounds used in the analysis.

Compound name	Abbreviation	Cas No.	Sample	Reference
Allyl Alcohol	AA	107-18-6	Steatosis	(Waterfield et al., 1993)
Amiodarone	AM	1951-25-3	Steatosis	(Antherieu et al., 2011)
Acetaminophen	APAP	103-90-2	Steatosis	(Fontana, 2008)
Acetamide	AAA	60-35-5	Steatosis	(Zhang et al., 2017)
Amitriptyline	AMT	50-48-6	Steatosis	(Xia et al., 2000)
Aspirin	ASA	50-78-2	Steatosis	(Shen et al., 2014)
Coumarin	CMA	91-64-5	Steatosis	(Sahini et al., 2014)
Colchicine	COL	64-86-8	Steatosis	(Seiliez et al., 2016)
Clomipramine	CPM	303-49-1	Steatosis	(Xia et al., 2000)
Cyclosporin A	CSA	59865-13-3	Steatosis	(Lopez-Riera et al., 2017)
Clozapine	CZP	5786-21-0	Steatosis	(Zhang et al., 2007)
Diltiazem	DIL	42399-41-7	Steatosis	(Dowman et al., 2010)
Disulfiram	DSF	97-77-8	Steatosis	(Balakirev and Zimmer, 2001)
Ethanol	ETN	64-17-5	Steatosis	(Donohue, 2007)
Ethinylestradiol	EE	57-63-6	Steatosis	(Morii et al., 2014)
Ethionamide	ETH	536-33-4	Steatosis	(Zhang et al., 2017)
Hydroxyzine	HYZ	68-88-2	Steatosis	(Sahini et al., 2014)
Imipramine	IMI	50-49-7	Steatosis	(Xia et al., 2000)
Lomustine	LS	13010-47-4	Steatosis	(King and Perry, 2001)
Methapyrilene	MP	91-80-5	Steatosis	(Craig et al., 2006)
Methyltestosterone	MTS	58-18-4	Steatosis	(Schoonen et al., 2007)
Phenylbutazone	PhB	50-33-9	Steatosis	(Bessone, 2010)
Rifampicin	RIF	13292-46-1	Steatosis	(Tostmann et al., 2008)
Terbinafine	TBF	91161-71-6	Steatosis	(Choudhary et al., 2014)
Tetracycline	TC	60-54-8	Steatosis	(Antherieu et al., 2011)
Vitamin A	VA	68-26-8	Steatosis	(Liu et al., 2016)
Valproic acid	VPA	99-66-1	Steatosis	(Vitins et al., 2014)
Pirinixic acid	WY	50892-23-4	Steatosis	(Cannon and Eacho, 1991)
Carbamazepine	CBZ	298-46-4	Negative Control	(Bessone, 2010)
Diclofenac	DFNa	15307-86-5	Negative Control	(Bessone, 2010)
Indomethacin	IM	53-86-1	Negative Control	(Dehpour et al., 1999)
Naproxen	NP	22204-53-1	Negative Control	(Bessone, 2010)
Nifedipine	NIF	21829-25-4	Negative Control	(Basile and Mascia, 1999)
Nimesulide	NIM	51803-78-2	Negative Control	(Bessone, 2010)
Sulindac	SUL	103-90-2	Negative Control	(Bessone, 2010)

The reference corresponds to the publication in which drug-induced hepatic steatosis or negative control has been reported.

analyzed. As a first step, a regression on time for each gene taking all the variables present in the model, hence using all the genes, was performed. A false discovery rate (FDR) method was used to select genes with a value less than 0.05. Moreover, for each gene the best regression model was selected using stepwise regression. A backward method was used; therefore all genes were used as variables to initialize the modeling (p-value <0.05 were considered). In a final step, the R-squared of the regression model was used as cut-off value in order to reduce the amount of false positive findings (genes). R-squared was set to 0.6 to allow flexibility to the regression model, since we are working with all the compounds associated with steatosis, as suggested by MasigPro. Overall, MasigPro provides information on genes that change over time and in respect to the control. Such analysis can be visualized, plotting DEG of every single gene for each compound studied according to time and dose.

Gene Set and Pathway Analysis

In addition to the DEG and the time-series analysis, a pathway analysis was performed based on our gene expression analysis for the 28 compounds. Compared with the individual gene/molecule-based approach, pathway analysis is more sensitive, consistent and informative on the outcomes studied (Luo et al., 2009). In our study, the parametric statistical analysis model (PAGE) was used (Kim and Volsky, 2005). The method is based on a modified Gene Set Enrichment Analysis (GSEA). A gene randomization test was applied to the gene expression data, in which the significance of gene sets is identified for pathways (computing permutations of gene labels or a parametric distribution over genes). The database used for the study of the pathways was KEGG, which is a database resource that integrates genomic, chemical and systemic functional information for a large set of pathways (Kanehisa et al., 2012). In order to obtain a quantitative result of the

compound's effect over the pathways, a Gene Fold Enrichment (GFE) score was calculated. This score divided the number of genes deregulated by the total number of genes of the pathway being analyzed and then multiply by the statistical mean for the same pathway. Pathview, a tool set for pathway-based data integration and visualization, was used within the GAGE package in R for visualization of the genes deregulated in the KEGG pathway (Luo and Brouwer, 2013). In our context, no specific pathway has been developed for steatosis in KEGG or other pathway databases. So, we have considered the NAFLD pathway, which is the closest pathway to steatosis for the visualization.

Clustering

Previous studies have shown that the use of gene expression clustering can group samples in clusters that may lead to a good prediction of the gene-outcome relationship (Alizadeh et al., 2000; Handen and Ganapathiraju, 2015). Therefore, a pathway analysis was also performed on the set of compounds after clustering. Clustering was based on the logarithm base 2 of the fold change of the gene expression at the different conditions. The Euclidian distance implemented in Ward.D2 in R method was used. The clustering method was performed in all compounds containing information for at least 2 timepoints, hence excluding clozapine (CZP) from the analysis, which has been studied only for 1 timepoint. The clustering has been performed separately for the different time points. The clustering shown in **Figure 5** was performed on the compounds at 24 h. The determination of the number of clusters was done by the *elbow method*. A range of k values from 1 to 10, k value being the number of chemical belonging to a cluster, were considered in our analysis. For each k value the sum of squared errors (SERs) was calculated and the selection of the number of cluster was based on a compromise between the number of clusters and low SER. $K = 4$ was selected as it showed a close to maximum separation of the samples and low SER.

RESULTS

DEG Analysis

Firstly, we analyzed DEGs under all different conditions versus control for the 28 compounds with a global normalization of the 255 experiments (all together analysis). 742 genes are highly deregulated in at least one condition, i.e., one compound at a specific time and concentration (logFC $\geq$ 1.5 and with a FDR Bonferroni-corrected value $\leq$0.05).

For the pathway analysis, we looked specifically at the NAFLD pathway, a general pathway related to fatty liver, and for which steatosis might be related for some genes. Through the pathway enrichment analysis, many genes involved in the NAFLD are deregulated (**Figure 1**). Mapping the genes deregulated by the set of 28 compounds on the NAFLD pathway led to the observation that some genes are up regulated, in red (INSR, adipR, or PPARα), by a large set of compounds (AAA, PhB, CPM, and HYZ), whereas another set of genes is more often down regulated, in green (LXR, PI3K, FAS, CASP8, IKKB, and BAX) by others compounds (VA, ASA, and APAP). There are several compounds that show opposite effects by up/down-regulating the same genes. This is the case of CYP2E, AMPK and other mitochondrial genes. This supposes that there are different mechanisms of action that can trigger steatosis.

Gene Ontology pathway enrichment was also performed with the 742 genes in order to get an impression of the biological processes that were affected (**Figure 2**). The enrichment in terms of pathways based on GO terms was used in this study. At the first level of the pathway hierarchy, the deregulated genes are related to several pathways, including cellular process, metabolic process, localization, developmental process and immune system process. The two most significant are the cellular processes and metabolic processes. Within metabolic processes, primary metabolic processes are the most significant pathways targeted by the deregulated genes. In primary metabolic processes, the two most targeted pathways are nucleobase-containing compound metabolic processes and lipid metabolic processes. The latest contains steroid metabolic process, phospholipids metabolic process and FA metabolic process. Looking into the specific pathways, the most represented in the GO analysis are FA β-oxidation and acetyl-CoA metabolic process. So, we can note that many genes deregulated by the set of compounds affect lipids and FAs and play a role in steatosis.

Time-Series Analysis

In the previous analysis, the outcomes were analyzed independently of time and dose. To investigate the evolution of the expression over time, a time-series analysis was carried out using the R package MasigPro. After removing the genes involved in cell cycle according to GO biological processes (see **Supplementary Table S3**) (Barron and Li, 2016), MasigPro detected 48 genes with significant temporal expression changes (**Table 2**). These genes are mainly involved in metabolic and immune system pathways. Among them, some genes have previously been reported to play a hepatotoxic role such as MYADM (Megger et al., 2014), SLC51B (Arab et al., 2017), PRDX6 ((Newton et al., 2009; Pacifici et al., 2014), OSBPL9 (Hong and Tontonoz, 2014), GPAT3 (Khatun et al., 2016), TMEM135 (Exil et al., 2010), DLGDA5 (Liao et al., 2013), BCO2 (Ip et al., 2015), IDH3G (Pan et al., 2014), NEURL1B (Lawan et al., 2015), and TSPAN6 (Wang et al., 2012). An extensive work done in rodents related to steatosis adverse outcome described how OSBPL proteins promote the development of NAFLD in mice (Stein et al., 2017). Finally, the role of GPTA proteins has been reported to play a role in the development of hepatic steatosis (Yu et al., 2018).

Our results confirm previous transcriptomics analysis in rodents with deregulation of genes such as GPAT, KIF, CXCL, and SLC family genes (Sahini et al., 2014) and OSBP family that alters the lipid metabolism in mice (Béaslas et al., 2013).

An example of the visualization of the time-series analysis is shown in **Figure 3** for neutralized E3 ubiquitin protein ligase 1B (NEURL1B) after exposure to the 28 compounds. We can observe that NEURL1B is regulated in positive direction over time for many compounds. Other examples are presented in supplementary information (**Supplementary Figure S1**).

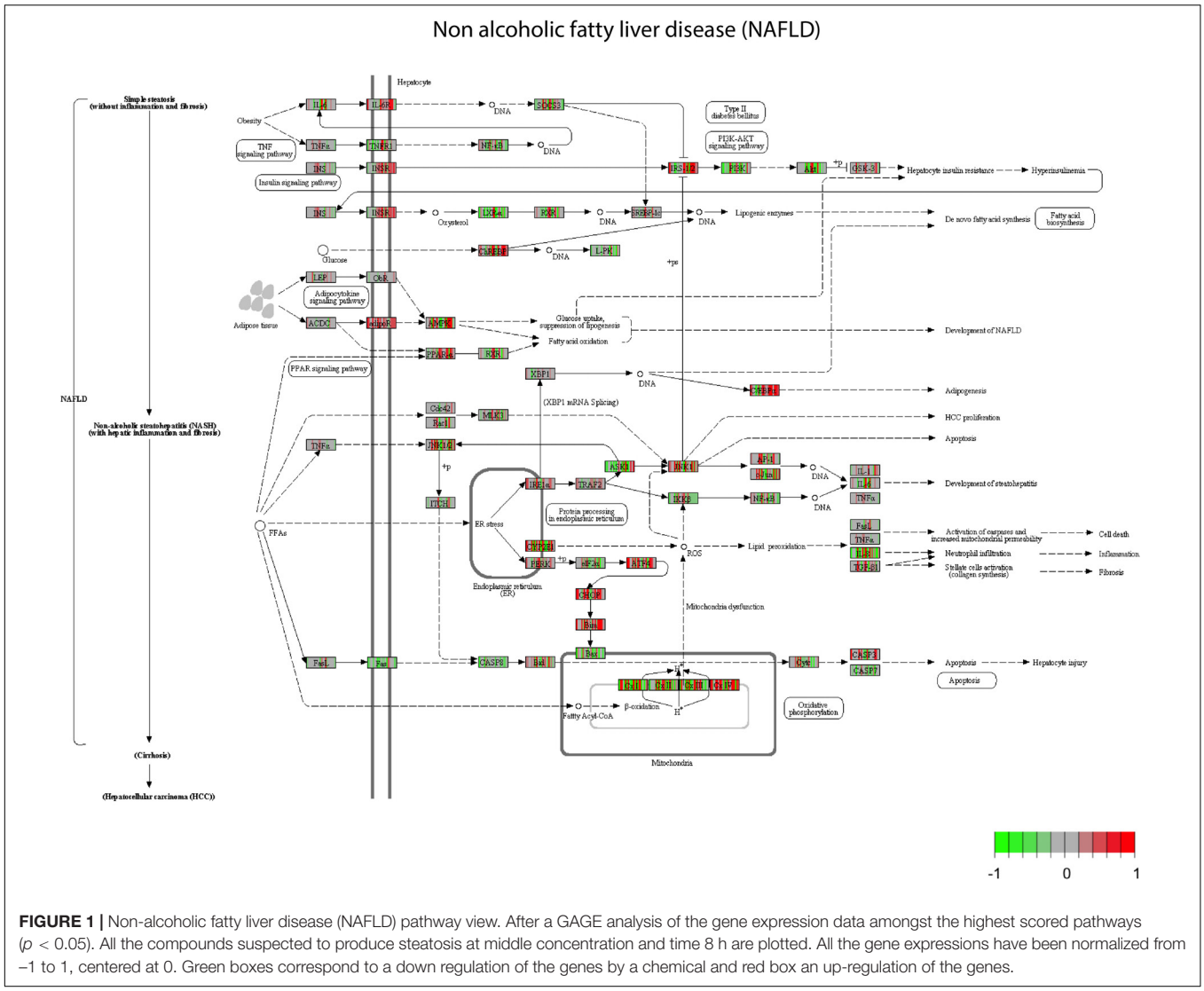


FIGURE 1 | Non-alcoholic fatty liver disease (NAFLD) pathway view. After a GAGE analysis of the gene expression data amongst the highest scored pathways ($p < 0.05$). All the compounds suspected to produce steatosis at middle concentration and time 8 h are plotted. All the gene expressions have been normalized from −1 to 1, centered at 0. Green boxes correspond to a down regulation of the genes by a chemical and red box an up-regulation of the genes.

From the gene list, other genes might play a role in steatosis and could be investigated further. For example, ALDH7A1 is highly deregulated in a time/dose-dependent manner. This gene is involved in oxidoreductase mechanisms and in protection of cell against oxidative stress by metabolizing a number of lipid peroxidation-derived aldehydes and could lead to steatosis. For some compounds, i.e., AM, APAP, LS, and MP, the de-regulation of ALDH7A1 is also dependent on the treatment concentration. The deregulation of these proteins will lead to a higher production of lipids, which together with a reduction of beta-oxidation of lipids could promote their accumulation in cells. Finally, a set of compounds deregulated some genes differently, suggesting that they trigger steatosis through another mechanism. This is the case for example for DSF and EE, which showed a weak deregulation of OSBL9 and a higher deregulation of the ALDH7A1.

Some genes known to be commonly associated to steatosis in human are not in this top list. This is the case for example of PNPLA3, which does not appear as one of the highest deregulated genes in our study. The genetic variation in PNPLA3 has been previously shown to play a role in the increase of FA accumulation in liver leading to steatosis (Romeo et al., 2008). So, the lack of the specific polymorphism related to susceptibility to steatosis in the cells used could explain the non-deregulation of this gene in our study.

Overall, this list of genes provides an insight into the mechanistic pathways already related to steatosis, as well as new hypotheses that can be analyzed further. Interestingly, the expression for many genes vary a little from control as a function of dose and the difference in the pattern of expression between control and treatment is relatively low. This confirmed a previous analysis showing that the doses differences between treatments in rat primary hepatocytes explain less than 0.1% of variation in all cases (Sutherland et al., 2016). One possible explanation is primary hepatocytes rapidly dedifferentiate (Lauschke et al., 2016) which could generate a gradual down regulation of hepatocyte function over time in culture.

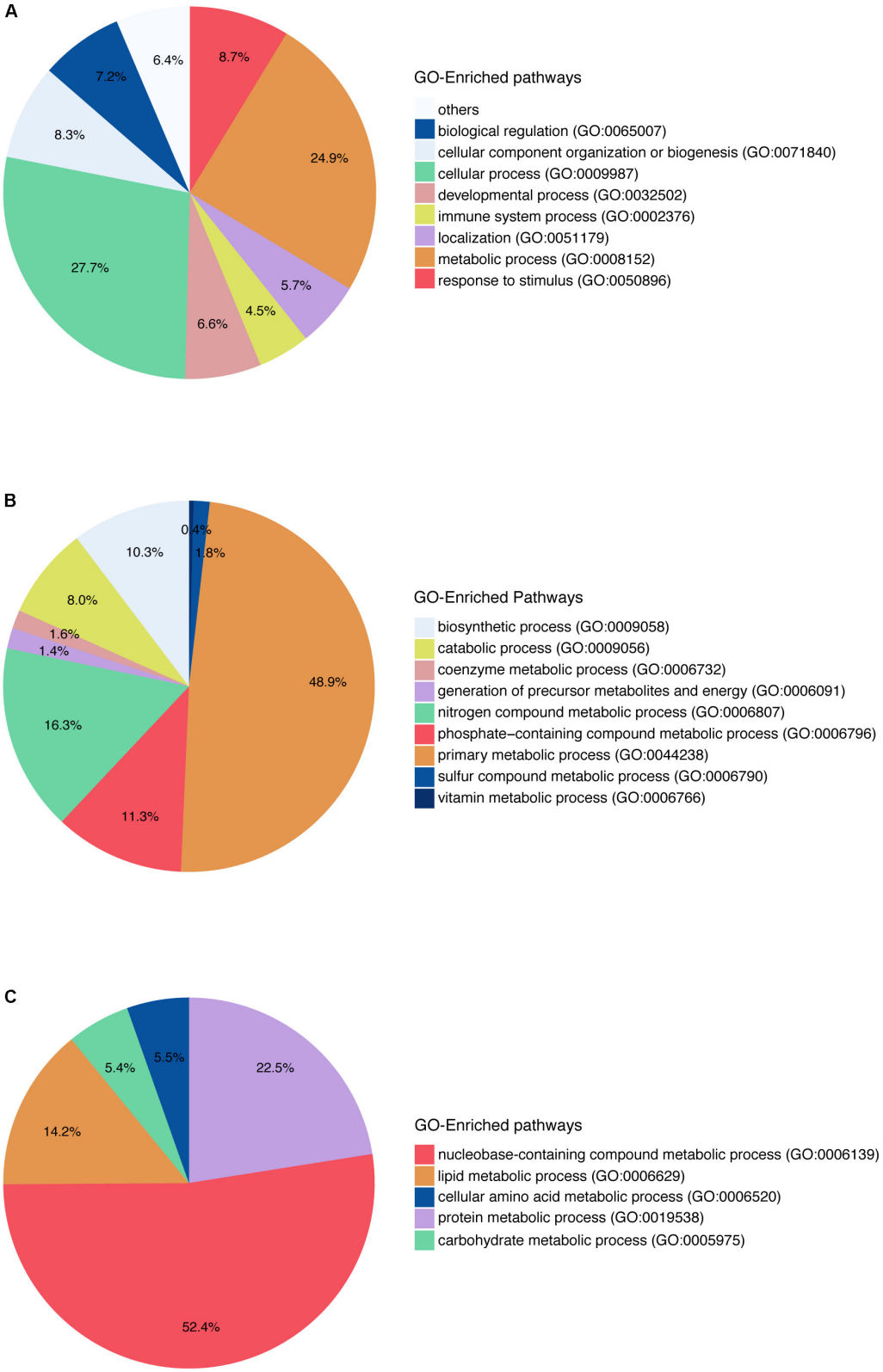


FIGURE 2 | Gene Ontology pathway enrichment for *in vitro* human hepatocytes. **(A)** The two main blocks of pathways that are deregulated according to the genes that have a log2FC above absolute value 1.5 and a *q*-value ≤0.05, are affecting the metabolic pathways as well as the cellular processes in general. **(B)** The pathways affected within the metabolic pathways are shown here, and they affect mainly the primary metabolic process. **(C)** Pathways represented within primary metabolic process.

TABLE 2 | Deregulated genes over time and dose.

Entrez-ID	Gene symbol	Gene full name
91663	MYADM	Myeloid associated differentiation marker
27286	SRPX2	Sushi repeat containing protein, X-linked 2
10491	CRTAP	Cartilage associated protein
84803	GPAT3	Glycerol-3-phosphate acyltransferase 3
9787	DLGAP5	DLG associated protein 5
83875	BCO2	Beta-carotene oxygenase 2
3161	HMMR	Hyaluronan mediated motility receptor
3421	IDH3G	Isocitrate dehydrogenase 3 (NAD(+)) gamma
6790	AURKA	Aurora kinase A
28998	MRPL13	Mitochondrial ribosomal protein L13
54492	NEURL1B	Neuralized E3 Ubiquitin Protein Ligase 1B
2633	GBP1	Guanylate binding protein 1
10112	KIF20A	Kinesin family member 20A
80114	BICC1	BicC family RNA binding protein 1
65084	TMEM135	Transmembrane protein 135
7105	TSPAN6	Tetraspanin 6
9615	GDA	Guanine deaminase
9488	PIGB	Phosphatidylinositol glycan anchor biosynthesis class B
55771	PRR11	Proline rich 11
11167	FSTL1	Follistatin like 1
2519	FUCA2	Fucosidase, alpha-L-2, plasma
9588	PRDX6	Peroxiredoxin 6
79594	MUL1	Mitochondrial E3 ubiquitin protein ligase 1
51292	GMPR2	Guanosine monophosphate reductase 2
81610	FAM83D	Family with sequence similarity 83 member D
55872	PBK	PDZ binding kinase
59	ACTA2	Actin, alpha 2, smooth muscle, aorta
7802	DNALI1	Dynein axonemal light intermediate chain 1
5445	PON2	Paraoxonase 2
3242	HPD	4-hydroxyphenylpyruvate dioxygenase
28998	MRPL13	Mitochondrial ribosomal protein L13
11004	KIF2C	Kinesin family member 2C
1606	DGKA	Diacylglycerol kinase alpha
10158	PDZK1IP1	PDZK1 interacting protein 1
9122	SLC16A4	Solute carrier family 16 member 4
23082	PPRC1	Peroxisome proliferator-activated receptor gamma, coactivator-related 1
123264	SLC51B	Solute carrier family 51 beta subunit
6372	CXCL6	C-X-C motif chemokine ligand 6
79053	ALG8	ALG8, alpha-1,3-glucosyltransferase
9928	KIF14	Kinesin family member 14
788	SLC25A20	Solute carrier family 25 member 20
114883	OSBPL9	Oxysterol binding protein like 9
55526	DHTKD1	Dehydrogenase E1 and transketolase domain containing 1
56922	MCCC1	Methylcrotonyl-CoA carboxylase 1
10351	ABCA8	ATP binding cassette subfamily A member 8
501	ALDH7A1	Aldehyde dehydrogenase 7 family member A1
516	ATP5G	ATP Synthase Membrane Subunit C Locus 1
9488	PIGB	Phosphatidylinositol glycan anchor biosynthesis class B

Pathway Time-Series Analysis

Due to the broad pharmacological and physicochemical characteristics of the compounds used for this study, we developed a new type of time-series analysis at the pathway level. For this purpose, all the compounds were analyzed to obtain the most significantly deregulated pathways including their corresponding GFE score (**Figure 4**).

At low concentration, pathways such as FA degradation and oxidative phosphorylation start to get down regulated. The deregulation of the oxidative phosphorylation is prominently affecting the mitochondria, therefore reducing the activity within this organelle, such as β-oxidation, which is the catabolic process through which FAs are broken down. The PIK3-Akt signaling pathway gets up regulated through the activation of the AMPK signaling pathway or downstream, Mtor signaling pathway, affecting the metabolism of the cell (Li et al., 2010). At higher concentrations, other important pathways are affected. The steroid biosynthesis is up-regulate at middle and high dose. FA biosynthesis and FA elongation are also among the up regulated pathways. Hence more FAs and lipids are produced. In contrast, the protein processing in ER, known to be related to lipid homeostasis, is down regulated. Interestingly, several studies have previously shown the existence of comorbidities between liver diseases and cardiovascular (CDV) diseases (Anstee et al., 2018). The deregulation of the renin-angiotensin system could explain part of the relation between steatosis and any possible CDV disease. Vitamin digestion and absorption is down regulated, which also points toward de-regulation in the FA β-oxidation. Also tyrosine metabolism, which is related to liver damage displays down-regulation. When the liver is damaged, phenylalanine cannot be converted to tyrosine. At this highest concentration, the adipocytokine signaling pathway and TNFα signaling pathway are deregulated, which indicates an activation of the cellular immune system. This immune system de-regulation may contribute the steatotic condition to move forward to other more severe drug-induced liver damages. Note that at high dose, cells often develop non-specific toxicity and the pathways altered may be not related solely to steatosis but also to other toxicity endpoints. The pathway analysis confirmed the little contribution of doses over time at the gene level observed previously, as the majority of the pathways deregulated in middle dose are also present in high dose.

Finally, to obtain a more characteristic view on the specific action points of the different compounds, we performed a similar analysis after clustering the compounds through the gene's signature similarity. Using the Euclidean distance based on the log2FC of the gene expression, all compounds were clustered in four sets (**Figure 5**).

VPA was clustered separately. MP and AA formed a different cluster as well as APAP and COL. A final cluster contained the remaining compounds. This last cluster contains essentially drugs used to treat a variety of conditions, acting as immunosuppressant's, antineoplastic agents, antibiotics, biguanides and butylpyrazolidines.

After clustering, the pathway time-series analysis was performed on each of the four clusters (**Figure 6**). For the compounds of the larger cluster, cluster 1 (TBF, AMT, DIL,

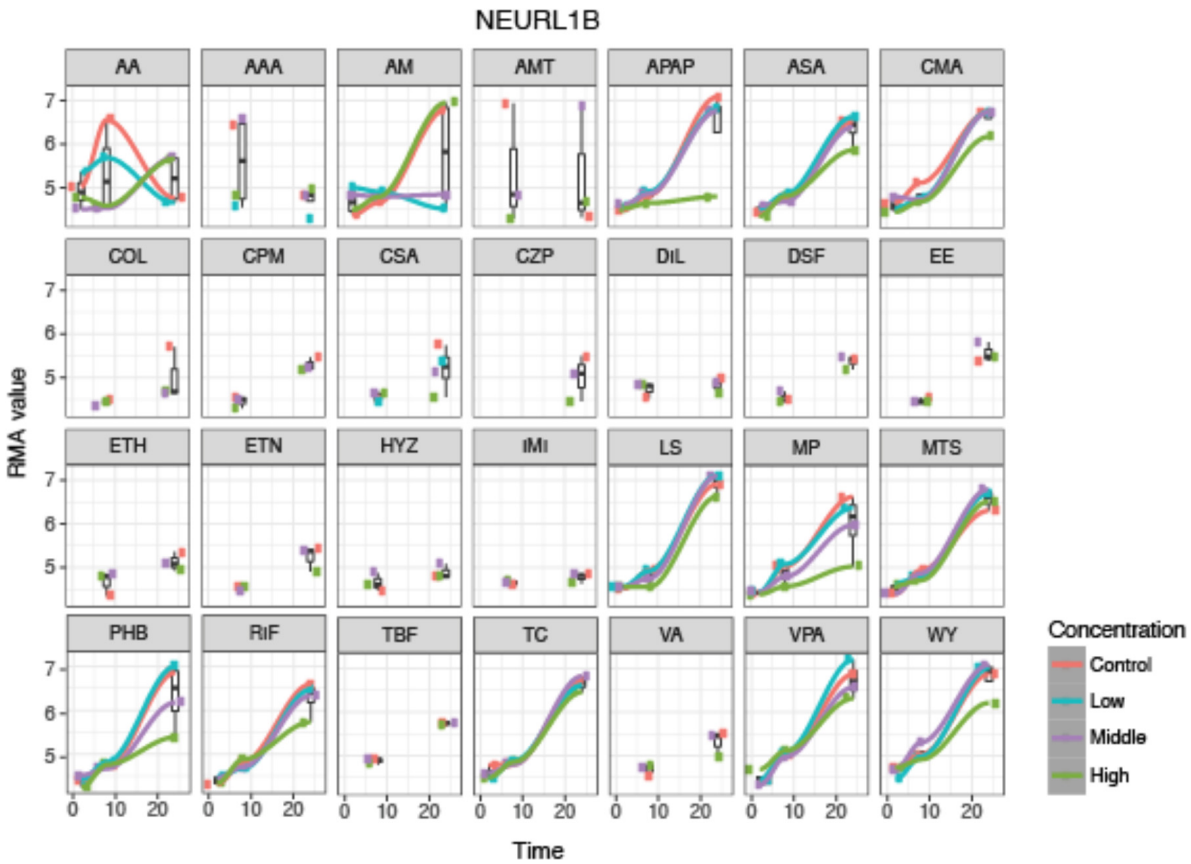


FIGURE 3 | De-regulation over time of the 28 steatotic chemicals for the selected protein, NEURL1B, from the time-series analysis. The *y*-axis indicates the normalized expression value of the gene at every time-point, 2, 8 and 24 h (*x*-axis), where the different colors indicate the dosages (red indicates control, blue low dose, purple middle dose, and green high dose). It shows a time dependency of the protein expression and compound specific dose effect. However, it does not show a systematic change on the gene expression due to dose. For some compounds only two time-point samples were taken.

PhB, HYZ, CSA, WY, RIF, ASA, AM, AAA, EE, VA, TC, MTS, IMI, CPM, DSF, CMA, ETH, LS, ETN), at low concentrations of treatment metabolic pathways such as fat digestion and absorption, linoleic and linolenic acid metabolism, calcium signaling pathway and others are up regulated over time. Other pathways, such as FA metabolism, peroxisome, retinol metabolism, and some steroid metabolic pathways are down regulated. At higher concentrations from time 2 to 24 h, the FA metabolism, calcium signaling pathway and steroid hormone biosynthesis increase over time, showing a de-regulation of these pathways promoted by the treatment. The FA degradation pathway is down regulated. It means that the FAs inside the cell are increasing and there are no pathways to deplete them. The oxidative phosphorylation becomes down regulated over time. So, the oxidative conditions in the mitochondria are starting to be reduced at this concentration. Finally, at the highest concentration tested, many signaling pathways known to be steatosis-producing related are targeted. FoxO, MAPK, PPAR signaling pathways, are highly up regulated. Steroid hormone biosynthesis, FA biosynthesis, glycerophospholipid metabolism, glycosphyngolipid biosynthesis and other lipid metabolic pathways are also up regulated. Moreover, SNARE interactions in vesicular transport are also up regulated over time, which could indicate the internalization of the FAs into vesicles, and so accumulation inside the cells. In contrast, pathways, such as oxidative phosphorylation, vitamin digestions and absorption and FA degradation are down regulated over time.

For cluster 2 (AA, MP) (**Figure 6**) at low concentrations the most highly up regulated pathways are the PI3K-Akt signaling pathway, the Mtor signaling pathway and the adipocytokine signaling pathway. Some metabolic pathways like FA degradation, peroxisome and retinol metabolism are down regulated. With the increasing concentration, steroid biosynthesis starts to be up regulated. At the highest concentration, glycolysis/gluconeogenesis, FA degradation, TNFα signaling pathway, tyrosine metabolism, peroxisome, PPAR signaling pathway and retinol metabolism become down-regulated over the time and FA metabolism, adipocytokine signaling pathway, MAPK signaling pathway, among others, become up-regulated. This could be explained by a lesser effect of these compounds. Therefore higher concentrations are needed in order to deregulate the cell to a steatotic pattern.

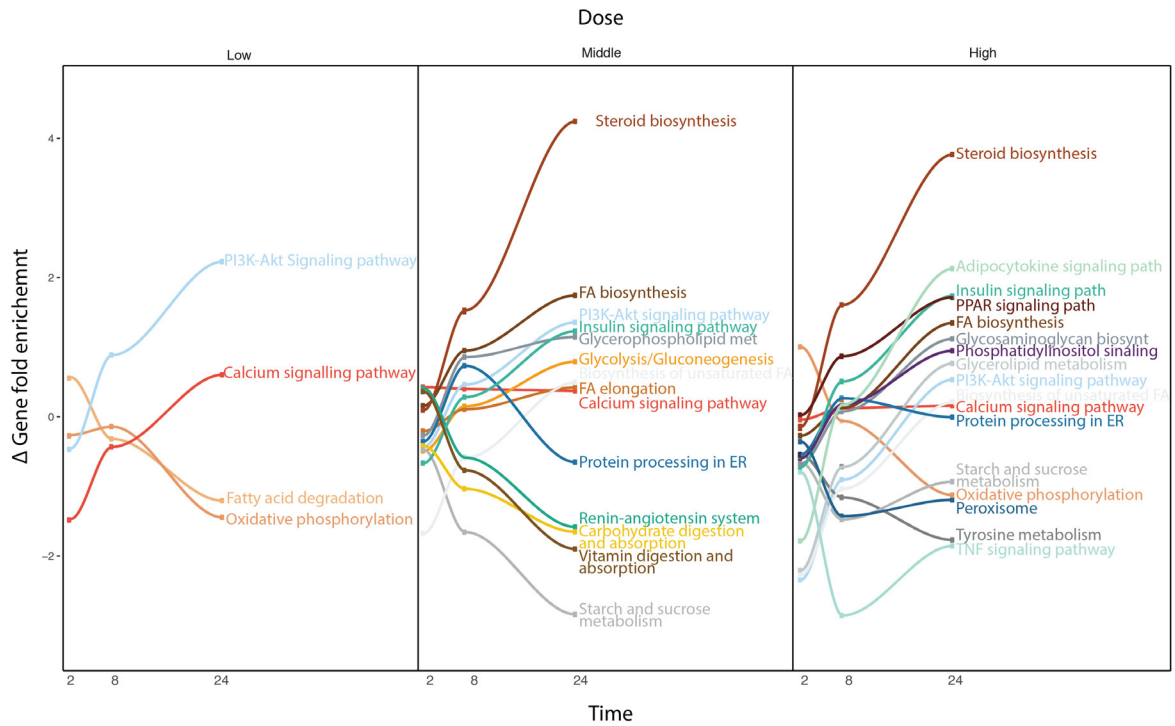


FIGURE 4 | Time-series pathway analysis based on GFE and depending on the concentration. The *y*-axis displays the overall Gene Fold Enrichment.

In the case of cluster 3 (APAP, COL) (**Figure 6**), at low concentrations, oxidative phosphorylation is highly down regulated, together with retinol metabolism and some lipid metabolism such as ether lipid metabolism and glycolysis/gluconeogenesis. On the other hand, FA elongation, TNFα signaling pathway, PI3K-Akt signaling pathway and sphingolipid signaling pathway and sphingolipid metabolism are highly up regulated. At higher concentrations, retinol metabolism continues to be down-regulated, glycolysis/gluconeogenesis, FA elongation are down-regulated and protein processing in the ER, steroid biosynthesis, SNARE interactions in vesicular transport among others are up regulated. These deregulations could affect the export of FAs to the exterior of the cell and their accumulation within organelles.

For the last cluster, containing only VPA (**Figure 6**) at low, middle and high doses, FA degradation, PPAR signaling pathway and retinol metabolism, all of them involved in the elimination of FAs are up regulated. This compound produces a strong effect on the metabolism of FAs and lipids and therefore the cells react with increasing the pathway activities associated with FA degradation.

So, it is interesting to see that, each of the cluster shows some pathway deregulation related lipid metabolism, FA degradation, glycolysis or PPAR signaling pathway, all related to steatosis.

DISCUSSION

The conventional assumption that a drug acts selectively on a single target is shifting toward "drug-holistic" systems based approaches. Similarly, a disease or a toxicity endpoint reflects not only the impairment of a unique gene. In fact, the disruption of many genes and pathways can lead to a disease or a specific toxicity. In the case of steatosis, we have focused the study on trying to understand the underlying mechanisms for steatosis using a set of diverse compounds. Considering the MIEs and KEs known to lead to the AOP steatosis (based on AOP-Wiki), our study confirms the deregulation of these biomarkers and highlighted new genes that produce steatosis. With the development of a time-series analysis combined with pathway analysis, it is possible to follow the evolution of the pathways over time and how they are connected to the different stages of steatosis.

Interestingly, the integration of a large and diverse set of compounds in the analysis pinpoints their specificity in leading to steatosis. However, our results show that the time seems to have a higher impact in the DEGs and pathways analysis than the concentration. The early dedifferentiation of PHH in 2D cultures might explain this observation. It is also possible that the global normalization reduces the specific signal of some genes. Additionally, for more than half of the compounds studied, only two times points have been tested experimentally, which might influence the results.

In our study, the compounds have been tested in PHH and the translation to human liver tissue would be of great interest to validate these outcomes. Some rats *in vivo* data beyond the 24 h time point are available in TG-GATEs and could be analyzed

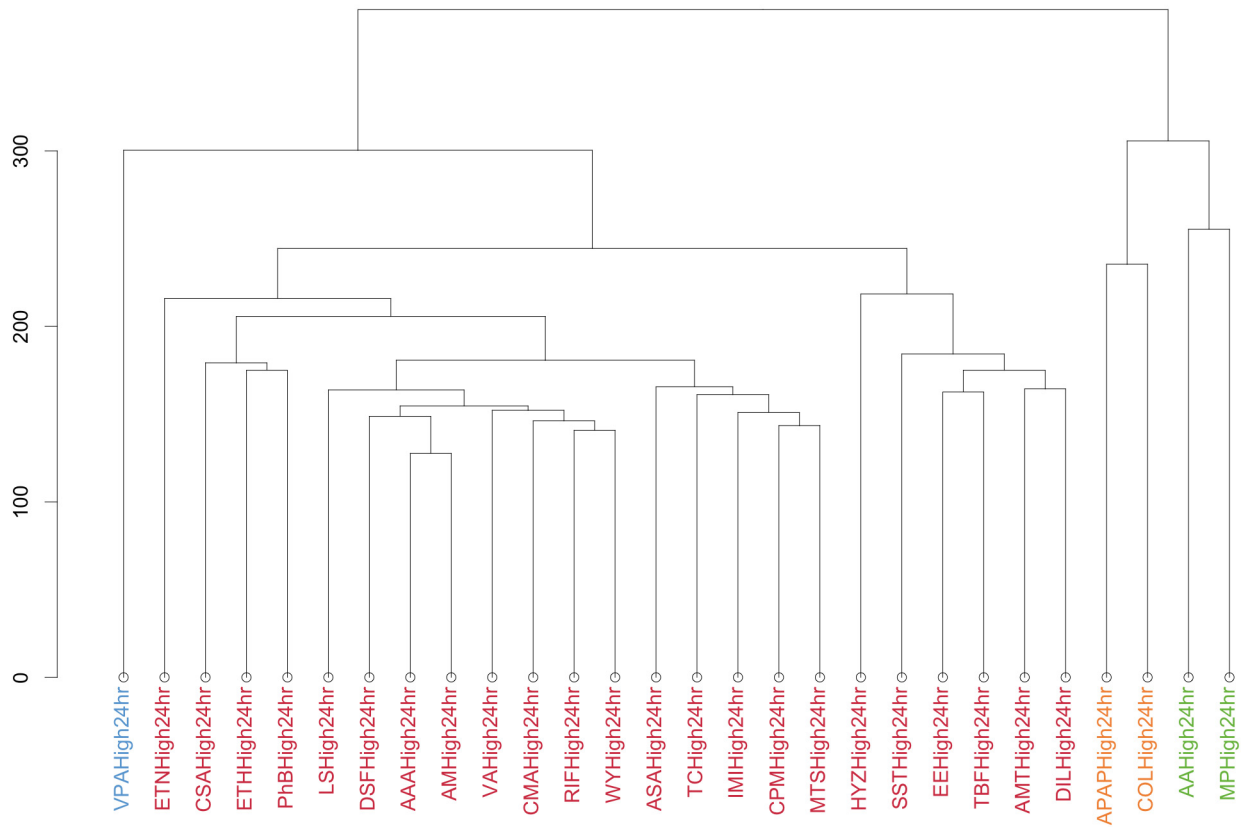


FIGURE 5 | Clustering by Euclidean distance of the different compounds based on the logarithm in base 2 Fold Change of the gene expression for the different conditions. Each color represents a cluster. The *y*-axis is a measure of closeness of either cluster or individual data points.

similarly in order to evaluate the overlap between *in vitro* and *in vivo* data. Finally, it has been reported that 3D cell cultures could be a more suitable system to mimic human organs that 2D cultures (Fey and Wrzesinski, 2012) and it would be interesting to assess the steatogenic effect of these compounds in this 3D spheroid system.

To summarize, most of the genes that are associated with a steatosis AOP, described in AOP wiki, have been found in our study. The integration of the previously published information on steatosis with the newly found genes and pathways from our analysis can enrich the knowledge of developed AOPs on steatosis. The **Figure 7A** represents an AOP network, i.e., the result of an accumulation of a number of individual AOPs listed on the AOP Wiki website. The list of AOPs used for the completion of the full steatosis pathway is: 34 (LXR activation leading to hepatic steatosis), 36 (Peroxisomal Fatty Acid Beta-Oxidation Inhibition Leading to Steatosis), 57 (AhR activation leading to hepatic steatosis), 58 (NR1I3 (CAR) suppression leading to hepatic steatosis), 60 (NR1I2 (Pregnane X Receptor, PXR) activation leading to hepatic steatosis), 61 (NFE2L2/FXR activation leading to hepatic steatosis). Besides, the capture of coenzyme A by VPA was added to the mechanistic pathway (Schumacher and Guo, 2015), as well as oxidative stress (Spahis et al., 2017). In this figure, we can see direct (and indirect) interaction between genes suggested by the analysis and known genes. For example, TSPAN6 deregulates oxidative phosphorylation, which acts on the mitochondrial β-oxidation. The deregulation of ALDH7A1 will lead to a higher production of lipids and impact the oxidative stress with a reduction of β-oxidation. In contrast, PON2 impacts the immune system. We looked also at the cellular compartmental level and how the genes deregulation can perturb the interaction with each other and lead to steatosis (**Figure 7B**). We can see that all the cell compartments can be involved in steatosis, many of which undertake functions within the mitochondria and the nucleus. More specifically, perturbation in endoplasmic reticulum and vesicles through the genes MUL1, TMEM135, OSBPL9, SCD1, SREBP-1C, GPAT3 can lead to steatosis.

Other studies have reported computational approaches to leverage large-scale toxicogenomic information, biological pathways and high throughput data for the identification of toxicity pathways. For example, Bell et al. (2016) described a computational approach in which curated biological pathways

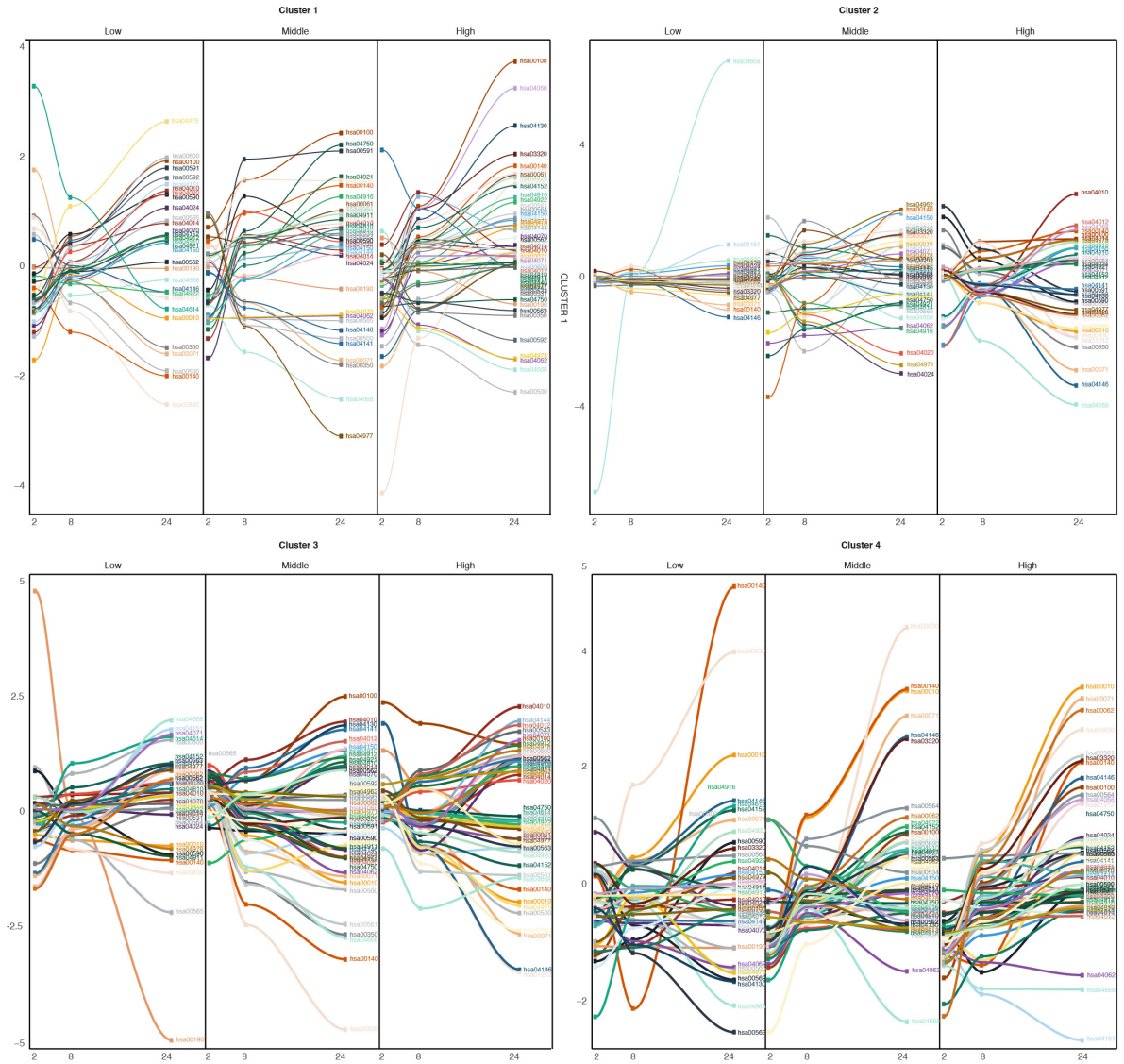


FIGURE 6 | Time-series pathway analysis based on GFE and depending on the concentration for the four clustered of compounds. The *y*-axis displays the overall Gene Fold Enrichment.

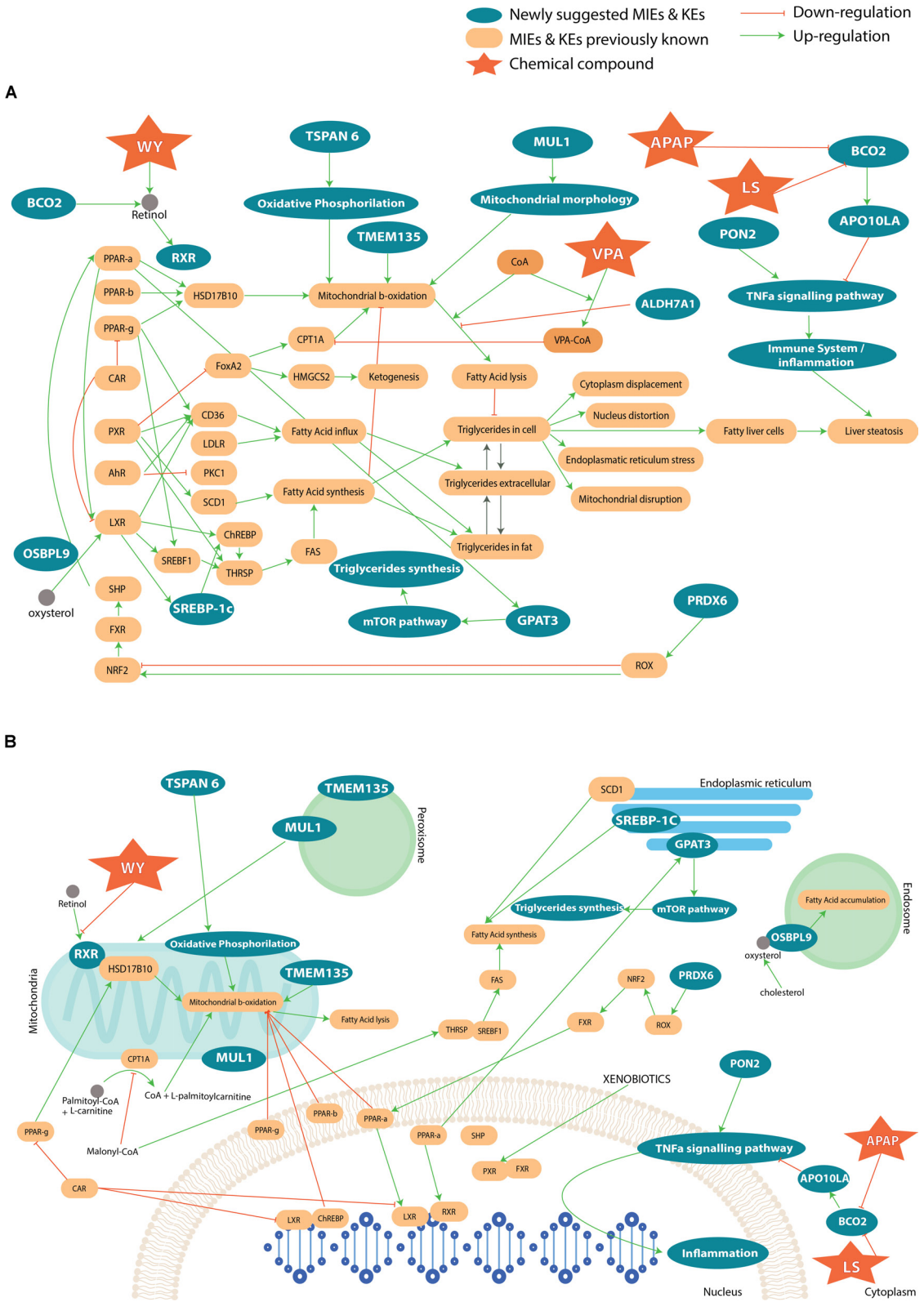


FIGURE 7 | (A) Adverse Outcome Pathway related to steatosis network, which summarizes a collection of information across several AOPs. The yellow boxes represent known genes and key events associated to steatosis. The blue color shows the newly suggested genes and pathways involved in the AOPs. Only compounds to which there is a higher likelihood to affect the MIE or KE than other compounds have been introduced in the AOP. In **(B)** the different cell compartments and how the interact with each other in the pathways that lead to steatosis.

and high-throughput toxicity data are used to identify toxicity pathways. This computational method uses a data-driven approach to assemble an AOP, which allows for the integration of biological information into pathway-based networks and can be updated with new information. Coupling both approaches could be interesting in the enrichment of the steatosis AOPs.

Overall, our findings illustrate how an integrative computational chemical system biology approach can be used to study steatosis and obtain new metabolic pathways that are deregulated during the process of liver injury by chemical exposure. Obviously, these findings need to be further validated with additional experimental studies. These associations are potentially not causative but more reflect biomarkers along the pathway to develop steatosis. In many case, changes in gene expression are a response to a stressor and it is only when these adaptive changes are overwhelmed that the adverse effect occurs.

AUTHOR CONTRIBUTIONS

AA-O performed the experiments. AA-O and OT analyzed the results and wrote the paper. FB and SB contributed in the writing of the final manuscript.

FUNDING

This work was part of the EU-ToxRisk project, which was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 681002. This work was also supported by the Novo Nordisk Foundation grant No. NNF14CC0001.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00396/full#supplementary-material

FIGURE S1 | Time-series analysis for a set gene involved in steatosis.

TABLE S1 | Histopathology findings in rat *in vivo* assays for the negative control compounds. The compounds are listed together with the dose level and time at which the pathological finding happened.

TABLE S2 | List of compounds used in the analysis with the specific time points, the exact dose and the dose level.

TABLE S3 | Genes removed from results table after time-series analysis for the role in cell cycle pathways. Filter carried out by selecting genes that appear in any cell cycle related pathway from GO Biological processes.

REFERENCES

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511. doi: 10.1038/35000501

Anstee, Q. M., Mantovani, A., Tilg, H., and Targher, G. (2018). Risk of cardiomyopathy and cardiac arrhythmias in patients with nonalcoholic fatty liver disease. *Nat. Rev. Gastronterol. Hepatol.* 15, 425–439. doi: 10.1038/s41575-018-0010-0

Anthierieu, S., Rogue, A., Fromenty, B., Guillouzo, A., and Robin, M. A. (2011). Induction of vesicular steatosis by amiodarone and tetracycline is associated with up-regulation of lipogenic genes in HepaRG cells. *Hepatology* 53, 1895–1905. doi: 10.1002/hep.24290

Arab, J. P., Karpen, S. J., Dawson, P. A., Arrese, M., and Trauner, M. (2017). Bile acids and nonalcoholic fatty liver disease: molecular insights and therapeutic perspectives. *Hepatology* 65, 350–362. doi: 10.1002/hep.28709

Balakirev, M. Y., and Zimmer, G. (2001). Mitochondrial injury by disulfiram: two different mechanisms of the mitochondrial permeability transition. *Chem. Biol. Interact.* 138, 299–311. doi: 10.1016/S0009-2797(01)00283-6

Barron, M., and Li, J. (2016). Identifying and removing the cell-cycle effect from single-cell RNA-sequencing data. *Sci. Rep.* 6:33892. doi: 10.1038/srep33892

Basile, C., and Mascia, E. (1999). Dihydropyridine calcium channel blockers: a rare and reversible cause of hepatotoxicity with cholestasis in a CAPD patient. *Nephrol. Dial. Transplant.* 14, 2776–2777. doi: 10.1093/ndt/14.11.2776

Béaslas, O., Metso, J., Nissilä, E., Laurila, P.-P., Kaiharju, E., Batchu, K. C., et al. (2013). Osbpl8 deficiency in mouse causes an elevation of high-density lipoproteins and gender-specific alterations of lipid metabolism. *PLoS One* 8:e58856. doi: 10.1371/journal.pone.0058856

Bessone, F. (2010). Non-steroidal anti-inflammatory drugs: what is the actual risk of liver damage? *World J. Gastroenterol.* 16, 5651–5661. doi: 10.3748/wjg.v16.i45.5651

Bell, S. M., Angrish, M. M., Wood, C. E., and Edwards, S. W. (2016). Integrating publicly available data to generate computationally predicted adverse outcome pathways for fatty liver. *Toxicol. Sci.* 150, 510–520. doi: 10.1093/toxsci/kfw017

Cannon, J. R., and Eacho, P. I. (1991). Interaction of LY171883 and other peroxisome proliferators with fatty-acid-binding protein isolated from rat liver. *Biochem. J.* 280, 387–391. doi: 10.1042/bj2800387

Choudhary, N. S., Kotecha, H., Saraf, N., Gautam, D., and Saigal, S. (2014). Terbinafine induced liver injury: a case report. *J. Clin. Exp. Hepatol.* 4, 264–265. doi: 10.1016/j.jceh.2014.03.040

Craig, A., Sidaway, J., Holmes, E., Orton, T., Jackson, D., Rowlinson, R., et al. (2006). Systems toxicology: integrated genomic, proteomic and metabonomic analysis of methapyrilene induced hepatotoxicity in the rat. *J. Proteome Res.* 5, 1586–1601. doi: 10.1021/pr0503376

Dehpour, A. R., Mani, A. R., Amanlou, M., Nahavandi, A., Amanpour, S., and Bahadori, M. (1999). Naloxone is protective against indomethacin-induced gastric damage in cholestatic rats. *J. Gastroenterol.* 34, 178–181. doi: 10.1007/s005350050240

Donohue, T. M. Jr. (2007). Alcohol-induced steatosis in liver cells. *World J. Gastroenterol.* 13, 4974–4978. doi: 10.3748/wjg.v13.i37.4974

Dowman, J. K., Tomlinson, J. W., and Newsome, P. N. (2010). Pathogenesis of non-alcoholic fatty liver disease. *QJM* 103, 71–83. doi: 10.1093/qjmed/hcp158

Exil, V. J., Silva Avila, D., Benedetto, A., Exil, E. A., Adams, M. R., Au, C., et al. (2010). Stressed-induced TMEM135 protein is part of a conserved genetic network involved in fat storage and longevity regulation in *Caenorhabditis elegans*. *PLoS One* 5:e14228. doi: 10.1371/journal.pone.0014228

Fey, S. J., and Wrzesinski, K. (2012). Determination of drug toxicity using 3D speroids constructed from an immortal human hepatocyte cell line. *Toxicol. Sci.* 127, 403–411. doi: 10.1093/toxsci/kfs122

Fontana, R. J. (2008). Acute liver failure including acetaminophen overdose. *Med. Clin. North Am.* 92, 761–794. doi: 10.1016/j.mcna.2008.03.005

Fuchs, C. D., Claudel, T., and Trauner, M. (2014). Role of metabolic lipases and lipolytic metabolites in the pathogenesis of NAFLD. *Trends Endocrinol. Metab.* 25, 576–585. doi: 10.1016/j.tem.2014.08.001

Grimberg, M., Stöber, R. M., Edlund, K., Rempel, E., Godoy, P., Reif, R., et al. (2014). Toxicogenomics directory of chemically exposed human hepatocytes. *Arch. Toxicol.* 88, 2261–2287. doi: 10.1007/s00204-014-1400-x

Handen, A., and Ganapathiraju, M. K. (2015). LENS: web-based lens for enrichment and network studies of human proteins. *BMC Med. Genomics* 8:S2. doi: 10.1186/1755-8794-8-S4-S2

Hebels, D. G., Jetten, M. J., Aerts, H. J., Herwin, R., Theunissen, D. H., Gaj, S., et al. (2014). Evaluation of database-derived pathway development for enabling biomarker discovery for hepatotoxicity. *Biomark. Med.* 8, 185–200. doi: 10.2217/bmm.13.154

Hong, C., and Tontonoz, P. (2014). Liver X receptors in lipid metabolism: opportunities for drug discovery. *Nat. Rev. Drug Discov.* 13, 433–444. doi: 10.1038/nrd4280

Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., et al. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43, D921–D927. doi: 10.1093/nar/gku955

Ip, B. C., Liu, C., Lichtenstein, A. H., von Lintig, J., and Wang, X. D. (2015). Lycopene and apo-10'-lycopenoic acid have differential mechanisms of protection against hepatic steatosis in beta-carotene-9',10'-oxygenase knockout male mice. *J. Nutr.* 145, 268–276. doi: 10.3945/jn.114.200238

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Shcerf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Jozefczuk, J., Kashofer, K., Ummanni, R., Henjes, F., Rehman, S., Geenen, S., et al. (2012). A systems biology approach to deciphering the etiology of steatosis employing patient-derived dermal fibroblasts and Ips cells. *Front. Physiol.* 3:339. doi: 10.3389/fphys.2012.00339

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

Khatun, I., Clark, R. W., Vera, N. B., Kow, K., Erion, D. M., Coskran, T., et al. (2016). Characterization of a novel intestinal glycerol-3-phosphate acyltransferase pathway and its role in lipid homeostasis. *J. Biol. Chem.* 291, 2602–2615. doi: 10.1074/jbc.M115.683359

Kim, S. Y., and Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 6:144. doi: 10.1186/1471-2105-6-144

King, P. D., and Perry, M. C. (2001). Hepatotoxicity of chemotherapy. *Oncologist* 6, 162–176. doi: 10.1634/theoncologist.6-2-162

Krug, A. K., Kolde, R., Gaspar, J. A., Rempel, E., Balmer, N. V., Meganathan, K., et al. (2013). Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Arch. Toxicol.* 87, 123–143. doi: 10.1007/s00204-012-0967-3

Lake, A. D., Novak, P., Fisher, C. D., Jackson, J. P., Hardwick, R. N., Billheimer, D. D., et al. (2011). Analysis of global and absorption, distribution, metabolism, and elimination gene expression in the progressive stages of human nonalcoholic fatty liver disease. *Drug Metab. Dispos.* 39, 1954–1960. doi: 10.1124/dmd.111.040592

Lauschke, V. M., Vorrink, S. U., Moro, S. M., Rezayee, F., Nordling, A., Hendriks, D. F., et al. (2016). Massive rearrangements of cellular microRNA signatures are key drivers of hepatocyte dedifferentiation. *Hepatology* 64, 1743–1756. doi: 10.1002/hep.28780

Lawan, A., Zhang, L., Gatzke, F., Min, K., Jurczak, M. J., Al-Mutairi, M., et al. (2015). Hepatic mitogen-activated protein kinase phosphatase 1 selectively regulates glucose metabolism and energy homeostasis. *Mol. Cell. Biol.* 35, 26–40. doi: 10.1128/MCB.00503-14

Li, S., Brown, M. S., and Goldstein, J. L. (2010). Bifurcation of insulin signaling pathway in rat liver: mtorc1 required for stimulation of lipogenesis, but not inhibition of gluconeogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 107, 3441–3446. doi: 10.1073/pnas.0914798107

Liao, W., Liu, W., Yuan, Q., Liu, X., Ou, Y., He, S., et al. (2013). Silencing of DLGAP5 by siRNA significantly inhibits the proliferation and invasion of hepatocellular carcinoma cells. *PLoS One* 8:e80789. doi: 10.1371/journal.pone.0080789

Liu, Y., Mu, D., Chen, H., Li, D., Song, J., Zhong, Y., et al. (2016). Retinol-binding protein 4 induces hepatic mitochondrial dysfunction and promotes hepatic steatosis. *J. Clin. Endocrinol. Metab.* 101, 4338–4348. doi: 10.1210/jc.2016-1320

Lopez-Riera, M., Conde, I., Tolosa, L., Zaragoza, A., Castell, J. V., Gomez-Lechon, M. J., et al. (2017). New microRNA biomarkers for drug-induced steatosis and their potential to predict the contribution of drugs to non-alcoholic fatty liver disease. *Front. Pharmacol.* 8:3. doi: 10.3389/fphar.2017.00003

Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 29, 1830–1831. doi: 10.1093/bioinformatics/btt285

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 10:161. doi: 10.1186/1471-2105-10-161

Marchesini, G., Bugianesi, E., Forlani, G., Cerrelli, F., Lenzi, M., Manini, R., et al. (2003). Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. *Hepatology* 37, 917–923. doi: 10.1053/jhep.2003.50161

Megger, D. A., Naboulsi, W., Meyer, H. E., and Sitek, B. (2014). Proteome analyses of hepatocellular carcinoma. *J. Clin. Transl. Hepatol.* 2, 23–30. doi: 10.14218/JCTH.2013.00022

Morii, K., Nishisaka, M., Nakamura, S., Oda, T., Aoyama, Y., Yamamoto, T., et al. (2014). A case of synthetic oestrogen-induced autoimmune hepatitis with microvesicular steatosis. *J. Clin. Pharm. Ther.* 39, 573–576. doi: 10.1111/jcpt.12191

Newton, B. W., Russell, W. K., Russell, D. H., Ramaiah, S. K., and Jayaraman, A. (2009). Liver proteome analysis in a rodent model of alcoholic steatosis. *J. Proteome Res.* 8, 1663–1671. doi: 10.1021/pr800905w

Nueda, M. J., Tarazona, S., and Conesa, A. (2014). Next maSigPro: updating maSigPro Bioconductor package for RNA-seq time series. *Bioinformatics* 30, 2598–2602. doi: 10.1093/bioinformatics/btu333

Pacifici, F., Arriga, R., Sorice, G. P., Capuani, B., Scioli, M. G., Pastore, D., et al. (2014). Peroxiredoxin 6, a novel player in the pathogenesis of diabetes. *Diabetes Metab. Res. Rev.* 63, 3210–3220. doi: 10.2337/db14-0144

Pan, H., Qin, K., Guo, Z., Ma, Y., April, C., Gao, X., et al. (2014). Negative elongation factor controls energy homeostasis in cardiomyocytes. *Cell Rep.* 7, 79–85. doi: 10.1016/j.celrep.2014.02.028

Rabinowich, L., and Shibolet, O. (2015). Drug induced steatohepatitis: an uncommon culprit of a common disease. *Biomed Res. Int.* 2015:168905. doi: 10.1155/2015/168905

Rector, R. S., Thyfault, J. P., Wei, Y., and Ibdah, J. A. (2008). Non-alcoholic fatty liver disease and the metabolic syndrome: an update. *World J. Gastroenterol.* 14, 185–192.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). *Limma* powers differential expression analysis for RNA-sequencing and microarray studies. *Nucleic Acid Res.* 43:e47. doi: 10.1093/nar/gkv007

Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L. A., et al. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat. Genet.* 40, 1461–1465. doi: 10.1038/ng.257

Sahini, N., and Borlak, J. (2014). Recent insights into the molecular pathophysiology of lipid droplet formation in hepatocytes. *Prog. Lipid Res.* 54, 86–112. doi: 10.1016/j.plipres.2014.02.002

Sahini, N., Selvaraj, S., and Borlak, J. (2014). Whole genome transcript profiling of drug induced steatosis in rats reveals a gene signature predictive of outcome. *PLoS One* 9:e114085. doi: 10.1371/journal.pone.0114085

Schoonen, W. G., Kloks, C. P., Ploemen, J. P., Smit, M. J., Zandberg, P., Horbach, G. J., et al. (2007). Uniform procedure of (1)H NMR analysis of rat urine and toxicometabonomics Part II: comparison of NMR profiles for classification of hepatotoxicity. *Toxicol. Sci.* 98, 286–297. doi: 10.1093/toxsci/kfm077

Schumacher, J., and Guo, G. (2015). Mechanistic review of drug-induced steatohepatitis. *Toxicol. Appl. Pharmacol.* 289, 40–47. doi: 10.1016/j.taap.2015.08.022

Seiliez, I., Belghit, I., Gao, Y., Skiba-Cassy, S., Dias, K., Cluzeaud, M., et al. (2016). Looking at the metabolic consequences of the colchicine-based in vivo autophagic flux assay. *Autophagy* 12, 343–356. doi: 10.1080/15548627.2015.1117732

Shen, H. F., Shahzad, G., Jawairia, M., Bostick, R. M., and Mustacchia, P. (2014). Association between aspirin use and prevalence of nonalcoholic fatty liver disease: a cross-sectional study from the third national health and nutrition examination survey. *Am. J. Gastroenterol.* 109, S160–S160. doi: 10.1111/apt.12944

Spahis, S., Delvin, E., Borys, J., and Levy, E. (2017). Oxidative stress as a critical factor in nonalcoholic fatty liver disease pathogenesis. *Antioxid. Redox Signal.* 26, 519–541. doi: 10.1089/ars.2016.6776

Starmann, J., Fälth, M., Spindelböck, W., Lanz, K. L., Lackner, C., Zatloukal, K., et al. (2012). Gene expression profiling unravels cancer-related hepatic molecular signatures in steatohepatitis but not in steatosis. *PLoS One* 7:e46584. doi: 10.1371/journal.pone.0046584

Stein, S., Lemos, V., Xu, P., Demagny, H., Wang, X., Ryu, D., et al. (2017). Impaired SUMOylation of nuclear receptor LRH-1 promotes nonalcoholic fatty liver disease. *J. Clin. Invest.* 127, 583–592. doi: 10.1172/JCI85499

Sutherland, J. J., Jolly, R. A., Goldstein, K. M., and Stevens, J. L. (2016). Assessing concordance of drug-induced transcriptional response in rodent liver and

culture hepatocytes. *PLoS Comput. Biol.* 12:e1004847. doi: 10.1371/journal.pcbi.1004847

Tostmann, A., Boeree, M. J., Aarnoutse, R. E., de Lange, W. C., van der Ven, A. J., and Dekhuijzen, R. (2008). Antituberculosis drug-induced hepatotoxicity: concise up-to-date review. *J. Gastroenterol. Hepatol.* 23, 192–202. doi: 10.1111/j.1440-1746.2007.05207.x

Vinken, M. (2013). The adverse outcome pathway concept: a pragmatic tool in toxicology. *Toxicology* 312, 158–165. doi: 10.1016/j.tox.2013.08.011

Vitins, A. P., Kienhuis, A. S., Speksnijder, E. N., Roodbergen, M., Luijten, M., and van der Ven, L. T. (2014). Mechanisms of amiodarone and valproic acid induced liver steatosis in mouse in vivo act as a template for other hepatotoxicity models. *Arch. Toxicol.* 88, 1573–1588. doi: 10.1007/s00204-014-1211-0

Wang, Y., Tong, X., Omoregie, E. S., Liu, W., Meng, S., and Ye, X. (2012). Tetraspanin 6 (TSPAN6) negatively regulates retinoic acid-inducible gene I-like receptor-mediated immune signaling in a ubiquitination-dependent manner. *J. Biol. Chem.* 287, 34626–34634. doi: 10.1074/jbc.M112.390401

Waterfield, C. J., Turton, J. A., Scales, M. D., and Timbrell, J. A. (1993). Investigations into the effects of various hepatotoxic compounds on urinary and liver taurine levels in rats. *Arch. Toxicol.* 67, 244–254. doi: 10.1007/BF01974343

Xia, Z., Ying, G., Hansson, A. L., Karlsson, H., Xie, Y., Bergstrand, A., et al. (2000). Antidepressant-induced lipidosis with special reference to tricyclic compounds. *Prog. Neurobiol.* 60, 501–512. doi: 10.1016/S0301-0082(99)00036-2

Yu, J., Loh, K., Song, Z., Yang, H., and Lin, S. (2018). Update on glycerol-3-phosphate acyltransferases: the roles in the development of insulin resistance. *Nutr. Diabetes* 8:34. doi: 10.1038/s41387-018-0045-x

Zhang, J. J., Meng, X., Li, Y., Zhou, Y., Xu, D. P., Li, S., et al. (2017). Effects of melatonin on liver injuries and diseases. *Int. J. Mol. Sci.* 18:E673. doi: 10.3390/ijms18040673

Zhang, W. V., Ramzan, I., and Murray, M. (2007). Impaired microsomal oxidation of the atypical antipsychotic agent clozapine in hepatic steatosis. *J. Pharmacol. Exp. Ther.* 322, 770–777. doi: 10.1124/jpet.107.124024

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 September 2018
doi: 10.3389/fgene.2018.00412

Network, Transcriptomic and Genomic Features Differentiate Genes Relevant for Drug Response

*Janet Piñero[1], Abel Gonzalez-Perez[2], Emre Guney[1], Joaquim Aguirre-Plans[3], Ferran Sanz[1], Baldo Oliva[3] and Laura I. Furlong[1]**

[1] Integrative Biomedical Informatics Group, Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain, [2] Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, Spain, [3] Structural Bioinformatics Group, Research Programme on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain*

Understanding the mechanisms underlying drug therapeutic action and toxicity is crucial for the prevention and management of drug adverse reactions, and paves the way for a more efficient and rational drug design. The characterization of drug targets, drug metabolism proteins, and proteins associated to side effects according to their expression patterns, their tolerance to genomic variation and their role in cellular networks, is a necessary step in this direction. In this contribution, we hypothesize that different classes of proteins involved in the therapeutic effect of drugs and in their adverse effects have distinctive transcriptomics, genomics and network features. We explored the properties of these proteins within global and organ-specific interactomes, using multi-scale network features, evaluated their gene expression profiles in different organs and tissues, and assessed their tolerance to loss-of-function variants leveraging data from 60K subjects. We found that drug targets that mediate side effects are more central in cellular networks, more intolerant to loss-of-function variation, and show a wider breadth of tissue expression than targets not mediating side effects. In contrast, drug metabolizing enzymes and transporters are less central in the interactome, more tolerant to deleterious variants, and are more constrained in their tissue expression pattern. Our findings highlight distinctive features of proteins related to drug action, which could be applied to prioritize drugs with fewer probabilities of causing side effects.

Keywords: drug response, pharmacogenomics, adverse drug reaction, genomics, network biology, gene expression

OPEN ACCESS

Edited by:
Danyel Jennen,
Maastricht University, Netherlands

Reviewed by:
Xia Yang,
University of California, Los Angeles, United States
Francesco Russo,
University of Copenhagen, Denmark

****Correspondence:***
Laura I. Furlong
laura.furlong@upf.edu;
lfurlong@imim.es

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *05 July 2018*
Accepted: *05 September 2018*
Published: *25 September 2018*

Citation:
Piñero J, Gonzalez-Perez A, Guney E, Aguirre-Plans J, Sanz F, Oliva B and Furlong LI (2018) Network, Transcriptomic and Genomic Features Differentiate Genes Relevant for Drug Response. Front. Genet. 9:412. doi: 10.3389/fgene.2018.00412

INTRODUCTION

Drugs exert their effect acting at different scales of biological organization. At the cellular level, the effect of a drug is the result of its interaction with the target(s), which in time may lead to a variety of cellular responses, such as the alteration of the expression of a set of genes, changes in intracellular signaling pathways, or changes in the localization of proteins, that result in specific

Abbreviations: LoF, loss-of-function variants, including variants affecting splice sites, or stop codons; METAB, proteins that are involved in the drug metabolism, absorption, distribution, metabolism, and excretion; OT, drug targets that do not mediate side effects; OTP, proteins associated to side effects that are not drug targets; TARGET, drug targets; TOXPROT, proteins associated to side effects; TT, drug targets that mediate side effects.

cell phenotypic responses. At the organism level, drug absorption, distribution, metabolism, and excretion (ADME) also contribute to modulate the response to the drug. Nevertheless, our understanding of the molecular events elicited by drugs, which result on their therapeutic effects or adverse reactions, is still very limited.

The response to drug treatment is also influenced by the genetic background of an individual (Madian et al., 2012). Nowadays, for some drugs, the impact of genetic variability is well established. More than 200 FDA approved drugs include pharmacogenomic labeling (US Food and Drug Administration[1]), and pharmacogenomic screenings for known biomarkers are routinely carried out in large hospitals (Roden et al., 2011; van der Wouden et al., 2017; Weinshilboum and Wang, 2017). In particular, the genomic variation of genes involved in drug metabolism and its impact on drug response has been extensively studied (Shenfield, 2004; Pinto and Dolan, 2011; Kozyra et al., 2017) (for recent reviews see Ahmed et al., 2016; Lauschke et al., 2018), Nevertheless, only few studies have probed the role of the genomic variability of drug targets. The results of these studies imply that there is a high frequency of variants impacting protein function in drug targets (Schärfe et al., 2017), pharmacogenes (Wright et al., 2018) and GPCRs (Hauser et al., 2018) in the population. In spite of these studies, we still lack a detailed characterization of the genomic variation of the full spectrum of genes relevant for drug response, including drug targets, ADME genes and genes associated to the side effects of drugs, and their impact on drug response phenotypes.

In the field of systems pharmacology, the study of the perturbations elicited by drugs within the context of cellular networks has provided insight into the molecular mechanisms leading to drug action, including their adverse reactions (Berger and Iyengar, 2011). Network analysis of omics data has been used to identify modules associated with drug response and toxicity (Berger et al., 2010; Bauer-Mehren et al., 2012), to characterize the therapeutic (Yildirim et al., 2007; Guney et al., 2016) and adverse effect of drugs (Guney, 2017), and to explain the similarity of side effects of different drugs (Brouwers et al., 2011).

A key goal of network analysis is to connect network structure to function. For example, multi-scale network analysis allowed distinction of different classes of disease genes based on their connectivity patterns in the human protein-protein interaction network, or interactome (Berenstein et al., 2015; Piñero et al., 2016a). The multi-scale network analysis involves the exploration of the network properties of the proteins at *local*, *meso* and *global* scales. *Local* properties of a protein in a network pertain to its direct interactions with other nodes (**Figure 1**). Examples of local properties are the degree of a node (the number of direct neighbors), or the clustering coefficient (the density of links in the node's immediate neighborhood). *Global* properties consider the links across the whole network. An example is the betweenness centrality (the proportion of shortest paths passing through a node in a network). Finally, the *meso-scale* network properties are related to the organization of the network into clusters or modules, that represent functional units in the cell (Hartwell et al., 1999). Exploring the connectivity of proteins at the meso-scale level can shed light on the modular organization of the interactome, potentially revealing the regulation of cellular processes.

Here we provide a comprehensive characterization of genomic, transcriptomic and network topological features of genes relevant to drug response. We carefully selected three sets of proteins relevant to pharmacokinetics and pharmacodynamics: drug targets, proteins associated to phenotypes of drug toxicity, and proteins involved in the transport and metabolism of drugs. By leveraging on data from large scale genomic and transcriptomic initiatives and the reconstructions of the human protein interactome, we characterized the tolerance to deleterious genomic variability across human populations, the multi-scale network properties, and the expression across human tissues of proteins involved in the therapeutic and toxic response to drugs.

MATERIALS AND METHODS

The Data

Drug Targets (TARGET)

We compiled a comprehensive set of drug target proteins (referred as TARGET hereafter) that mediates the therapeutic effects of the drugs by integrating data from several repositories: DrugBank, version 5.0.7 (Wishart et al., 2018), DrugCentral, data downloaded on September, 2017 (Ursu et al., 2017), DGIdb, version 3.0 (Cotto et al., 2017), and ChEMBL, version 23 (Bento et al., 2014). We then mapped all the drugs to DrugBank identifiers, and all proteins to NCBI Gene identifiers. From DrugBank, we included only targets for approved or investigational drugs. From DrugCentral, we kept only targets in the Tclin category. From DGIdb we considered drug-target associations from "ChEMBL," "GuideToPharmacology," "Tdg Clinical Trial," "FDA," "TEND," and "TTD." From ChEMBL, we kept the drug-target relationships for which we could find a corresponding DrugBank identifier. Finally, we removed any protein present in the METAB set (see below). The TARGET set was composed of 1,934 proteins, targeting 2,829 drugs (**Figure 2**).

Drug Carriers, Transporters and Metabolism Enzymes (METAB)

We retrieved the proteins that act as drug transporters, drug carriers, and enzymes involved in drug absorption, distribution, and metabolism from DrugBank. We mapped all proteins to NCBI Gene identifiers. We thus obtained the METAB set, composed of 470 proteins involved in the transport and metabolism of 1,519 drugs (**Figure 2**).

Proteins Associated to Drug Toxicity (TOXPROT)

We assembled a set of proteins associated to the side effects or toxicity phenotypes of the drugs included in this study. To do this, we first collected drug side effects, and drug therapeutic indications. The therapeutic indications were obtained from SIDER, version 4.1 (Kuhn et al., 2016), AEOLUS (Shah, 2016), CTD, revision 15142 (Davis et al., 2017), repoDB, version 1.2

[1]https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm

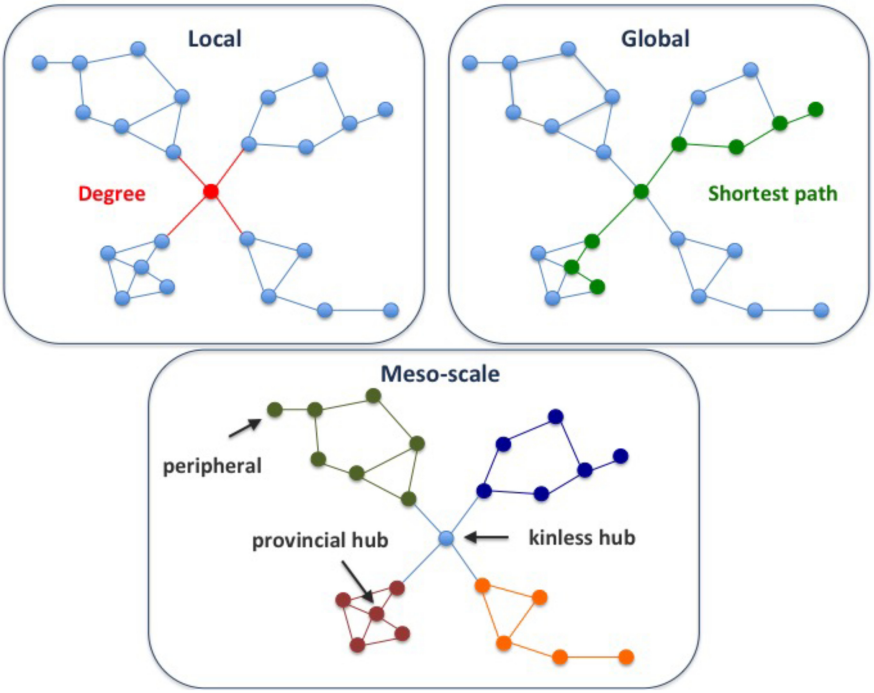


FIGURE 1 | Multi-scale network properties and cartographic roles. The multi-scale network analysis involves exploring the network properties of the proteins at different scales, namely local, meso-scale, and global. The degree of a node is a *local* network property, since it considers the first direct neighbors of a node, while the shortest path is a *global* network property, since we need to count paths between pairs of nodes across the whole network. Finally, the *meso-scale* network properties represent the organization of the network into clusters or modules. The meso-scale connectivity features of each protein can be characterized with the cartographic role classification scheme proposed by Guimerà and Amaral, 2005, namely ultra-peripheral, peripheral, non-hub connector, non-hub kinless, provincial hubs, connector hubs and kinless hubs (see **Supplementary Figure S1** for more information). Thus, focusing on how individual nodes are positioned in the modular (meso-scale) structure of the network, we can identify proteins that play different functions, such as mainly connected to other proteins within their modules (e.g., provincial hub), and those proteins that serve as bridges between modules (e.g., kinless hub).

(Brown and Patel, 2017), and ChEMBL, version 23. We mapped drugs to DrugBank identifiers, and disease identifiers to the Unified Medical Language System (UMLS, version 2016AB) Concept Unique Identifiers (CUIs) (Bodenreider, 2004). We only kept therapeutic indications reported by more than one source. The data of Adverse Drug Reaction (ADRs) was retrieved from 3 sources: Offsides (Tatonetti et al., 2012), AEOLUS, and ORGANDB (Mannil et al., 2015) (all files were downloaded on September, 2017). As we did with drug therapeutic indication data, we used UMLS CUIs to harmonize phenotypes and DrugBank identifiers to represent drugs.

Next, we filtered out phenotypes annotated to the UMLS semantic types "Patient or Disabled Group," "Professional or Occupational Group," "Therapeutic or Preventive Procedure," "Medical Device." To produce a high confidence dataset, we only kept associations reported by the three sources (Offsides, AEOLUS, and ORGANDB). From this set of drug-ADRs we removed the phenotypes/diseases that overlapped with the therapeutic indications of drugs. This produced a list of 12,213 drug-side effects pairs involving 593 drugs and 718 side effects. Finally, we used DisGeNET Curated (version 5.0) (Piñero et al., 2016b) to obtain a list of 4,160 genes associated to 452 ADRs, which we refer as TOXPROT throughout the text (**Figure 2**).

Due to the overlap between the TARGET and TOXPROT sets of proteins (see **Figure 3**) we separately assessed the properties of the overlapping subset of genes (TT), the genes annotated uniquely as drug targets (OT), and those annotated only as associated to drugs toxicity (OTP).

TARGET, TOXPROT, and METAB Protein Classes

We used data from Pharos, version 4.6.2 (Nguyen et al., 2017) to classify the drug targets in seven categories: GPCR, Transcription Factor, Enzyme, Kinase, Transporter, Ion Channel, and Nuclear Receptor (NHR). We extended this classification to the TOXPROT using the equivalent terms from the classification from Panther database, version 13.0 (Mi et al., 2017) in the file[2]. For METAB, we used the classification provided by DrugBank: transporters, carriers, enzymes.

The Network Analysis

Protein Interaction Data

We built two high-confidence protein-protein interaction networks (PIN) using data from INBIOMAP (Li et al., 2017) and from HIPPIE (Alanis-Lobato et al., 2017), two resources that

[2]ftp://ftp.pantherdb.org/sequence_classifications/current_release/PANTHER_Sequence_Classification_files/PTHR13.1_human

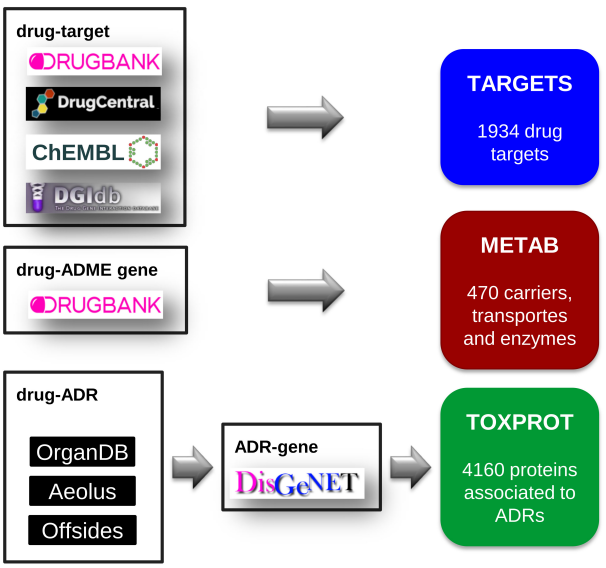


FIGURE 2 | The assembly of the three sets of proteins relevant for drug response: the drug target set (TARGET), the proteins involved in drug transport and metabolism (METAB), and the proteins associated to side effects (TOXPROT).

integrate information from several other sources, and provide a reliability score that allows to filter the interactions. To build the INBIOMAP network, we downloaded the file[3] (version 2016_05_31). We removed predicted interactions, and kept only interactions with score greater than 0.15. In the case of the HIPPIE-based network, we downloaded the file[4] (version 2.1). To produce a high confidence network, we filtered out all interactions with score smaller than or equal to 0.7 (keeping ∼25% of HIPPIE). From both PINs, to obtain a biologically meaningful modular representation of the network, we removed genes with degree higher than 300, such as chaperones and ubiquitins.

We also compiled 4 organ-specific interactomes using GTEx data (version 7.0) for brain, liver, kidney and heart. Briefly, we first mapped the ENSEMBL gene identifiers in the GTEx expression matrix to NCBI Gene identifiers. In the cases of brain, and heart, we merged the gene expression of different zones, and computed the median value of expression for each gene (Melé et al., 2015). Then, we removed from the PINs all interactions involving at least one gene with TPM < 1 in the corresponding tissue.

Network Cartographic Roles

To assign cartographic roles in the PINs to each protein, we computed the z (within-module degree) and P (participation coefficient) of each gene following the protocol described in Piñero et al. (2016a). Briefly, we clustered the PINs using the Infomap algorithm (Rosvall and Bergstrom, 2008) and calculated z and P using equations (1) and (2), respectively.

$$Z_i = \frac{k_i - \bar{k}_{ci}}{\sigma \bar{k}_{ci}} \quad (1)$$

where k_i is the number of links of node i to other nodes in its module, $\bar{k}_{ci}$ is the mean degree of all nodes in cluster c_i, and $\sigma \bar{k}_{ci}$ is the standard deviation of the degree of the nodes in the cluster c_i

$$P_i = 1 - \sum_{c=1}^{M} \left[\frac{k_{ic}}{k_i} \right]^2$$

where k_{ic} the number of links of node i to nodes in the cluster c, k_i is the total degree of node I, and M is the total number of modules in the network.

According to Guimerà and Amaral (2005) the genes were assigned to one of the following roles: ultra-peripheral nodes, peripheral, non-hub connector, non-hub kinless, provincial hubs, connector hubs, kinless hubs. These seven different roles are heuristically defined, using their localization in the different regions of the z–P parameter space (see **Supplementary Figure S1**). Nodes with $z > 2.5$ are classified as module hubs and nodes with $z < 2.5$ as non-hubs. Both hub and non-hub nodes are then further characterized by using their participation coefficient. Non-hub nodes can be divided into four different roles: ultra-peripheral nodes; that is, nodes with all their links within their module ($P \leq 0.05$); peripheral nodes; that is, nodes with most links within their module ($0.05 < P \leq 0.62$); non-hub connector nodes; that is, nodes with many links to other modules ($0.62 < P \leq 0.80$); and non-hub kinless nodes; that is, nodes with links homogeneously distributed among all modules ($P > 0.80$). Similarly, hub nodes are assigned to: provincial hubs; that is, hub nodes with the vast majority of links within their module ($P \leq 0.30$); connector hubs; that is, hubs with many links to most of the other modules ($0.30 < P \leq 0.75$); and kinless hubs; that is, hubs with links homogeneously distributed among all modules ($P > 0.75$).

The Analysis of Genomic Features

We used the data on germline variants detected across 60,706 exomes from the Exome Aggregation Consortium, version 0.3.1(Lek et al., 2016). To evaluate the tolerance of different sets of genes to variants in the human germline, we downloaded the data of Functional Gene Constraint[5].

Specifically, for each human protein coding gene, we obtained the pLI (defined as the probability of being loss-of-function intolerant, including both heterozygous and homozygous LoF variants), and the pNull (defined as the probability of being tolerant to both heterozygous and homozygous LoF variants).

[3]https://www.intomics.com/inbio/map/#downloads

[4]http://cbdm-01.zdv.uni-mainz.de/∼mschaefer/hippie/hippie_current.txt

[5]ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint/README_fordist_cleaned_exac_r03_z_data_pLI_2016_01_13.txt

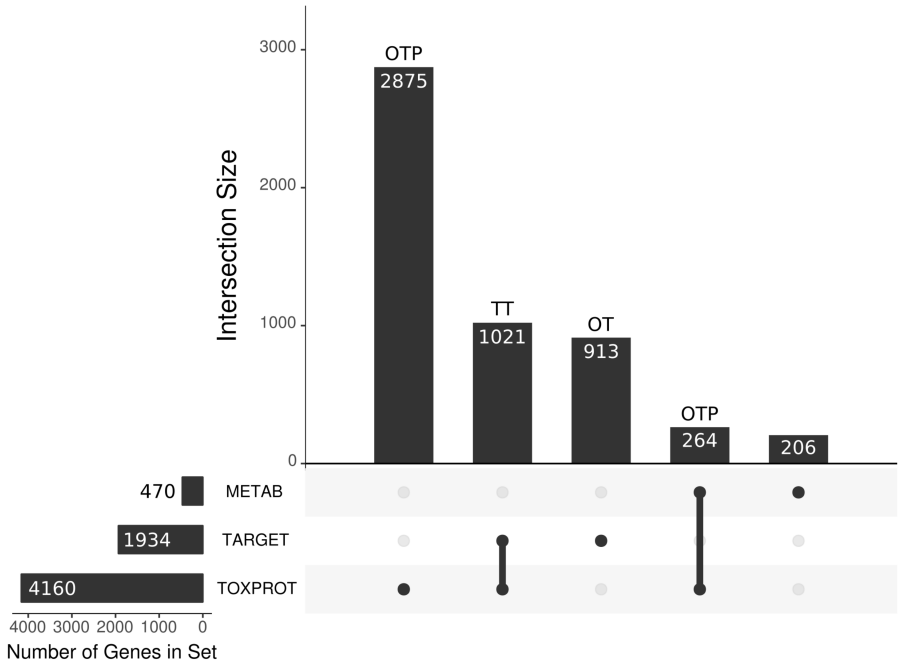


FIGURE 3 | We present the main sets of proteins under study, and their overlaps. The horizontal bars on the left correspond to the three larger categories: the METAB set, constituted of 470 proteins that are involved in drug absorption, distribution, metabolism, and excretion; the drug target set (TARGET), composed of 1,930 proteins; and the set of 4,160 proteins associated to side effects (TOXPROT); the intersections among these categories are represented with the filled dots in the matrix. The TARGET set shares 1,021 proteins with the TOXPROT set and we refer to those drug targets associated toxicity as the TT set. The OT (only targets) set are the TARGET proteins that are not included in the TOXPROT set (913 proteins). The only TOXPROT proteins (OTP) set is composed by 3,139 TOXPROT proteins that are not in the TARGET set. In the figure, it corresponds to the two bars marked as OTP.

Gene Expression Data in Healthy Tissues/Organs Across Individuals

We used gene expression data from GTEX (version 7.0) to analyze the pattern of expression of the different sets of genes. For GTEX tissues, we mapped the ENSEMBL gene identifiers to NCBI Gene identifiers, and kept the genes with TPM > = 1. We used the information for 53 tissues in GTEX, which represent all tissues covered except Cells.EBV.transformed.lymphocytes and Cells.Transformed.fibroblasts.

Statistical Analysis

To compute the deviation of the value of each network feature for each set of genes, we randomly sampled 10,000 sets of genes from the network of the same size of the set under analysis in each case. Then, we computed the mean value of each sampled feature (degree, betweenness, clustering coefficient, participation coefficient, and within-module degree) for each of the 10,000 randomly sampled gene sets. From this distribution of means a z-score was estimated for every gene set for every feature. The same was done to compute the deviation of the value of the genomic features and of the expression of each gene set from their expected distribution, but genes were sampled from the entire list of human protein coding genes in this case. The statistical analysis were carried out using R, version 3.4.0 (R Core Team, 2017) and the network analysis were performed using the iGraph Library, version igraph_1.1.2 (Csardi and Nepusz, 2006). Additionally, the following packages were employed: UpSetR_1.3.3 (Conway et al., 2017), to produce **Figure 3**, showing the intersects between the different protein sets evaluated in the paper, and clusterProfiler_3.6.0 (Yu et al., 2012) to compare the pathways in which are involved the proteins in the class "enzyme" in the TARGET set and in the METAB set.

RESULTS

More Than Half of the Proteins That Are Targets of Drugs, or Involved in Drug Metabolism Are Associated to Side Effects

We compiled three sets of drug-associated proteins as detailed in Methods (**Figure 2**). The first comprised 1,934 proteins that are well-established drug targets (TARGET set); the second comprised 470 proteins involved in drug transport and metabolism (METAB set); the third was composed of 4,160 proteins associated to Adverse Drug Reactions, or ADRs (TOXPROT set). Twenty-five percent of the proteins in the TOXPROT set are also targets of drugs (TOXPROT-TARGET, TT set). More than half of drug target proteins and of proteins involved in drug metabolism are associated to side effects (**Figure 3**).

The Distribution of Cartographic Roles Is Preserved Across Organ-Specific Interactomes

For this study, we assembled two different global human interactomes and several organ-specific interactomes from two different resources, INBIOMAP (Li et al., 2017) and HIPPIE (Alanis-Lobato et al., 2017). We focused on brain, heart, kidney, and liver due to the relevance of these organs in drug toxicity. Throughout the paper, we illustrate the results obtained with the INBIOMAP interactomes, but all analyses were replicated in the HIPPIE-based interactomes.

The INBIOMAP global interactome is composed of 12,967 nodes and 107,787 edges. The number of proteins in organ-specific interactomes varies between 8,800 and 9,800 nodes, and the final networks contain around 80% of the interactions of the global interactome (**Supplementary Table S1**).

To uncover the modular organization of the global human interactome and the organ-specific interactomes, we employed the Infomap procedure (Rosvall and Bergstrom, 2008), one of the best performing network community recognition methodologies, which has produced biologically relevant partitions of the human interactome (Berenstein et al., 2015). After partitioning the interactome into modules, we characterized the meso-scale connectivity features for each protein in the network using the within-module degree (z) and the participation coefficient (P) parameters (Guimerà and Amaral, 2005). The z parameter standardizes the degree of a node in relation with the degree of nodes that belong to the same cluster, and the P parameter quantifies the fraction of links that a given node projects to other clusters. We further categorized each network node according to the universal cartographic role classification scheme proposed by Guimerà and Amaral (2005): ultra-peripheral, peripheral, non-hub connector, non-hub kinless, provincial hubs, connector hubs and kinless hubs (**Supplementary Figure S2**). Thus, focusing on how individual nodes are positioned in the modular (meso-scale) structure of the network, we can identify proteins that play different functions, such as those only connected to proteins within their modules, and those proteins that serve as bridges between different modules.

The cartographic analysis of the global human interactome and four organ-specific interactomes (brain, heart, kidney and liver) is shown in **Supplementary Figures S1**, **S2**, respectively, and summarized in **Table 1** and **Supplementary Table S2**. Most of the proteins in the global network have roles with within-module degree smaller than 2.5, that is kinless (14.7%), connector (28.4%), peripheral (27.5%) and ultra-peripheral (26.6%). Nodes with hub roles account for 2.8% of the network. The nodes with the higher z, also have high P, resulting in their classification as connector hub or kinless hub nodes. This distribution of genes across cartographic roles is preserved in the organ-specific networks (**Table 1**). In other words, the proportion of nodes with different roles in the network does not change substantially when we take into account only the genes expressed in each tissue to construct the networks, although there is a small decrease in the percentage of nodes in the ultra-peripheral role in organ-specific networks. A similar behavior is observed for the HIPPIE global interactome, and its organ-specific interactomes (**Supplementary Table S2**). Taken together, these findings point to a conserved network structure and connectivity patterns at the meso-scale level in the interactome across tissues.

Targets That Mediate Side Effects and Side Effect Proteins Are Important for Connecting Different Modules in the Network

Next, we studied the multi-scale network properties of the sets of genes relevant for drug response within the context of the global and the organ-specific interactomes. The coverage for the different gene sets in the interactomes varies between 70 and 90% in the INBIOMAP global interactome (**Supplementary Table S3**). Eighty percent of the TARGET and TOXPROT sets are present in the global networks, while METAB proteins coverage is the lowest (around 70%). The coverage in the organ-specific networks ranges between 50 and 60% depending on the protein set and the tissue. Similar coverage is observed for the HIPPIE-based interactomes (**Supplementary Table S3**).

The analysis of the proteins belonging to each set according to their cartographic role showed that TARGET proteins are significantly enriched for nodes that play kinless and kinless hub roles in the network (**Table 2**). The enrichment of TOXPROT proteins is more apparent for nodes of the network that play kinless, kinless hub and marginally connector roles. As a matter of fact, the overrepresentation of targets in the kinless and kinless hub nodes is almost completely explained by the subset of TT amongst them (targets that are associated to side effects).

TABLE 1 | Cartographic partition of the nodes in the INBIOMAP interactomes (the global interactome, and the four organ-specific PINs).

Cartographic role	Global	Brain	Heart	Kidney	Liver
Provincial hub	9 (0.07%)	5 (0.05%)	6 (0.06%)	8 (0.08%)	6 (0.07%)
Connector hub	119 (0.92%)	68 (0.69%)	72 (0.78%)	67 (0.69%)	73 (0.82%)
Kinless hub	236 (1.82%)	216 (2.2%)	196 (2.12%)	202 (2.07%)	187 (2.1%)
Kinless	1903 (14.68%)	1718 (17.52%)	1560 (16.88%)	1603 (16.43%)	1524 (17.13%)
Connector	3686 (28.43%)	2891 (29.48%)	2808 (30.38%)	2911 (29.84%)	2692 (30.26%)
Peripheral	3570 (27.53%)	2604 (26.56%)	2528 (27.35%)	2660 (27.27%)	2403 (27.01%)
Ultra-peripheral	3444 (26.56%)	2303 (23.49%)	2074 (22.44%)	2303 (23.61%)	2011 (22.61%)

The number of nodes in each cartographic role for each network is shown, with its percentage between parentheses.

TABLE 2 | Enrichment analysis of the cartographic roles of each set of genes in the INBIOMAP global interactome.

Cartographic role	TARGET	TT	OT	METAB	TOXPROT	OTP
Kinless hub	3.4 (1.2*e* – 14)	4.2 (8.7*e* – 15)	1.7 (0.09)	0.66 (1.0)	2.6 (1.1*e* – 11)	1.3 (1.1*e* – 01)
Connector hub	1.4 (2.2*e* – 01)	2.2 (1.9*e* – 02)	0.46 (1.0)	0.32 (1.0)	1.4 (1.0*e* – 01)	1 (7.8*e* – 01)
Provincial hub	3.5 (2.0*e* – 01)	1.6 (7.9*e* – 01)	5.1 (0.19)	0 (1.0)	1.4 (7.8*e* – 01)	1.2 (8.5*e* – 01)
Kinless	1.7 (1.5*e* – 13)	1.9 (1.7*e* – 12)	1.3 (0.02)	1 (0.78)	2 (3.5*e* – 40)	1.8 (3.8*e* – 23)
Connector	1.1 (1.0*e* – 01)	1.1 (1.1*e* – 01)	1.1 (0.54)	0.86 (1.0)	1.2 (1.7*e* – 03)	1.1 (1.6*e* – 02)
Peripheral	0.85 (1.0)	0.74 (1.0)	1 (0.6736)	0.93 (1.0)	0.8 (1.0)	0.86 (1.0)
Ultra-peripheral	0.55 (1.0)	0.49 (1.0)	0.69 (1.0)	1.3 (0.06)	0.55 (1.0)	0.62 (1.0)

The fold enrichment of the Fisher's exact test is shown with the corresponding p-value corrected by the Benjamini and Hochberg method between parentheses. The cartographic partition of the different gene sets in the INBIOMAP and HIPPIE interactomes is provided in ***Supplementary Table S3****.*

METAB proteins are not particularly enriched in any role in the network. The results are similar in organ-specific interactomes (**Supplementary Figure S3**) and for networks derived from HIPPIE, except for the case of METAB proteins, which play peripheral roles in the global and the liver HIPPIE PINs (**Supplementary Figure S3**). The distribution of the gene sets across the seven cartographic roles is shown in **Supplementary Table S4**.

A more detailed analysis of other network properties of the sets of genes shows that TARGET, TOXPROT, and TT sets tend to have a significantly higher degree, participation coefficient, within-module degree, and betweenness than the other genes in the network (**Figure 4**). They also have a lower clustering coefficient. We note, however, that most of the effect observed for the TARGET and TOXPROT sets is explained by their shared TT subset. On the other hand, METAB proteins have significantly lower degree, and within module degree than expected and significantly lower participation coefficient in most organ-specific interactomes (**Figure 4**). METAB proteins are more specialized, thus it would make sense that they play less central roles in the network, with less interaction partners in the interactome. Similar results are obtained for HIPPIE interactomes (**Supplementary Figure S4**).

Drug Targets, and Toxicity Proteins Are Highly Sensitive to Loss of Function Mutations, While Proteins Involved in Drug Metabolism Are Tolerant

Next, we analyzed the tolerance of drug related proteins to LoF variants using exome sequence data from 60K "healthy" subjects provided by the ExAC consortium (Lek et al., 2016). We employed two gene constraint metrics developed by the ExAC team: pLI and pNull. pLI is the probability of a gene to be intolerant to heterozygous LoF mutations (LoF variants are nonsense and essential splice site variants). It separates genes into LoF intolerant (pLI $\geq$ 0.9) or LoF tolerant (pLI $\leq$ 0.1). On the other hand, pNull is the probability of a gene to be tolerant to both heterozigous and homozigous LoF variation.

We found that METAB genes have significantly lower pLI than the other genes in the genome (**Table 3**). In other words, METAB genes are more tolerant to LoF variation than the average human genes. On the other hand, TARGET and TOXPROT genes have significantly greater pLI value than average human genes. Since genes intolerant to LoF variation are likely to be dosage sensitive (Lek et al., 2016), TARGET and TOXPROT sets might contain haploinsufficient genes. The results for the pNull are consistent with those of the pLI, but with the opposite meaning: genes with high pNull are tolerant to LoF variation.

In order to explore in more detail the features of different classes of TARGET proteins, we classified them using categories from the drug target ontology (Lin et al., 2017) (**Figure 5**). We used similar categories to classify the TOXPROT set (for more details see Methods section), and we classified METAB genes into carriers, enzymes, and transporters using the information from DrugBank. We found that among METAB genes, enzymes display the lowest pLI and highest pNull, while carriers and transporters are not significantly different than expected in terms of pLI (**Figure 5**). In the TARGET set, kinases are the most intolerant subset to LoF variation, with a mean value more than 12 SD greater than the expected mean pLI value, followed by transcription factors (*z*-score = 9.04) and TARGET enzymes (*z*-score = 7.51). It is worth noting that the enzymes within the TARGET set are related to signaling pathways, and core cellular metabolic processes, while the enzymes in the METAB set are proteins mainly participating in the metabolism of xenobiotics (**Supplementary Figure S5**). The remaining groups of TARGET genes are also intolerant to LoF but to a lesser extent, with the exception of GPCRs, that are more tolerant to LoF variation than expected (**Figure 5**). Again, the results for pNull are consistent with those of pLI, except for the case of ion channels, which are marginally intolerant to LoF variation, but do not show differences with the rest of the genes with respect to pNull.

Within the TOXPROT set, transcription factors exhibited the highest intolerance to LoF variants, followed by kinases. Nevertheless, the enzymes were not different than the rest of the genes, but they do show a lower pNull, indicating that they are less tolerant to LoF than the background.

Proteins Associated to Side Effects Are Highly Expressed Across Tissues

Next, we characterized the expression patterns of each gene set across normal human tissues, using GTEx data (**Figure 6**). TOXPROT genes are more expressed than other genes in the genome across all tissues, with the exception of some areas of the

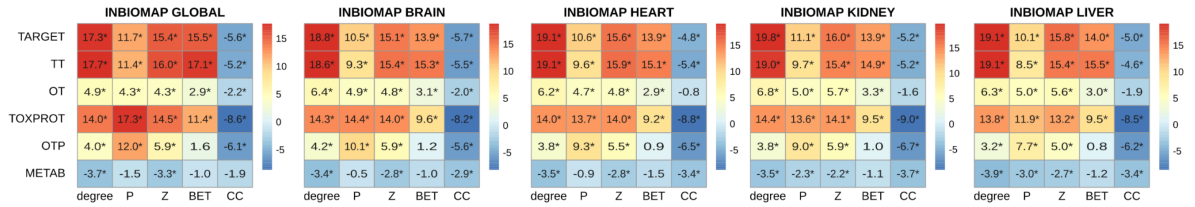


FIGURE 4 | Multi-scale network features of the gene sets (INBIOMAP interactomes). We plot the *z*-score of the network features (degree, P: participation coefficient, Z: within-module degree, BET: betweenness, CC: clustering coefficient) resulting from 10,000 randomizations. The asterisks indicate that the *z*-score is statically significant (p-value < 0.05). TARGET: drug targets, METAB: proteins that are involved in the drug metabolism, absorption, distribution, metabolism, and excretion. TOXPROT: proteins associated to side effects. TT: genes in common between drug targets and toxicity genes. OT: only TARGET proteins. OTP: only TOXPROT proteins.

TABLE 3 | Genomic features of the sets of genes.

Gene set	pLI			pNull		
	pLI	***z*-score**	***p*-value**	**pNull**	***z*-score**	***p*-value**
TARGET	0.380	8.87	7.31E-19	0.167	−5.829	5.58E-09
TOXPROT	0.365	10.9	1.15E-27	0.146	−13.604	3.79E-42
METAB	0.214	−4.92	8.65E-07	0.260	3.843	1.22E-04
TT	0.399	7.84	4.51E-15	0.153	−5.607	2.06E-08
OT	0.358	4.15	3.32E-05	0.183	−2.239	0.0251559
OTP	0.354	7.45	9.33E-14	0.143	−12.108	9.58E-34

We show the value of the feature, the z-score resulting from 10,000 randomizations, and its associated two-sided p-value.

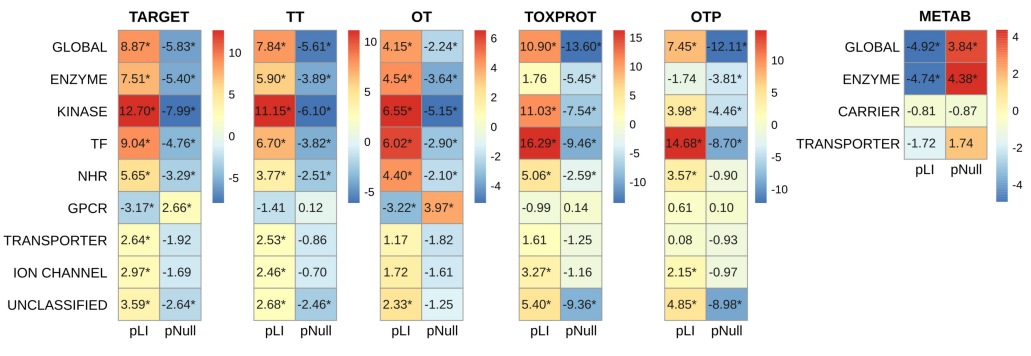


FIGURE 5 | Genomic features of the sets of genes, and of the protein classes. We plot the *z*-score resulting from 10,000 randomizations. The asterisks indicate that the *z*-score is statically significant (p-value < 0.05). TARGET: drug targets, METAB: proteins that are involved in the drug metabolism, absorption, distribution, metabolism, and excretion. TOXPROT: proteins associated to side effects. TT: genes in common between drug targets and toxicity genes. OT: only TARGET proteins. OTP: only TOXPROT proteins. See **Figure 1** for more details. It is worth noting that while 90% genes from the total of 1,934 TARGET genes are classified in the 7 target classes, 40% of the TOXPROT set belongs to other protein classes (Unclassified set).

brain (that do not show statistically significant differences with the other genes), and the testis (where they show a lower level of expression than other genes of the genome). TARGETs also tend to be more highly expressed than other genes of the genome across most tissues. The tissues with the most significantly higher expression are blood, lung, spleen, liver, adipose tissue, and heart. Drug targets are not significantly over or under expressed in any brain area. A closer look at this set shows that TT tend to behave like the TARGET set, with the exception of few brain areas, such as cerebellum and cerebellar hemisphere. On the other hand, OTs are not expressed at higher levels than other genes of the genome, with very few exceptions, which exhibit marginal significance. Probably, drug targets that are expressed in more tissues throughout the body, at higher levels than the rest of the proteins are more likely to elicit side effects. The broader the expression of the target, the higher is the risk of adverse reactions when the drug is administered systemically (Gashaw et al., 2011).

As expected, metabolic enzymes exhibited most significantly higher expression in liver, and to some extent in kidney, but they tend to show significantly lower expression in most tissues. Interestingly, the levels of expression of all sets of proteins (except OTP) in testis are significantly lower than the other genes in the genome.

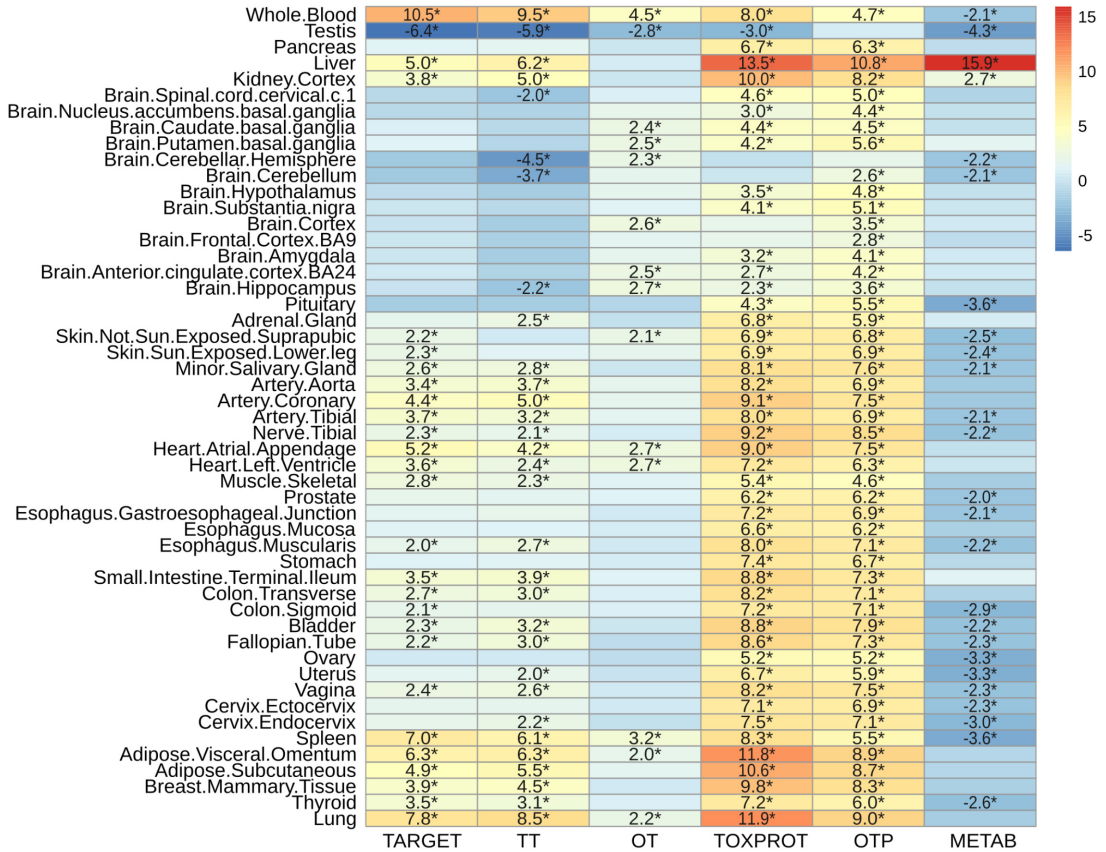


FIGURE 6 | Characterization of the expression level by tissue of the sets of genes. We plot the *z*-score resulting from 10,000 randomizations. The asterisks indicate that the *z*-score is statically significant (p-value < 0.05). TARGET: drug targets, METAB: proteins that are involved in the drug metabolism, absorption, distribution, metabolism, and excretion. TOXPROT: proteins associated to side effects. TT: genes in common between drug targets and toxicity genes. OT: only TARGET proteins. OTP: only TOXPROT proteins.

DISCUSSION

There is a pressing need to identify "good" targets for safer drug development, patient treatment, and better management of drug toxicity. In this contribution, we propose that leveraging large scale genomic, transcriptomic, and interactomic data can support this goal. We show that drug targets, targets associated to side effects, and proteins associated to side effect display higher level of centrality measures (degree, betweenness, z, P and lower clustering coefficient) in the protein interaction network, indicating that they occupy central positions in the network. This centrality is evidenced not only at the local network level (evidenced by the degree and clustering coefficient), but more importantly, they are key nodes for connecting different modules in the network. On the other side, proteins participating in drug metabolism are characterized by lower degree and within-module degree, and their role is confined to their own modules in the network.

We then assessed how the observed differences in these network properties are related to the tolerance of each set of genes to LoF variation. We observed that genes which play more central roles in the network exhibit significantly lower tolerance to LoF variants, indicating that they are under stronger purifying selection. These results are in agreement with observations across genes related to different disease classes (Piñero et al., 2016a). On detail, we found that disease genes that play central roles in the network, such as cancer genes and genes associated to autosomal dominant diseases show less tolerance to likely deleterious variants than genes associated to autosomal recessive diseases, which play peripheral roles in the network (Piñero et al., 2016a). **Figure 7** shows the separation between METAB genes, and TARGET and TOXPROT in terms of the gene constraint metrics (pLI and pNull) and the multi-scale network features.

We found that TT genes are more central in the network, (indicated by their higher z, degree and P, and lower clustering coefficient), than OT genes. In particular, the observed higher P indicates that these proteins play an important role in connecting different modules within the network, suggesting that they are pleiotropic and participate in diverse biological processes, which could explain why they are mediators of both, therapeutic and side effects of drugs. These results are in line with those of Perez-Lopez et al. (2015) who showed that drug targets that mediate side effects are better spreaders of perturbations in a human global interactome, than targets of drugs having no reported side effects,

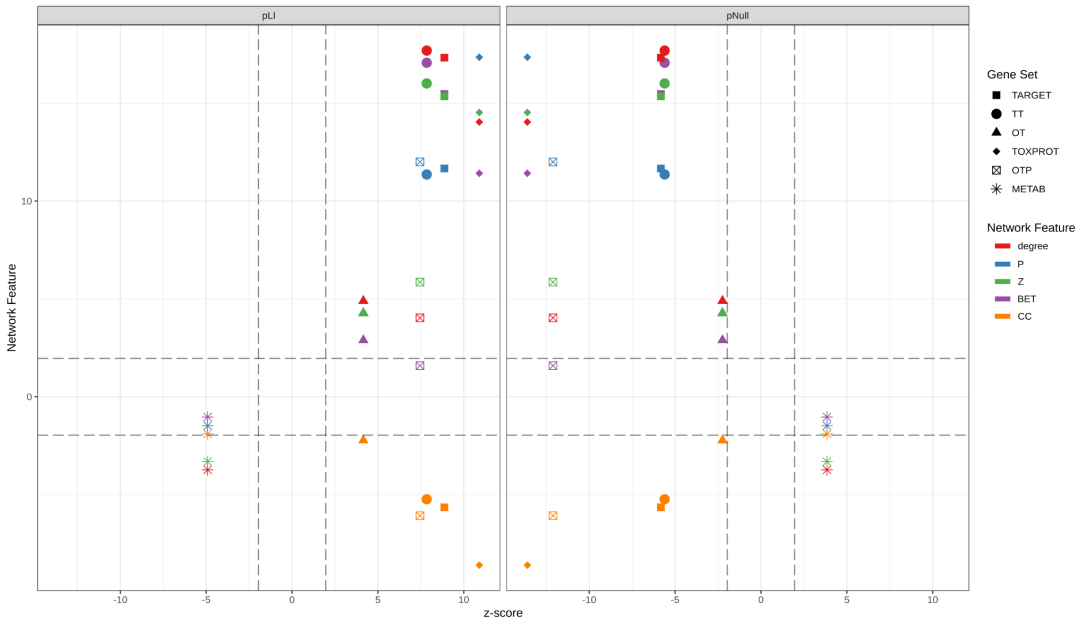


FIGURE 7 | Relationship between the network features (degree, betweenness, clustering coefficient, participation coefficient, and within-module degree) and the tolerance to LoF variation (pLI and pNull) of the gene sets in the INBIOMAP global interactome. We plot the *z*-scores of the genomic and the network features resulting from 10,000 randomizations. TARGET: drug targets, METAB: proteins that are involved in the drug metabolism, absorption, distribution, metabolism, and excretion. TOXPROT: proteins associated to side effects. TT: genes in common between drug targets and toxicity genes. OT: only TARGET proteins. OTP: only TOXPROT proteins.

and non-target proteins. Our results also support those of Kotlyar et al. (2012) who showed that drug targets and drug-regulated genes have higher degree and betweenness, and lower clustering coefficients.

Drug targets and drug targets that cause side effects are significantly LoF intolerant, while METAB proteins are relatively tolerant to LoF variants and to homozygous LoF variants (**Figure 7**). These results agree with those of a recent study (Wright et al., 2018) that found high-confidence LoF variants in more than half of the pharmacogenes under analysis. The relatively high tolerance of METAB proteins may be at least partially explained by their degree of paralogy (Pan et al., 2016), and overlapping substrate specificity across these enzymes (Zhou, 2008). In detail, there are 32 cytochromes in the METAB dataset, and their drug specificity ranges from 1 to over 600 drugs. They are all characterized by their relatively high tolerance to LoF mutations (e.g., low pLI values and high pNull values). An example of redundancy in these enzymes is CYP3A5 (pLI = 5.2 e-11). It has been reported that CYP3A5 deficiency occurs in approximately 75% of white persons and 50% of African descent populations because of a single nucleotide polymorphism (CYP3A5*3, 6986A > G) within intron 3 that introduces a premature stop codon and truncation of the protein (Kuehl et al., 2001). Because many drugs metabolized by CYP3A5 are also substrates of CYP3A4, truncating mutations in either of these proteins might produce no visible phenotype.

The intolerance to LoF variation observed in the TARGET set is mainly driven by TT genes, as shown by the smaller z-scores of pLI, and pNull of OT genes (**Figure 7**). GPCRs behave differently than the other TARGET classes. They possess lower pLI and higher pNull values than the rest of the genes. A closer look at this set of proteins shows that GPCRs that do not directly mediate side effects (OT set, 89 genes with ExAC data) are responsible for this trend, since GPCRs in the TT set (123 genes with ExAC data) have no significantly different pLI values than the rest of the genes. GPCRs do not seem to play central roles in the network at the global and meso-scale level (although they display low clustering coefficient, see **Supplementary Figure S6**), which suggests that they are not under strong negative selection and therefore would be more tolerant to functional variants. A recent study of the pharmacogenomics of 108 GPCRs targeted by FDA approved drugs (Hauser et al., 2018) showed that GPCRs have, on average, LoF mutations in 9 different positions per receptor, and at least 1 LoF variant has been observed in each of the GPCRs under study. The mechanisms that might explain the compatibility of these drastic genomic alterations with normal phenotypes could be heterozygosity, epistasis, and allele-specific expression. Nevertheless, it is also possible that some of the receptors with low pLI have functional redundancy.

The fact that drug targets that mediate side effects tend to be more intolerant to LoF variation is in line with the finding that the inter individual genomic variability of drug targets is a strong predictor of the withdrawal of drugs (Lee et al., 2016). This study, using several metrics to estimate the deleteriousness of variants in 2,504 publicly available genomes from the 1000 Genomes Project, found a high person-to-person variability of deleterious variants among drug-related genes. They also designed a genomic deleteriousness score that they found to be significantly lower for

withdrawn drugs, and US FDA pharmacogenomic biomarkers than for other drug-related proteins.

Finally, we have characterized the expression of drug related genes across healthy human tissue, showing differences in the pattern of expression among the different gene sets. To the best of our knowledge, this is the first study that performs such analysis.

Our results show that there is a relationship between the role in the cellular network of genes involved in different drug effects and their tolerance to LoF variation. We have uncovered a scenario in which proteins that mediate side effects are more central, tend to be more intolerant to LoF mutations, and are highly expressed in most of the human tissues. The subset of drug targets that mediate drug adverse reactions occupy more central positions in the network –not only because they have a high degree, but because they connect different network modules–, and they also exhibit higher sensitivity to LoF variants. In contrast, drug targets that do not mediate side effects do not exhibit any significant pattern of network centrality, and appear to be under weaker negative selection. The case of ADME proteins is particular, because they are less central, tolerate LoF mutations, and show a very specific tissue expression pattern. The integrated analysis of different omics data reveals distinct features of proteins associated to drug response, which is relevant in the context of drug development and pharmacogenomics.

AUTHOR CONTRIBUTIONS

JP and LF conceived and designed the experiments. JP performed the experiments. JP, AG-P, EG, and LF analyzed the data. JP, AG-P, EG, LF, JA-P, FS, and BO reviewed and discussed the results. JP, LF, and AG-P wrote the paper. All the authors reviewed the final version of the manuscript.

FUNDING

We received support from ISCIII-FEDER (PI13/00082, CP10/00524, and CPII16/00026), IMI-JU under grant agreements no. 116030 (TransQST) and no. 777365 (eTRANSAFE) resources of which are composed of financial contribution from the EU-FP7 (FP7/2007-2013) and EFPIA companies in kind contribution, and the EU H2020 Program 2014–2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and was supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. The DCEXS is a "Unidad de Excelencia María de Maeztu", funded by the MINECO (ref: MDM-2014-0370). AG-P was supported by a Ramón y Cajal contract (RYC-2013-14554). BO and JA-P were supported by MINECO grant BIO2014-57518-R.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00412/full#supplementary-material

REFERENCES

Ahmed, S., Zhou, Z., Zhou, J., and Chen, S.-Q. (2016). Pharmacogenomics of drug metabolizing enzymes and transporters: relevance to precision medicine. *Genomics Proteomics Bioinformatics* 14, 298–313. doi: 10.1016/j.gpb.2016.03.008

Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res.* 45, D408–D414. doi: 10.1093/nar/gkw985

Bauer-Mehren, A., van Mullingen, E. M., Avillach, P., Carrascosa, M. D. C., Garcia-Serna, R., Piñero, J., et al. (2012). Automatic filtering and substantiation of drug safety signals. *PLoS Comput. Biol.* 8:e1002457. doi: 10.1371/journal.pcbi.1002457

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., et al. (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090. doi: 10.1093/nar/gkt1031

Berenstein, A. J., Piñero, J., Furlong, L. I., and Chernomoretz, A. (2015). Mining the modular structure of protein interaction networks. *PLoS One* 10:e0122477. doi: 10.1371/journal.pone.0122477

Berger, S. I., Ma'ayan, A., and Iyengar, R. (2010). Systems pharmacology of arrhythmias. *Sci. Signal.* 3:ra30. doi: 10.1126/scisignal.2000723

Berger, S. S., and Iyengar, R. (2011). Role of systems pharmacology in understanding drug adverse events. *Rev. Syst. Biol. Med.* 3, 129–135. doi: 10.1002/wsbm.114.Role

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, 267D–270D. doi: 10.1093/nar/gkh061

Brouwers, L., Iskar, M., Zeller, G., van Noort, V., and Bork, P. (2011). Network neighbors of drug targets contribute to drug side-effect similarity. *PLoS One* 6:e22187. doi: 10.1371/journal.pone.0022187

Brown, A. S., and Patel, C. J. (2017). A standard database for drug repositioning. *Sci. Data* 4:170029. doi: 10.1038/sdata.2017.29

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

Cotto, K. C., Wagner, A. H., Feng, Y.-Y., Kiwala, S., Coffman, A. C., Spies, G., et al. (2017). DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Res.* 46, D1068–D1073. doi: 10.1093/nar/gkx1143

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Interj. Complex Syst.* 1695, 1–9.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., et al. (2017). The comparative Toxicogenomics database: update 2017. *Nucleic Acids Res.* 45, D972–D978. doi: 10.1093/nar/gkw838

Gashaw, I., Ellinghaus, P., Sommer, A., and Asadullah, K. (2011). What makes a good drug target? *Drug Discov. Today* 16, 1037–1043. doi: 10.1016/J.DRUDIS.2011.09.007

Guimerà, R., and Amaral, L. A. N. (2005). Cartography of complex networks: modules and universal roles. *J. Stat. Mech.* 2005:niha35573. doi: 10.1088/1742-5468/2005/02/P02001

Guney, E. (2017). *Investigating Side Effect Modules in the Interactome and Their Use in Drug Adverse Effect Discovery*. Cham: Springer, 239–250. doi: 10.1007/978-3-319-54241-6_21

Guney, E., Menche, J., Vidal, M., and Barábasi, A.-L. (2016). Network-based in silico drug efficacy screening. *Nat. Commun.* 7:10331. doi: 10.1038/ncomms10331

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402, C47–C52. doi: 10.1038/35011540

Hauser, A. S., Chavali, S., Masuho, I., Jahn, L. J., Martemyanov, K. A., Gloriam, D. E., et al. (2018). Pharmacogenomics of GPCR drug targets. *Cell* 172, 41.e19–54.e19. doi: 10.1016/j.cell.2017.11.033

Kotlyar, M., Fortney, K., and Jurisica, I. (2012). Network-based characterization of drug-regulated genes, drug targets, and toxicity. *Methods* 57, 499–507. doi: 10.1016/j.ymeth.2012.06.003

Kozyra, M., Ingelman-Sundberg, M., and Lauschke, V. M. (2017). Rare genetic variants in cellular transporters, metabolic enzymes and nuclear receptors can be important determinants of interindividual differences in drug response. *Genet. Med.* 19, 20–29. doi: 10.1038/gim.2016.33

Kuehl, P., Zhang, J., Lin, Y., Lamba, J., Assem, M., Schuetz, J., et al. (2001). Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat. Genet.* 27, 383–391. doi: 10.1038/86882

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Res.* 44, D1075–D1079. doi: 10.1093/nar/gkv1075

Lauschke, V. M., Milani, L., and Ingelman-Sundberg, M. (2018). Pharmacogenomic biomarkers for improved drug therapy—recent progress and future developments. *AAPS J.* 20:4. doi: 10.1208/s12248-017-0161-x

Lee, K. H., Baik, S. Y., Lee, S. Y., Park, C. H., Park, P. J., and Kim, J. H. (2016). Genome sequence variability predicts drug precautions and withdrawals from the market. *PLoS One* 11:e0162135. doi: 10.1371/journal.pone.0162135

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi: 10.1038/nature19057

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., et al. (2017). A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* 14, 61–64. doi: 10.1038/nmeth.4083

Lin, Y., Mehta, S., Küçük-McGinty, H., Turner, J. P., Vidovic, D., Forlin, M., et al. (2017). Drug target ontology to classify and integrate drug discovery data. *J. Biomed. Semantics* 8:50. doi: 10.1186/s13326-017-0161-x

Madian, A. G., Wheeler, H. E., Jones, R. B., and Dolan, M. E. (2012). Relating human genetic variation to variation in drug responses. *Trends Genet.* 28, 487–495. doi: 10.1016/j.tig.2012.06.008

Mannil, D., Vogt, I., Prinz, J., and Campillos, M. (2015). Organ system heterogeneity DB: a database for the visualization of phenotypes at the organ system level. *Nucleic Acids Res.* 43, D900–D906. doi: 10.1093/nar/gku948

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science* 348, 660–665. doi: 10.1126/science.aaa0355

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., et al. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* 45, D183–D189. doi: 10.1093/nar/gkw1138

Nguyen, D.-T., Mathias, S., Bologa, C., Brunak, S., Fernandez, N., Gaulton, A., et al. (2017). Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* 45, D995–D1002. doi: 10.1093/nar/gkw1072

Pan, S.-T., Xue, D., Li, Z.-L., Zhou, Z.-W., He, Z.-X., Yang, Y., et al. (2016). Computational identification of the paralogs and orthologs of human cytochrome P450 superfamily and the implication in drug discovery. *Int. J. Mol. Sci.* 17:1020. doi: 10.3390/ijms17071020

Perez-Lopez, Á. R., Szalay, K. Z., Türei, D., Módos, D., Lenti, K., Korcsmáros, T., et al. (2015). Targets of drugs are generally and targets of drugs having side effects are specifically good spreaders of human interactome perturbations. *Sci. Rep.* 5:10182. doi: 10.1038/srep10182

Piñero, J., Berenstein, A., Gonzalez-Perez, A., Chernomoretz, A., and Furlong, L. I. (2016a). Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. *Sci. Rep.* 6:24570. doi: 10.1038/srep24570

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2016b). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45:gkw943. doi: 10.1093/nar/gkw943

Pinto, N., and Dolan, M. E. (2011). Clinically relevant genetic variations in drug metabolizing enzymes. *Curr. Drug Metab.* 12, 487–497.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Available at: https://www.r-project.org/

Roden, D. M., Wilke, R. A., Kroemer, H. K., and Stein, C. M. (2011). Pharmacogenomics: the genetics of variable drug responses. *Circulation* 123, 1661–1670. doi: 10.1161/CIRCULATIONAHA.109.914820

Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1118–1123. doi: 10.1073/pnas.0706851105

Schärfe, C. P. I., Tremmel, R., Schwab, M., Kohlbacher, O., and Marks, D. S. (2017). Genetic variation in human drug-related genes. *Genome Med.* 9:117. doi: 10.1186/s13073-017-0502-5

Shah, N. H. (2016). Data descriptor: a curated and standardized adverse drug event resource to accelerate drug safety research. *Sci. Data* 3:160026. doi: 10.1038/sdata.2016.26

Shenfield, G. M. (2004). Genetic polymorphisms, drug metabolism and drug concentrations. *Clin. Biochem. Rev.* 25, 203–206.

Tatonetti, N. P., Ye, P. P., Daneshjou, R., and Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* 14:125ra31. doi: 10.1126/scitranslmed.3003377

Ursu, O., Holmes, J., Knockel, J., Bologa, C. G., Yang, J. J., Mathias, S. L., et al. (2017). DrugCentral: online drug compendium. *Nucleic Acids Res.* 45, D932–D939. doi: 10.1093/nar/gkw993

van der Wouden, C., Cambon-Thomsen, A., Cecchin, E., Cheung, K., Dávila-Fajardo, C., Deneer, V., et al. (2017). Implementing pharmacogenomics in Europe: design and implementation strategy of the ubiquitous pharmacogenomics consortium. *Clin. Pharmacol. Ther.* 101, 341–358. doi: 10.1002/cpt.602

Weinshilboum, R. M., and Wang, L. (2017). Pharmacogenomics: precision medicine and drug response. *Mayo Clin. Proc.* 92, 1711–1722. doi: 10.1016/j.mayocp.2017.09.001

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Wright, G. E. B., Carleton, B., Hayden, M. R., and Ross, C. J. D. (2018). The global spectrum of protein-coding pharmacogenomic diversity. *Pharmacogenomics J.* 18, 187–195. doi: 10.1038/tpj.2016.77

Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabasi, A.-L., and Vidal, M. (2007). Drug—target network. *Nat. Biotechnol.* 25, 1119–1126. doi: 10.1038/nbt1338

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhou, S.-F. (2008). Drugs behave as substrates, inhibitors and inducers of human cytochrome P450 3A4. *Curr. Drug Metab.* 9, 310–322. doi: 10.2174/138920008784220664

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 October 2018
doi: 10.3389/fgene.2018.00429

Investigation of Nrf2, AhR and ATF4 Activation in Toxicogenomic Databases

Elias Zgheib[1], Alice Limonciel[2], Xiaoqi Jiang[3], Anja Wilmes[2], Steven Wink[4], Bob van de Water[4], Annette Kopp-Schneider[3], Frederic Y. Bois[5]* and Paul Jennings[2]*

[1] Laboratoire de Biomécanique et Bio-ingénierie, Sorbonne Universités – Université de Technologie de Compiègne, Compiègne, France, [2] Division of Molecular and Computational Toxicology, Amsterdam Institute for Molecules, Medicines and Systems, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [3] Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany, [4] Division of Drug Discovery and Safety, Leiden Cell Observatory High Content Imaging Screening Facility, Leiden Academic Center for Drug Research, Leiden University, Leiden, Netherlands, [5] Models for Ecotoxicology and Toxicology Unit (DRC/VIVA/METO), Institut National de l'Environnement Industriel et des Risques, Verneuil-en-Halatte, France

OPEN ACCESS

Edited by:
Hideko Sone, National Institute for Environmental Studies, Japan

Reviewed by:
Rodrigo Juliani Siqueira Dalmolin, Federal University of Rio Grande do Norte, Brazil
Immacolata Porreca, Wellcome Trust Sanger Institute (WT), United Kingdom

***Correspondence:**
Frederic Y. Bois
frederic.bois@certara.com
Elias Zgheib
elias.zgheib.pro@gmail.com

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Genetics

Received: *30 May 2018*
Accepted: *11 September 2018*
Published: *02 October 2018*

Citation:
Zgheib E, Limonciel A, Jiang X, Wilmes A, Wink S, van de Water B, Kopp-Schneider A, Bois FY and Jennings P (2018) Investigation of Nrf2, AhR and ATF4 Activation in Toxicogenomic Databases. Front. Genet. 9:429. doi: 10.3389/fgene.2018.00429

Toxicological responses to chemical insult are largely regulated by transcriptionally activated pathways that may be independent, correlated and partially or fully overlapping. Investigating the dynamics of the interactions between stress responsive transcription factors from toxicogenomic data and defining the signature of each of them is an additional step toward a system level understanding of perturbation driven mechanisms. To this end, we investigated the segregation of the genes belonging to the three following transcriptionally regulated pathways: the AhR pathway, the Nrf2 pathway and the ATF4 pathway. Toxicogenomic datasets from three projects (carcinoGENOMICS, Predict-IV and TG-GATEs) obtained in various experimental conditions (in human and rat *in vitro* liver and kidney models and rat *in vivo*, with bolus administration and with repeated doses) were combined and consolidated where overlaps between datasets existed. A bioinformatic analysis was performed to refine pathways' signatures and to create chemical activation capacity scores to classify chemicals by their potency and selectivity of activation of each pathway. With some refinement such an approach may improve chemical safety classification and allow biological read across on a pathway level.

Keywords: transcriptomics, Nrf2, AhR, ATF4, toxicity pathways, toxicogenomic, oxidative stress

INTRODUCTION

Many transcriptionally activated pathways are intimately involved in responses to chemical induced perturbations and toxicological outcomes (Jennings et al., 2013). Here we focus on three such pathways. (1) The Nrf2 pathway (Nuclear Factor (Erythroid-derived 2)-Like 2 NFE2L2) which regulates the response to oxidative stress, (2) the ATF4 (Activating Transcription Factor 4) branch of the unfolded protein response and (3) the dioxin response or AhR pathway (Aryl Hydrocarbon Receptor). While these pathways have specific non-overlapping activation mechanisms and specific non-overlapping DNA binding elements reviewed in (Jennings et al., 2013), they also have overlapping downstream target genes. Adding to this complexity, converging toxicological mechanisms may lead to co-activation.

Oxidative stress is a major cause of chemical-induced injury and associated chronic diseases (e.g., cancer or Parkinson's disease) (Taguchi et al., 2011; Kong et al., 2014). The Nrf2 pathway the

main adaptive response to oxidative stress. The Nrf2 protein exists in an inactive, cytoplasm-localized state, that is bound to the cytoskeleton-associated KEAP1 which facilitates Nrf2 ubiquitination and degradation. Upon oxidative stress, a conformational change in KEAP1 makes its binding to Nrf2 less favorable. Nrf2 stabilizes and translocates to the nucleus where it binds the antioxidant response element and drives the transcription of a genes involved in glutathione synthesis and recycling, xenobiotic metabolism and transport, and antioxidant genes (Jennings et al., 2013). ATF4 is a major branch of the unfolded protein response and is activated in response to endoplasmic reticulum (ER) disturbances or proteotoxicity where unfolded proteins accumulate in the ER and compete with PERK for the inhibitory protein BiP (Leonard et al., 2014). Activated PERK phosphorylates eIF2a which inhibits general protein translation while inducing AT4 translation. ATF4 in turn binds to the CARE consensus sequence and drives transcription of genes involved in amino acid synthesis, amino acid transport and aminoacyl-tRNA synthesis (Leonard et al., 2014). Xenobiotics can also activate specific genes through the AhR pathway. Upon ligand (xenobiotic) binding, the AhR transcription factor (TF) shuttles into the nucleus where it dimerizes with the "AhR nuclear translocator" and binds to so-called xenobiotic-responsive elements (XRE), aka dioxin response element (DRE), in the promoter region down stream targets including cytochrome P1-450 A1 (*CYP1A1*) (Haarmann-Stemmann et al., 2012).

Measuring the activation of transcriptionally regulated pathways such Nrf2, AhR, and ATF4 using transcriptomic approaches has great potential in increasing mechanistic understanding of chemical perturbations and to develop better prediction tools (Aschauer et al., 2015; Limonciel et al., 2015). Also, such an approach could be used for biological read across. However, there is still a knowledge gap pertaining to the interplay between the Nrf2, AhR, and ATF4 pathways. It is known that several of their downstream targets have promotor sequences for more than one of these TFs. For example, NQO1 is driven by both AhR and Nrf2. Also, it is likely that the pathways may cooperate in redressing certain hoemeostatic perturbations. For example, we have shown that Nrf2 and ATF4 cooperate on the level of glutathione, where ATF4 promotes the uptake of glutathione amino acid building blocks including glutamine and cysteine and promotes glutamate production via induction of asparagine synthetase. Nrf2 in turn through induction of glutamate cysteine ligase and glutathione synthase produce new glutathione (Wilmes et al., 2013). Very little is known about species differences, tissue specificity, chemical specificity, or other subtleties in the activation of these pathways.

To investigate this further, we performed a transcriptomic analysis of large and medium size toxicogenomic datasets from the EU 6th and 7th framework projects carcinoGENOMICS and Predict-IV, as well as from TG-GATEs. Within these studies we also identified some potentially useful specific activators of the pathways investigated. Potassium bromate and phorone have been used to experimentally activate Nrf2. Potassium bromate is an oxidizing agent causing ROS injury and oxidative stress-induced DNA damage (Ballmaier and Epe, 1995; Limonciel et al., 2012). In a recent study we showed that potassium bromate activated the Nrf2 and p53 response without activation of the ATF4 response (Limonciel et al., 2018). Phorone can similarly activate Nrf2 due to glutathione depletion (Younes et al., 1986; Iannone et al., 1990; Oguro et al., 1996). Tunicamycin is a prototypical activator of the unfolded protein response (including the ATF4 branch) by causing an accumulation of misfolded glycoproteins in the ER (Oslowski and Urano, 2011). More specifically, tunicamycin inhibits the N-glycosylation of newly formed proteins by DPAGT1, leading to an interruption in glycoprotein production (Bassik and Kampmann, 2011). Benzo(a)pyrene and omeprazole have been used to activate AhR. Benzo(a)pyrene is a polycyclic aromatic hydrocarbon and a prototypical AhR agonist (Nebert et al., 2004). Omeprazole, a proton pump inhibitor (Howden, 1991, 199) is also an AhR activator (Jin et al., 2012, 2014).

The aim of the investigation was to investigate potential co-dependences of ATF4, Nrf2 and/or AhR, to develop a signature panel for each pathway and to develop a chemical activity scoring system, for chemical grouping.

Abbreviations: AhR, Aryl Hydrocarbon Receptor; ATF4, Activating Transcription Factor 4; CAC, Chemical Activation Capacity; ChIP-seq, Chromatin Immunoprecipitation followed by DNA sequencing; CYP1A1, Cytochrome P1-450 A1; FC, Fold Change; Nrf2, Nuclear Factor (Erythroid-derived 2)-Like 2 (NFE2L2); TF, Transcription Factors.

MATERIALS AND METHODS

Generation of Target Gene Lists

For each of the three TF of interest (AhR, Nrf2, and ATF4), the following three search strategies, from the works of (Limonciel et al., 2015), were applied in PubMed to retrieve TF target genes: (i) search for TF name and ChIP-sequencing or ChIP-microarray studies, (ii) search for TF name and TF-specific response element and "Electrophilic Mobility Shift Assay" or ChIP studies, and (iii) search for TF name and TF-specific DNA response element and name of a target gene known. In the first tier of this strategy, high-throughput sequencing datasets were retrieved, which provided extensive lists of genes shown to have the TF bind in their promoter region. In the second tier, lower throughput investigations were included, providing target genes that were more deeply investigated in the article with proven TF binding of the promoter region. These first two tiers provided an unbiased source of target genes that was completed in the third tier with manually added target genes for which at least one study showed binding of the TF in their promoter region.

PubMed searches were performed on 24.11.2014 for Nrf2 and 17.12.2014 for ATF4 and AhR. Gene lists are reported in **Supplementary Table 1** and illustrated on **Supplementary Figure 1**.

Construction of a Chemical-Effects Transcriptomics Database

The database of chemical-induced transcriptomic changes comes from three projects: carcinoGENOMICS (Vinken et al., 2008), Predict-IV (Mueller et al., 2015) and TG-GATEs (Igarashi et al.,

TABLE 1 | Number of chemicals used in each experimental category.

Project	Species	Tissue	Setting	Mode	Time-points	Number of chemicals	Notes
All dataset [211]*							(1–2)
Carcino-GENOMICS [31]	Human	Kidney	*in vitro*	Bolus	6h, 24h, 72h	30	(3–4)
	Rat	Kidney	*in vitro*	Bolus	6h, 24h, 72h	15	
PREDICT-IV [22]	Human	Kidney	*in vitro*	Repeated doses	1d, 3d, 14d	12	(5–6)
	Human and Rat	Liver	*in vitro*	Repeated doses	1d, 3d, 14d	11	(7)
TG-GATEs [171]	Human	Liver	*in vitro*	Bolus	2h, 8h, 24h	160	(8)
	Rat	Liver	*in vitro*	Bolus	2h, 8h, 24h	145	(9)
		Liver	*in vivo*	Bolus	3h, 6h, 9h, 24h	158	(10–11)
		Liver	*in vivo*	Repeated doses	4d, 8d, 15d, 29d	143	–
		Kidney	*in vivo*	Bolus	3h, 6h, 9h, 24h	41	(12)
		Kidney	*in vivo*	Repeated doses	4d, 8d, 15d, 29d	41	

(1) Number of chemicals assayed in at least one of the three source projects.
(2) ***Cyclosporine A*** *is the only chemical that was used in the three projects.* ***Cyclosporine A*** *appears in every single experimental category and sub-category (except carcinoGENOMICS's Rat tests).*
*(3) In carcinoGENOMICS, all 15 chemicals tested on rat cells, except one (****Dimethylnitrosamine****), were also tested on human cells.*
(4) Beside ***Cyclosporine A****, and five of the chemicals that appear in TG-GATEs as well, all chemicals are specific to carcinoGENOMICS [****2-Nitrofluorene*** *and* ***N-nitrosomorpholine*** *(TG-GATEs "Human liver in vitro bolus" and "Rat liver in vivo bolus"); and* ***Diclofenac****,* ***Nifedipine*** *and* ***Tolbutamide*** *(all liver categories of TG-GATEs)].*
(5) The 12 chemicals tested on kidney cells and the 11 tested on liver cells in PREDICT-IV are distinct; Only ***Cyclosporine A*** *is presented in these two categories.*
(6) Among the chemicals tested on kidney cells in PREDICT-IV, only ***Cisplatin*** *appears elsewhere (in TG-GATEs rat tests).*
(7) Among the chemicals tested on liver cells in PREDICT-IV, only ***Acetaminophen*** *and* ***Valproic acid*** *appear in all TG-GATEs categories;* ***Amiodarone****,* ***Chlorpromazine****,* ***Fenofibrate****,* ***Ibuprofen*** *and* ***Metformin*** *were tested on liver cells of TG-GATEs, and* ***Rosiglitazone*** *as well (except in "Rat liver in vitro bolus").*
*(8) In TG-GATEs, five chemicals were tested on human cells only (****HGF****,* ***IL1beta****,* ***IL6****,* ***INFalpha****,* ***Nefazodone,*** *and* ***TGFbeta1****) and six others on animal categories only (****Carboplatin****,* ***Cephalotin****,* ***Cisplatin****,* ***Gentamicin****,* ***TNFalpha,*** *and* ***Trimethadione****).*
(9) Five chemicals appear in liver in vitro bolus categories only (human and rat): ***Alpidem****,* ***Buspirone****,* ***Clozapine****,* ***Nefazodone*** *and* ***Venlafaxine****.*
(10) ***3-Methylcholantrene****,* ***Bortezomib****,* ***Gefitinib****,* ***Imatinib,*** *and* ***Puromycin*** *appear in the "Rat liver in vivo bolus" category exclusively.*
(11) ***2-Nitrofluorene****,* ***Aflatoxin B1****,* ***Dexamethasone****,* ***N-methyl-N-nitrosourea*** *and* ***TNF*** *are common to TG-GATEs' "Human" and "Rat liver in vivo bolus" categories and were not tested in other conditions.*
(12) The 41 chemicals that are used for TG-GATEs kidney in vivo testing are the same for both modes (bolus and repeated doses) and are common for all other categories (exceptions: ***Gentamicin****,* ***Carboplatin****,* ***Cephalotin****,* ***Cisplatin****,* ***Desmopressin acetate****,* ***Amphotricine B,*** *and* ***Acetamide)****.*
**The number between brackets refers to the number of chemicals per project.*

TABLE 2 | Chosen pathway specific chemical through the dataset.

Pathway	Species	Kidney		Liver	
		in vitro	*in vivo*	*in vitro*	*in vivo*
AhR	Human	Benzo(a)pyrene		Omeprazole	
	Rat				
Nrf2	Human	Potassium Bromate		Phorone	
	Rat				
ATF4	Human			Tunicamycin	
	Rat				

TABLE 3 | Number of conditions (chemicals, concentrations, time-points) tested per category.

Pathway	Species	Kidney		Liver		TOTAL
		in vitro	*in vivo*	*in vitro*	*in vivo*	
	Human	85	0	963	0	1048
	Rat	30	487	1282	1838	3637
Total		602		4083		4685

2015). In carcinoGENOMICS, human and rat kidney cells were exposed to bolus concentrations of up to 31 chemicals in *in vitro* settings for up to 72 h. In Predict-IV, human kidney cells and liver cells from human and rat were exposed daily *in vitro* for up to 14 days to up to 22 chemicals. Up to 171 chemicals from TG-GATEs were tested in various rat *in vivo* and *in vitro systems*, with various treating regimes. **Table 1** summarizes this and shows the 211 chemicals tested and dispatched in different categories of one or more of the three projects. **Supplementary Table 2** presents the exhaustive lists of chemicals by category.

Data Sources

The carcinoGENOMICs and Predict-IV data are publicly accessible on the diXa database hosted by The European Bioinformatics Institute[1]. In carcinoGENOMICS, *in vitro* renal cell experiments were performed using the human cell lines RPTEC/TERT1 (human, telomerase transfected) and NRK-52E (rat). The study no. is DIXA-003. Differentiated cell cultures were exposed to a single bolus of non or low cytotoxic (<IC10) concentration of chemical for 6, 24, or 72 h before lysis in TRIZOL, RNA purification and transcriptomic analysis

[1]http://wwwdev.ebi.ac.uk/fg/dixa/

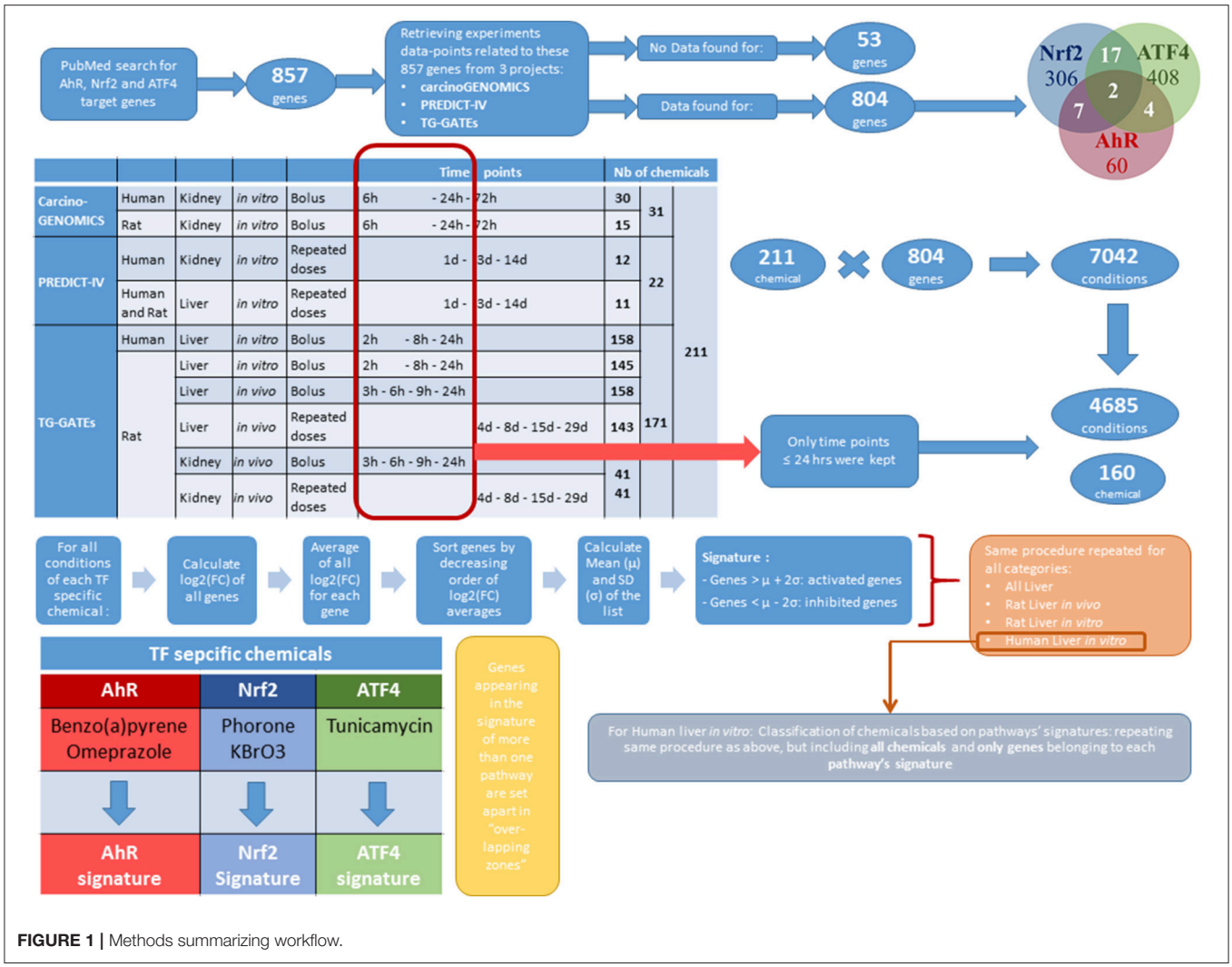


FIGURE 1 | Methods summarizing workflow.

on Affymetrix microarrays as described (Limonciel et al., 2012). Affymetrix Human Genome U133 Plus 2.0 GeneChIP arrays were used for human samples and Rat Genome 230 2.0 GeneChIP for rat samples. Normalization quality controls, including scaling factors, average intensities, present calls, background intensities, noise and raw Q-values were within acceptable limits for all chips. Hybridization controls were identified on all chips and yielded the expected increases in intensities. All subsequent analyses were based on normalized expression values generated using the MAS5 normalization algorithm. It is noted that RMA or GCRMA normalization would have been preferred. Normalized data was imported into GeneSpring (Agilent) to identify log2 fold change values for selected genes.

Within PREDICT-IV, *in vitro* testing of nephrotoxic and hepatotoxic compounds were performed on RPTEC/TERT1 cells (renal model), primary human hepatocytes, and rat hepatocytes (PHH and PRH, respectively). The study no. on the diXa database is DIXA-095. Differentiated cell cultures were exposed daily to a high (≤10% cell death) or low concentration of chemical for 1, 3 or 14 days, as described (Wilmes et al., 2013, 2014; Aschauer et al., 2015; Crean et al., 2015; Limonciel et al., 2015). Transcriptomic analysis was carried out on Illumina® HT 12 v4 BeadChip arrays for kidney and PHH human samples, except RPTEC/TERT1 exposed to CsA (HT 12 v3 chips). PRH samples were analyzed with Illumina® RatRef-12 v1 BeadChIP arrays. Results were normalized by quantile normalization and expressed as log2 fold over time-matched control. Where several probes existed for a given gene, the probe with the highest variation across the dataset was selected.

The TG-GATEs datasets comprised *in vivo* rat data from liver and kidney tissue, as well as data from *in vitro* primary rat and human hepatocyte cultures, after a single administration of chemical and repeat dosing (see **Table 1**)[2]. CEL files were downloaded from the Open TG-GATEs

[2]https://dbarchive.biosciencedbc.jp/en/open-tggates/desc.html

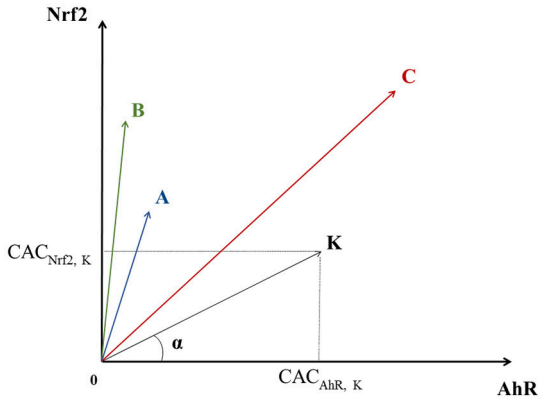


FIGURE 2 | Geometric representation of chemical specificity and potency for the Nrf2 and AhR pathways. *K* represents a chemical and its coordinates are ($CAC_{AhR,\ K}$, $CAC_{Nrf2,\ K}$). *K* also defines the vector $\overrightarrow{OK}$ linking the origin *O* (0, 0) to point *K*. The absolute value of the cosine of the angle α between $\overrightarrow{OK}$ and a pathway's axis can be used to measure the specificity of a chemical for the given pathway (the smaller α, the more specific the chemical). On the other hand, the overall activation potency of a chemical increases proportionally with the length of $\overrightarrow{OK}$. Points *A*, *B,* and *C* represent three other chemicals with different specificities and potencies for pathways' activation (see text).

database of the Toxicogenomics Project and Toxicogenomics Informatics Project under CC Attribution-Share Alike 2.1 Japan. Probe annotation for the primary human hepatocyte data was performed using the hthgu133pluspmhsentrezg.db package version 17.1.0 and probe mapping was performed with hthgu133pluspmhsentrezgcdf downloaded from NuGO[3]. Probe annotation for the rat data was performed using the rat2302rnentrezg.db package version 19.0.0 and probe mapping was performed with the rat2302rnentrezgcdf package version 19.0.0 downloaded from NuGO. These mappings summarize the corresponding probes to a single probe set per gene. Probe-wise background correction (Robust Multi-Array Average expression measure), between-array normalization within each treatment group (quantile normalization) and probe set summaries (median polish algorithm) were calculated with the RMA function of the Affy package (Affy package, version 1.38.1) (Irizarry et al., 2003). The normalized data were statistically analyzed for differential gene expression using a linear model with coefficients for each experimental group within a treatment group (Wolfinger et al., 2001). A contrast analysis was applied to compare each exposure with the corresponding vehicle control. For hypothesis testing the moderated t-statistics by empirical Bayes moderation was used followed by an implementation of the multiple testing correction of Benjamini and Hochberg (Hochberg and Benjamini, 1990) using the LIMMA package (Smyth et al., 2005).

All interspecies gene conversions where done using the provided human gene symbols which were converted to human or rat gene identifiers using the online conversion tool of bioDBnet[4].

[3]http://nmg-r.bioinformatics.nl/NuGO_R.html
[4]https://biodbnet-abcc.ncifcrf.gov/

Altogether, the collected data concern 804 genes from the 857 genes identified in PubMed as targets of AhR, Nrf2 and ATF4. The 53 target genes that are not covered with data from any of the three projects were excluded from this study. These genes are listed in the last row of **Supplementary Table 1**.

Bioinformatics Methods

Data Selection

The heterogeneity of the sources of information of our database widens its coverage and strengthens its capacity to represent multiple conditions. However, this richness makes the database's structure complex. To simplify the analysis without losing potentially important information, we focused on conditions providing the best background to study the three pathways individually. The effects observed following exposure to a chemical could vary greatly depending on exposure duration. Exposures lasting more than 24 h tend to cause mixed stress responses that make it difficult to delineate the activation of specific molecular pathways and the initial mechanisms of toxicity of chemicals. These conditions could be a potential source of noise for the analysis and were thus excluded. Excluding all data obtained after 24 h reduced the dataset from 7,042 to 4,685 testing conditions. We chose not to eliminate the early kidney *in vivo* time points (at 3 and 6 h), even though they may be more reflective of background levels in case of slow absorption of the chemical administered.

Pathway Specific Chemicals

In order to distribute the genes to pathways and pathway overlapping zones, log2 genes fold changes (FC) were ranked in decreasing order and examined on reduced datasets containing conditions relative to pathway specific activators. We define a pathway specific activator as a chemical where the mode of action is known, that the mode of action activates the specific pathways and that this mode of action is not expected to activate the other pathways under investigation. Thus, at relatively short exposures, to relatively low concentrations these chemicals will only act on their specific target. It is however possible at higher concentrations or longer time exposure, other targets will be affected due to increasing toxicity. As shown in **Table 1**, some chemicals were not tested in all categories and tissue types. Thus, it was not possible to find pathway specific activators able to cover the entire database. **Table 2** shows the coverage of the datasets by the pathway specific activators selected as reference for analysis. Although none of the toxicogenomic databases analyzed here were designed to specifically address any of our three pathways of interest, most datasets included at least one chemical that could be considered as a specific pathway activator. Two specific chemicals were selected for AhR (Benzo(a)pyrene and Omeprazole) and Nrf2 (Potassium Bromate and Phorone) and one for ATF4 (Tunicamycin). However, within "Rat Kidney *in vivo*" category, no Nrf2 specific chemicals were found, and for all kidney data no ATF4 specific chemical were found either.

TABLE 4 | Pathways' global signatures for AhR, Nrf2 and ATF4 pathways and the signatures of their overlapping zones (AhR-Nrf2, Nrf2-ATF4, AhR-ATF4, and AhR-Nrf2-ATF4) for all available data.

Activated genes	AhR Signature			Nrf2 Signature			ATF4 Signature		
	Genes	log_2 (FC) averages	*A priori* pathway	Genes	log_2 (FC) averages	*A priori* pathway	Genes	log_2 (FC) averages	*A priori* pathway
	CYP1A1	4.35	AhR	*HMOX1*	1.12	Nrf2	*DDIT3*	1.59	ATF4
	DLL1	1.36	AhR	*SRXN1*	0.97	ATF4 Nrf2	*TSLP*	1.51	ATF4
	RUNX2	1.03	AhR	*MAFF*	0.78	AhR Nrf2	*AKNA*	1.30	ATF4
	SLC16A9	0.92	Nrf2	*OSGIN1*	0.67	Nrf2	*HERPUD1*	1.23	ATF4
	FAM65C	0.79	AhR	*DUSP5*	0.66	ATF4	*SLC1A4*	1.15	ATF4
	FLRT1	0.78	ATF4	*TXNRD1*	0.63	ATF4	*IL23A*	1.05	ATF4
	FIBIN	0.77	ATF4	*GCLC*	0.60	ATF4	*CHAC1*	0.99	ATF4
	TIPARP	0.73	AhR	*PPP1R15A*	0.57	ATF4	*FGF21*	0.95	ATF4
	CYP1A2	0.69	AhR	*GCLM*	0.57	Nrf2	*HSPA5*	0.94	ATF4
	ASB3	0.67	Nrf2	*HSPA1B*	0.56	Nrf2	*NUPR1*	0.94	ATF4
	PDE1A	0.66	ATF4	*FBXO30*	0.55	ATF4	*GTPBP2*	0.91	ATF4
	PBX1	0.64	Nrf2	*GSTP1*	0.53	Nrf2	*PDIA4*	0.87	Nrf2
				PHGDH	0.46	Nrf2	*FAM129A*	0.87	ATF4
				TMEFF2	0.46	ATF4	*LONP1*	0.80	ATF4
				RUNX3	0.46	Nrf2	*VNN3*	0.78	ATF4
							SESN2	0.75	ATF4
							MTHFD2	0.73	ATF4
							PYCR1	0.72	ATF4
							BACH1	0.68	Nrf2
Inhibited genes	*SLC1A7*	−1.57	ATF4	*TMEM189*	−1.48	ATF4	*COCH*	−1.25	Nrf2
	PSG5	− 1.43	AhR	*NREP*	−0.99	ATF4	*SNAI2*	− 1.20	ATF4
	PRKAR2B	−1.23	Nrf2	*KIFC1*	−0.79	ATF4	*INSIG1*	−1.02	Nrf2
	SOAT2	−0.80	ATF4	*DLX2*	− 0.78	Nrf2	*AKR1B10*	−0.96	Nrf2
	DAAM2	−0.78	Nrf2	*BMF*	−0.73	ATF4	*PMAIP1*	−0.88	Nrf2
	WDR63	− 0.70	AhR	*TGFB2*	−0.72	ATF4	*ANGPTL4*	− 0.87	ATF4
	FAM69A	−0.68	Nrf2	*DDC*	− 0.71	Nrf2	*SNRNP35*	− 0.77	ATF4
	CDH11	−0.67	Nrf2	*GLI2*	−0.71	ATF4	*SERPINE1*	−0.68	Nrf2
	LCN2	−0.66	ATF4	*AURKB*	−0.69	ATF4	*PRC1*	−0.65	Nrf2
	PLA2G4A	−0.66	Nrf2	*NEDD9*	−0.67	ATF4	*LMCD1*	−0.64	AhR
	CXCL5	−0.64	Nrf2	*TFPI*	−0.65	ATF4	*LBH*	−0.61	Nrf2
	WISP1	−0.62	ATF4	*OSMR*	− 0.59	Nrf2			

Activated or Inhibited genes	AhR-Nrf2 Overlapping signature			Nrf2-ATF4 Overlapping signature		
	Genes	AhR log_2(FC) averages	Nrf2 log_2(FC) averages	Genes	Nrf2 log_2(FC) averages	ATF4 Log2 FC average
	NQO1	0.7	0.83	*ATF3*	0.73	0.90
	DLGAP5	−0.64	−0.56	*SLC7A11*	0.70	0.69
	CFTR	−0.69	−0.73	*TRIB3*	0.70	1.02
	RAB39B	−0.92	−0.52	*CABC1*	0.56	2.90
	GSTA1	−1.43	−0.83	*GDF15*	0.48	0.80
				CCL2	−0.61	−1.28
				KCNT2	−0.9	0.76

Activated or Inhibited genes	AhR-ATF4 Overlapping signature			AhR-Nrf2-ATF4 Overlapping signature			
	Genes	AhR log_2(FC) averages	ATF4 log_2(FC) averages	Genes	AhR Log2 FC average	Nrf2 log_2(FC) averages	ATF4 log_2(FC) averages
	CYP1B1	3.56	−0.63	*TPX2*	−0.75	−0.8	−2.38

*Gray background indicates genes that appear in the signature of the pathway from previous studies (**Supplementary Table 1**) and confirmed here. Non-grayed out values are novel allocations from this analysis.*

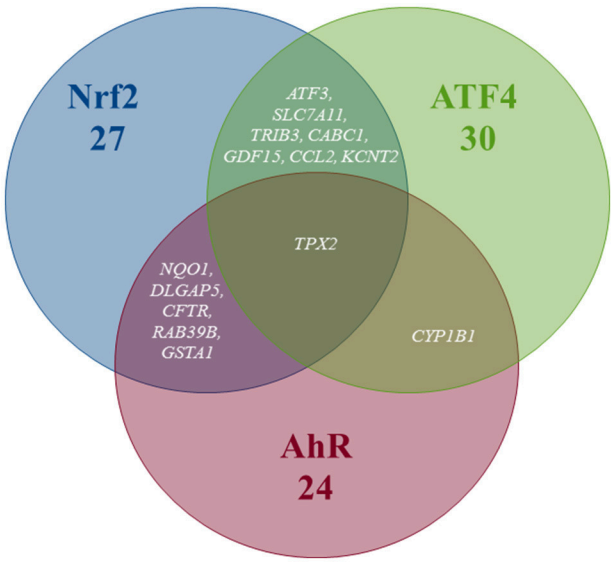


FIGURE 3 | Venn diagram of the number of genes per pathway's global signatures and names of genes of overlapping zones.

Construction of Pathway Signatures

For each of the pathway specific chemicals, all testing conditions were selected. For every gene, the mean of log2(FC) throughout all those conditions was calculated, to form the average activation value of each gene by each of the pathway specific activator. For AhR and Nrf2, the two average activation values obtained (one for each of the pathway specific activator) were themselves averaged. Genes were then sorted in decreasing order of average activation values per pathway. It is important to note that, since the expression of some genes can be inhibited (down regulated) by some chemicals or in certain conditions, some of the average activation values were negative. In order to select the most sensitive genes for each pathway, we computed the mean (μ) and the standard deviation (σ) of the genes' average activation values in each list. A pathways signature was formed by the genes whose average activation values were greater than $\mu + 2\sigma$ or smaller than $\mu - 2\sigma$ for this pathway. Genes appearing in the signature of more than one pathway were set apart in "overlapping signatures."

Furthermore, we stratified signatures by original databases' categories ("Rat liver cells *in vitro,*" "Rat liver cells *in vivo,*" "Human liver cells *in vitro*" etc.) (which correspond to primary cells), to check if there would be any species-specific or *in vitro*/*in vivo* differences among signatures. We chose to work only with liver data since more data were available for liver (602 conditions in kidney vs. 4,083 tested in liver, see **Table 3**). Following the same procedure as above, we constructed pathway signatures for AhR, Nrf2, and ATF4 in each of the following liver categories: (a) Rat liver cells *in vitro*, (b) Rat liver cells *in vivo*, and (c) Human liver cells *in vitro*.

In all cases, general or stratified, some genes were excluded for having no data on effect of the chosen pathway specific chemicals. A list of those genes appears in **Supplementary Table 3.**

A summary of the above-described protocols and the following procedures of Methods are presented in the workflow of **Figure 1**.

Pathway's Signature-Based Prioritization of Chemicals

Among the three liver categories where signatures were stratified, we chose to focus on the "Human liver cells *in vitro*" sub-category exclusively since the ultimate goal of our toxicity pathways' analyses and models is risk assessment of human cells' exposure to xenobiotics. We considered only the genes belonging to the signature of each of the three pathways, but not their overlapping zones. This selection of experimental category and genes reduces the number of studied chemicals from 211 to 160 for the lack of data on the rest of chemicals in this section. Then, for each of the 160 chemicals investigated, we averaged log2(FC) of the pathway signature genes over experimental conditions. Therefore, for each of the three pathways, we obtained a "chemical activation capacity" (CAC) value per chemical. This value reflects how strongly a chemical can activate a given toxicity pathway. Those CACs can be negative for chemicals inhibiting the majority of the genes of a pathway. We used CACs to estimate the pathway's selectivity of chemicals as well as the importance of their impact.

Each chemical can be considered as a point having three CACs as coordinates in a 3-dimensional space which axes correspond to a given pathway. Let us consider a chemical K that has a point in a bi-dimensional graph where the X-axis corresponds to AhR and the Y-axis to Nrf2. In this graph, K's coordinates would be: ($CAC_{AhR,\ K}$, $CAC_{Nrf2,\ K}$), see **Figure 2**. K also defines the vector $\overrightarrow{OK}$ linking the origin O (0, 0) to the point K.

The specificity of a chemical for a given pathway can be measured by the proximity of its point K to the axis representing that pathway. Proximity can be mathematically evaluated by the absolute value of the cosine of the angle (α) between the pathway's axis and $\overrightarrow{OK}$. The more K is specific to AhR, the closer it is to the AhR's axis, the smaller α is, and the bigger **cos** ($\boldsymbol{\alpha}$). In theory, in a 3-dimensional space, a point is closer to an axis than to the two others when its **cos** ($\boldsymbol{\alpha}$) with this axis is greater than $\frac{1}{\sqrt{3}}$. Thus, the value of 0.57735 ($\frac{1}{\sqrt{3}}$) was chosen as a cut-off point for **cos** ($\boldsymbol{\alpha}$). On the other hand, the activation potency of a chemical proportionally increases with the module of the vector $\overrightarrow{OK}$ vector noted $\left\|\overrightarrow{OK}\right\|$ (the distance between the origin and the chemical's point). The value of 0.5 was chosen as a cut-off point for $\left\|\overrightarrow{OK}\right\|$. For instance, chemicals A and B in **Figure 2** are both quite specific of Nrf2, but A's activation potency is relatively limited compared to B's ($\left\|\overrightarrow{OA}\right\| < \left\|\overrightarrow{OB}\right\|$).

Similarly, even though C seems to have a greater activation potency than A and B (greater module), it is equidistant to both axes and therefore is not specific of any of the two pathways. The same logic applies for a 3-dimensional space, adding one extra axis for the ATF4 pathway.

In our signature-based classification of chemicals, for each pathway, after applying the chosen cut-off points, we sorted chemicals by the result of the product $\mathbf{cos}(\boldsymbol{\alpha}) \times \left\|\overrightarrow{OK}\right\|$. Thus,

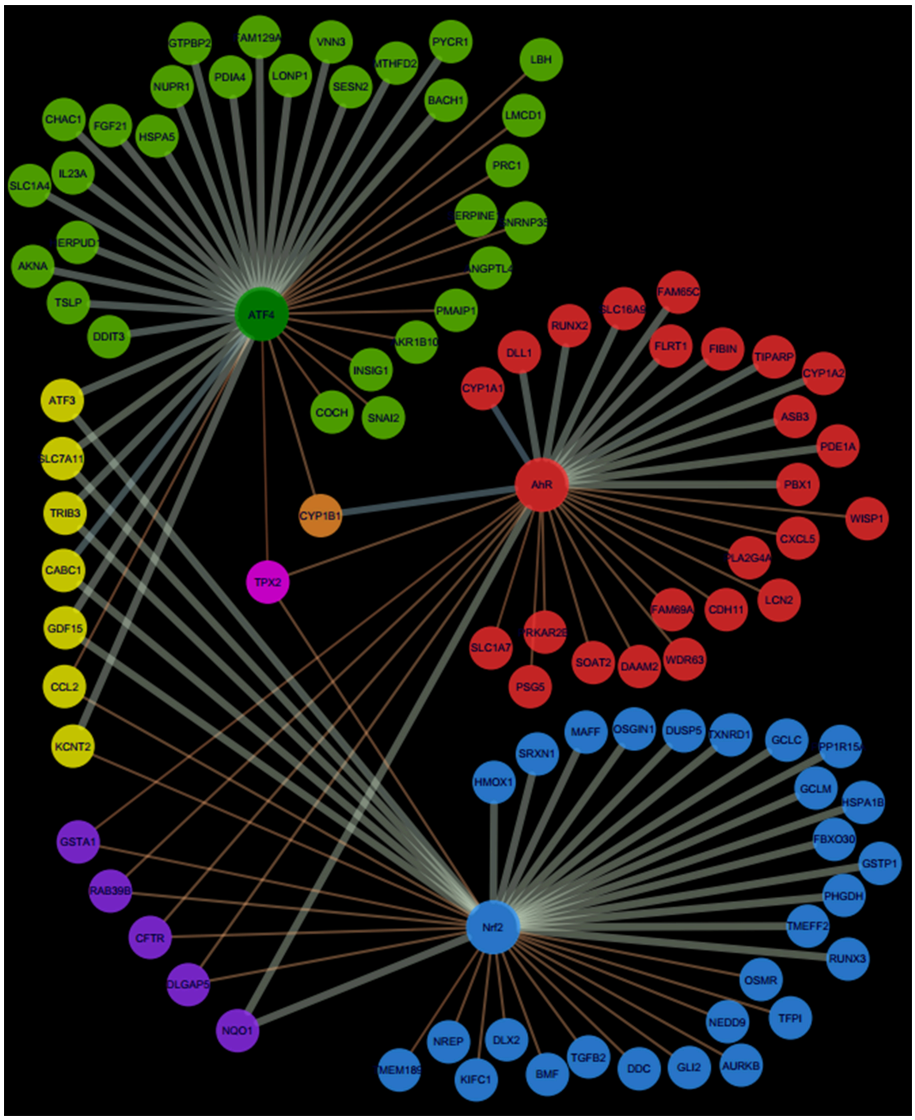


FIGURE 4 | Network representation of AhR, Nrf2 and ATF4 pathway signatures and their overlapping zones.

chemicals which are both pathway specific (high **cos** (**α**)) and potent (high $\left\|\overrightarrow{OK}\right\|$) show up first in our lists.

RESULTS

A visual depiction of the workflow is provided in **Figure 1**.

Pathways' Global Signatures

Pathway's signatures defined on the basis of the whole data set are listed in **Table 4**. Each signature has two parts: "Activated genes" (those having positive log2(FC) averages and are greater than $\mu + 2\sigma$) and "Inhibited genes" (those having negative log2(FC) averages and are smaller than $\mu - 2\sigma$); The two parts are merged in one in the overlapping signatures. In all lists, genes are sorted by the decreasing absolute value of the genes' log2(FC) averages.

The number of genes in the obtained pathway's signature was 24 for AhR, 27 for Nrf2 and 30 for ATF4. In each pathway, at least half (12 for AhR, 15 for Nrf2 and 19 for ATF4) were "Activated genes." The *a priori* pathway is the one for which the gene has come up in PubMed searches; **Table 4** shows that most of activated genes were *a priori* suspected to belong to the target pathway (for example: *CYP1A1*, *RUNX2,* and *CYP1A2* were known to be activated by AhR, *HMOX1* and *SRXN1* by Nrf2 and *DDIT3* and *HERPUD1* by ATF4; those genes are highlighted in gray) while this wasn't the case of the "Inhibited genes" part of the lists. **Figure 3** shows the overlapping zones. Among the five genes that are in the AhR-Nrf2 overlapping zone (*NQO1*, *DLGAP5*, *CFTR*, *RAB39B* and *GSTA1*), only *NQO1* is a mainly activated gene while this was the case of most seven genes of the Nrf2-ATF4 overlapping zone (*ATF3*, *SLC7A11*, *TRIB3*, *CABC1*,

TABLE 5 | AhR, Nrf2 and ATF4 pathways' signatures stratified in liver data and by all liver data sub-categories ("Rat Liver *in vitro*" data, "Rat Liver *in vivo*" data and "Human Liver *in vitro*" data).

	All liver data		Rat liver *in vitro*		Rat liver *in vivo*		Human liver *in vitro*	
	Genes	**Log_2(FC) averages**	**Genes**	**Log_2(FC) averages**	**Genes**	**Log_2(FC) averages**	**Genes**	**Log_2(FC) averages**
AhR SIGNATURES								
Activated genes	*CYP1A1*	4.55	*CYP1A1*	1.30	*CYP1A1*	6.86	*CYP1A1*	4.72
	CYP1A2	1.47			*CYP1A2*	1.71	*CYP1A2*	2.44
	TIPARP	0.64	*TIPARP*	0.40			*TIPARP*	1.21
			ABCC4	0.25	*ABCC4*	0.97		
			IL1R1	0.24	*HTATIP2*	1.19	*CYP1B1*	3.49
			TAF15	0.22			*SLC20A1*	0.78
Inhibited genes			*PRKAR2B*	−0.20			*KCNT2*	−0.60
			ANXA1	−0.18				
			ANGPTL4	−0.17				
Nrf2 SIGNATURES								
Activated genes	*MAFF*	1.42	*MAFF*	0.67	*MAFF*	2.37		
	FBXO30	0.92					*FBXO30*	0.35
	HSPA1B	0.82	*HSPA1B*	0.37			*HSPA1B*	0.63
	PPP1R15A	0.77			*PPP1R15A*	1.16		
	GSTP1	0.67			*GSTP1*	1.24		
	GCLC	0.66	*GCLC*	0.35				
	PSAT1	0.64			*PSAT1*	1.54		
	DUSP5	0.62	*DUSP5*	0.64				
	SLC3A2	0.60			*SLC3A2*	1.09	*SLC3A2*	0.40
	OSGIN1	0.58			*OSGIN1*	0.91	*OSGIN1*	0.42
	SLC6A9	0.57			*SLC6A9*	1.06		
	SLC20A1	0.52	*SLC20A1*	0.41				
	ABCC3	0.52			*ABCC3*	1.00		
			YPEL5	0.47			*YPEL5*	0.37
			CPT1A	0.38			*CPT1A*	0.36
	ASNS	0.75	*SRXN1*	0.66	*HMOX1*	2.03	*ATF5*	0.37
	PHGDH	0.55	*PHLDA1*	0.53	*SLC7A11*	1.74	*AP5Z1*	0.35
	PLA2G12A	0.50	*TXNRD1*	0.41	*GDF15*	1.30		
	SLC7A1	0.48	*ABCC2*	0.39	*BTG2*	0.89		
			PIR	0.34				
			FLVCR2	0.33				
			GSR	0.33				
			GABARAPL1	0.33				
			AGPAT9	0.57				
			TBCEL	0.48				
			MMD	0.33				
Inhibited genes							*MMD*	−0.4
	LCN2	−0.45			*LCN2*	−0.97		
			TGFB2	−0.34			*TGFB2*	−0.44
	MID1IP1	−0.48	*TNFAIP2*	−0.44	*BMF*	−0.88	*ALDH1A1*	−0.61
	IL33	−0.46	*VASN*	−0.39	*DHRS7*	−0.69	*DDC*	−0.42
	NREP	−0.45	*AURKB*	−0.38			*DUT*	−0.35
	SERPINB9	−0.42	*RAB32*	−0.36			*IFIT3*	−0.33
			CD36	−0.36			*UGT1A6*	−0.32
			DCN	−0.34				
			CTSC	−0.34				

(Continued)

TABLE 5 | Continued

	All liver data		Rat liver *in vitro*		Rat liver *in vivo*		Human liver *in vitro*	
	Genes	**Log_2(FC) averages**	**Genes**	**Log_2(FC) averages**	**Genes**	**Log_2(FC) averages**	**Genes**	**Log_2(FC) averages**
			LBH	−0.32				
			CXCL3	−0.32				
ATF4 SIGNATURES								
Activated genes	*TSLP*	1.51					*TSLP*	1.51
	AKNA	1.30					*AKNA*	1.30
	HERPUD1	1.23	*HERPUD1*	1.28	*HERPUD1*	0.61	*HERPUD1*	2.39
	IL23A	1.05	*IL23A*	1.69			*IL23A*	1.86
	HSPA5	0.94					*HSPA5*	3.28
	GTPBP2	0.91	*GTPBP2*	1.12			*GTPBP2*	1.89
	PDIA4	0.87	*PDIA4*	0.92			*PDIA4*	2.18
	FAM129A	0.87					*FAM129A*	2.92
	PYCR1	0.72	*PYCR1*	0.91				
			CHAC1	1.40	*CHAC1*	0.50		
			KLF15	0.81	*KLF15*	0.43		
	SLC1A4	1.15	*TRIB3*	1.12	*HES1*	0.57	*FIBIN*	2.72
	NUPR1	0.94	*BCAT2*	0.97	*USP2*	0.55	*LCN2*	1.91
	LONP1	0.80	*ARHGEF2*	0.93	*ENC1*	0.48	*CTH*	1.62
	VNN3	0.78	*CASP4*	0.84	*TSC22D3*	0.44	*NFE2L1*	1.2
	SESN2	0.75	*KLF4*	0.82	*DDIT4*	0.39		
	BACH1	0.68	*BET1*	0.82	*SLC38A2*	0.38		
			WARS	0.80	*IP6K2*	0.62		
			PCK2	0.73				
			SLC25A33	0.71				
			SLC7A5	0.71				
			ACOT2	0.83				
			MANEA	0.75				
Inhibited genes	*PRC1*	−0.65	*PRC1*	−0.61				
	LMCD1	−0.64	*LMCD1*	−0.80			*LMCD1*	−1.73
	LBH	−0.61					*LBH*	−2.56
	SNAI2	−1.20	*DPYSL2*	−0.98	*FOXA2*	−0.61	*FRMD6*	−1.52
	AKR1B10	−0.96	*DUSP6*	−0.97	*ABCG2*	−0.49	*SLC39A10*	−1.35
	PMAIP1	−0.88	*IFIT3*	−0.72	*NEDD9*	−0.43	*GPNMB*	−1.26
	SNRNP35	−0.77	*EMILIN1*	−0.69	*TMEM159*	−0.37	*ANKRD1*	−1.16
	SERPINE1	−0.68	*FCER1G*	−0.65			*PHLDA1*	−1.16
			SQRDL	−0.61				
			IFI44	−0.61				

Genes that appear in more than one column are highlighted in gray.

GDF15) with two exceptions (*CCL2* has negative averages for both pathways and *KCNT2* for Nrf2). *CYP1B1* is the only mutual gene for AhR (strong activation) and ATF4 (inhibition) and *TPX2* is the only mutual gene for all three pathways (inhibition). **Figure 4** shows a network representation of the three signatures and their overlapping zones.

Pathways' Stratified Signatures in Liver

The Three Main Pathways' Stratified Signatures in Liver

Table 5 shows the stratified signatures in liver of each pathway in four columns (categories): each containing the genes' names and their log2(FC) averages. Genes that appear in more than one column are highlighted in gray and empty lines were left in order to display those genes on the same line in all the categories where they appear. Genes of the first column, sorted by the decreasing absolute values of their log2(FC) averages, appear first, followed by genes appearing in more than one category but not the first column and then the rest of the genes sorted by the decreasing absolute values of their log2(FC) averages as well.

AhR stratified signatures

Table 5 shows that *CYP1A1* is clearly, by far the most activated gene in this pathway. Three other genes appear in the AhR

signature in more than one column: *CYP1A2* everywhere except "Rat liver *in vitro*," *TIPARP* everywhere except "Rat liver *in vivo*" and *ABCC4* shows up in these two categories only. "Rat liver *in vitro*" AhR signature is completed by five additional genes, "Rat liver *in vivo*" by one more and "Human liver *in vitro*" by three.

Nrf2 stratified signatures

Nrf2 signatures are bigger: 22 genes in the all liver data signature, 28 for "Rat Liver *in vitro*" and 15 for each of "Rat Liver *in vivo*" and "Human Liver *in vitro*". Around two third of those genes are "Activated genes" and the rest have negative log2(FC) averages. *MAFF*, *SLC3A2*, *OSGIN2* are among the "Activated genes" that appear in three out of the four categories we are studying. Other important genes show up in two columns (*HSPA1B*, *PPP1R15A*, and *GCLC*) and some, in only one (*SRXN1* in "Rat Liver *in vitro*" and *HMOX1* in "Rat Liver *in vivo*"). The values of the "Rat liver *in vivo*" are also higher than the "Rat liver *in vitro*" and "Human liver *in vitro*" categories.

ATF4 stratified signatures

ATF4 signatures size is similar to Nrf2's signatures with a comparable proportion of activated genes: 23 genes in the all liver data signature, 28 for "Rat liver *in vitro*" and 14 for each of "Rat liver *in vivo*" and 19 for "Human liver *in vitro*." *HERPUD1* is an important gene in this pathway; it is part of the signature of every single category we are examining and exhibits values as high as 2.39 in "Human Liver *in vitro*" (among the highest in ATF4 signatures). Other genes also are present in the majority of the categories: *IL23A*, *GTPBP2,* and *PDIA4*. It is noteworthy that the ATF4 signature of "Rat Liver *in vivo*" results don't have a lot in common with the other three categories and its log2(FC) averages are lower than the rest (the highest value is 0.61 for *HERPUD1*).

The Overlapping Zones Stratified Signatures

Figure 5 shows that the AhR-ATF4 overlapping zone is the least populated (four genes maximum in all liver data, no genes for "Rat Liver *in vivo*" and two genes in the two other categories). The number of genes in the AhR-Nrf2 overlapping signatures ranges from four to eight, with many typical key Nrf2 genes

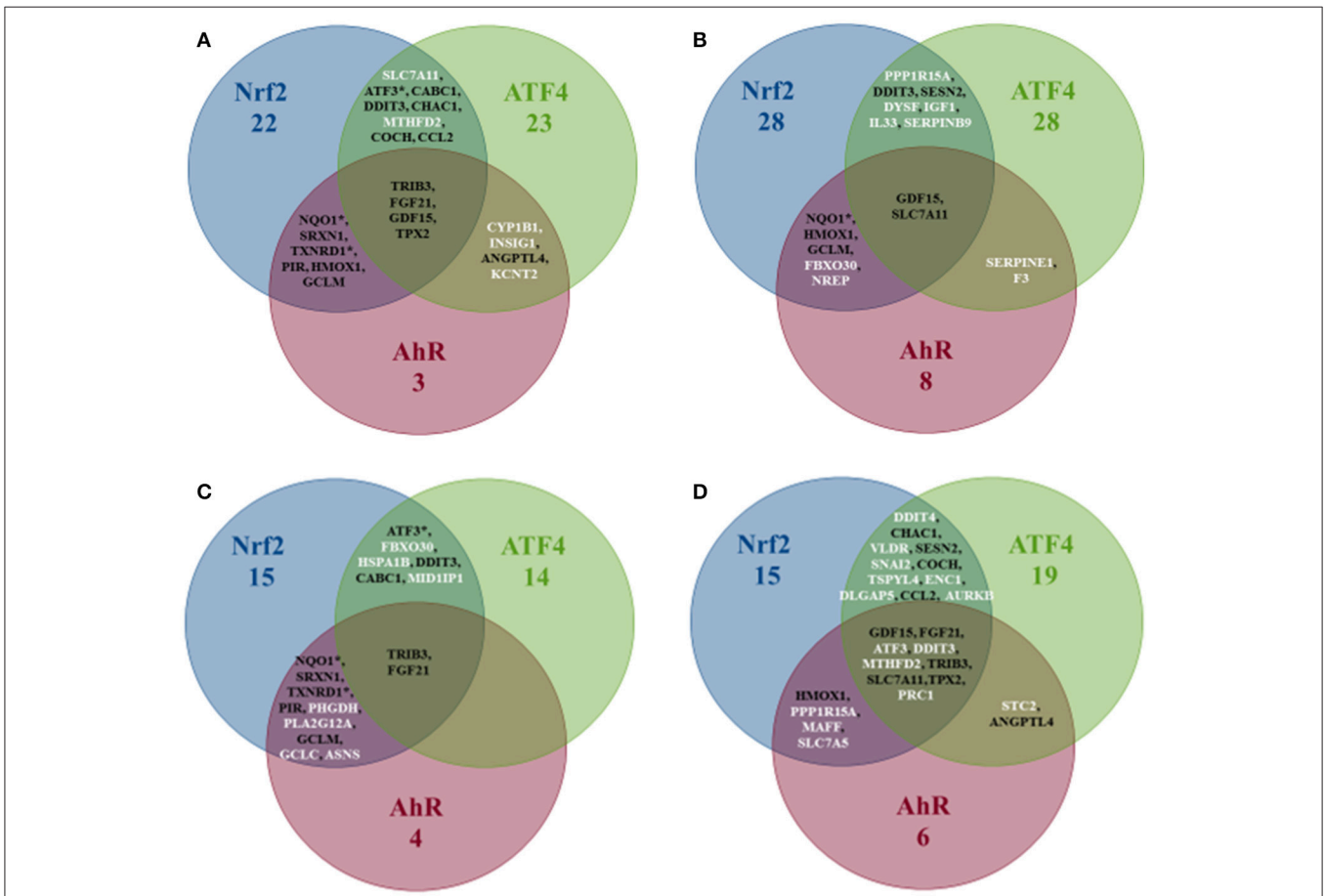


FIGURE 5 | Venn diagram of the number of genes per pathway's stratified signatures and names of genes of overlapping zones. Categories: **(A)** All liver data, **(B)** Rat Liver *in vitro* data, **(C)** Rat Liver *in vivo* data, **(D)** Human Liver *in vitro* data. *Refers to genes that were known to be part of the same overlapping zone according to **Supplement Table 1** lists. White is the color of gene names that appear in an overlapping zone of only one of the four categories studied, and black is the color of gene names that appear in more than one category (two, three or four).

(*NQO1*, *SRXN1*, *HMOX1*, *TXNRD1,* and *GCLM*) appearing in more than one category. The Nrf2-ATF4 overlapping signatures contain six to eleven genes (*DDIT3*, *ATF3,* and *CHAC1* are among the repetitive genes). Finally, *TRIB3*, *FGF21*, *GDF15*, *SLC7A11,* and *TPX2* are in the signature of the zone mutual to all three pathways for at least two of the four categories studied.

Pathways' Stratified Signatures in Liver

Figures 6, **7**, **8** plot the 160 chemicals' vector modules vs. the absolute value of **cos** (**α**), which represents the pathway activation scores of chemicals that activate each pathway both selectively and strongly. Chemicals are represented by a number that corresponds to their rank in the alphabetically ordered list. The blue dashed lines mark the vertical ($\mathbf{cos}(\boldsymbol{\alpha}) = \frac{\mathbf{1}}{\sqrt{\mathbf{3}}}$) and horizontal ($\left\|\overrightarrow{\boldsymbol{OK}}\right\| = \mathbf{0.5}$) limits we set. The number chemicals that are off these limits is 34 for AhR, one for Nrf2 and four for ATF4; these chemicals are in red and their names are listed in the legend on the right by the order of the decreased values of the product result $\mathbf{cos}(\boldsymbol{\alpha}) \times \left\|\overrightarrow{\boldsymbol{OK}}\right\|$. As we can see in these figures' legends, "pathway specific activators" show up first in the lists of AhR (Omeprazole)

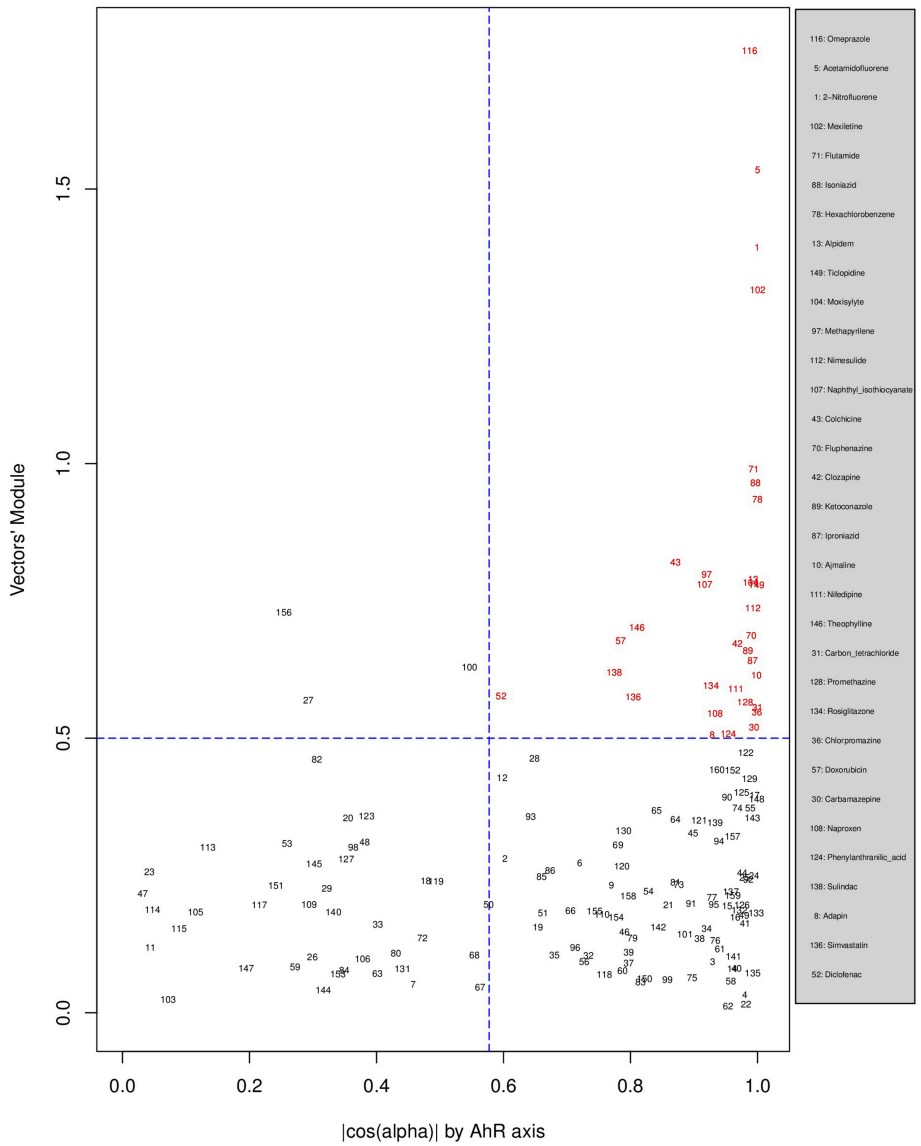


FIGURE 6 | Distribution of chemicals by potency (Y-axis: module $\left\|\overrightarrow{\mathbf{OK}}\right\|$ of the vector linking the origin O(0,0) to the chemical's point in a 3D space) and specificity to the AhR pathway (X-axis: the absolute value of the $|\mathbf{cos}(\boldsymbol{\alpha})|$ of the angle between $\overrightarrow{\mathbf{OK}}$ and the AhR axis in a 3D space). Chemicals are represented by their rank in the alphabetically ordered list. Chemicals that are both strong (horizontal blue dashed line: $\left\|\overrightarrow{\boldsymbol{OK}}\right\| > \mathbf{0.5}$) and AhR specific (vertical blue dashed line: $\mathbf{cos}(\boldsymbol{\alpha}) = \frac{\mathbf{1}}{\sqrt{\mathbf{3}}}$) are in red and their names are listed in the legend on the right.

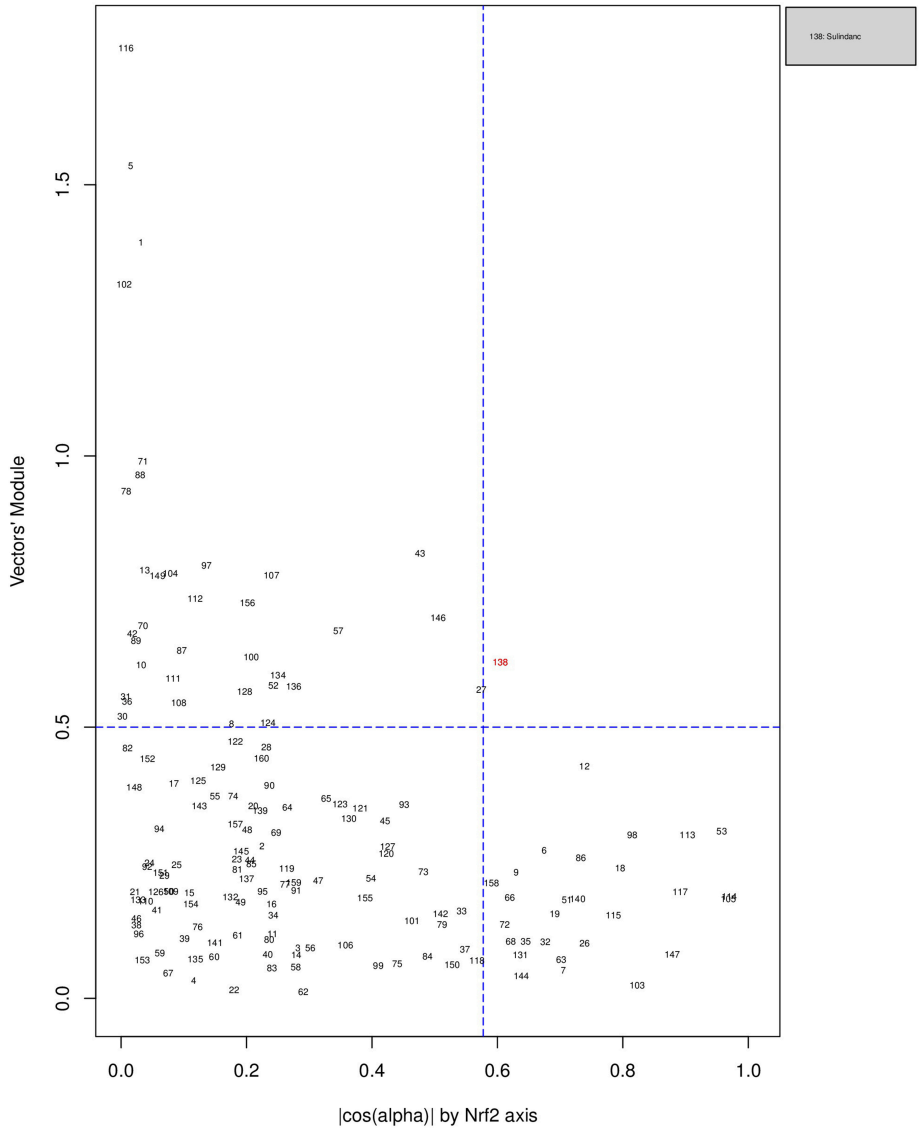


FIGURE 7 | Distribution of chemicals by potency (Y-axis: module $\left\|\overrightarrow{OK}\right\|$ of the vector linking the origin O(0,0) to the chemical's point in a 3D space) and specificity to the Nrf2 pathway (X-axis: the absolute value of the $|\cos(\alpha)|$ of the angle between $\overrightarrow{OK}$ and the AhR axis in a 3D space). Chemicals are represented by their rank in the alphabetically ordered list. The only chemical that is both strong (horizontal blue dashed line: $\left\|\overrightarrow{OK}\right\| > 0.5$) and AhR specific (vertical blue dashed line: $\cos(\alpha) = \frac{1}{\sqrt{3}}$) Sulindac, is in red and it is listed in the legend on the right.

and ATF4 (Tunicamycin), but do not appear at all in the list of Nrf2 (Phorone).

DISCUSSION

Nrf2, ATF4, and AhR are important TFs in toxicological contexts and have well described downstream gene targets (Jennings et al., 2013). Each of these TFs have distinct unrelated upstream activation points, unique gene targets, but also have direct (i.e., via multiple upstream promoter regions) and likely indirect overlaps on some specific gene targets. The AhR protein is a cytosolic protein receptor, where activation via chemical ligand binding causes nuclear translocation, DNA binding to it consensus sequence and RNA transcription. Several toxic compounds including dioxin-like compounds activate AhR. The TF Nrf2 is liberated from its cytosolic inhibitor KEAP1, where the latter is sensitive to electrophiles and ROS. The TF ATF4 is activated via PERK, where PERK is activated when its inhibitor BiP, dissociates from PERK to bind unfolded proteins. All sorts of Endoplasmic Reticulum disturbances can cause an increase in unfolded proteins.

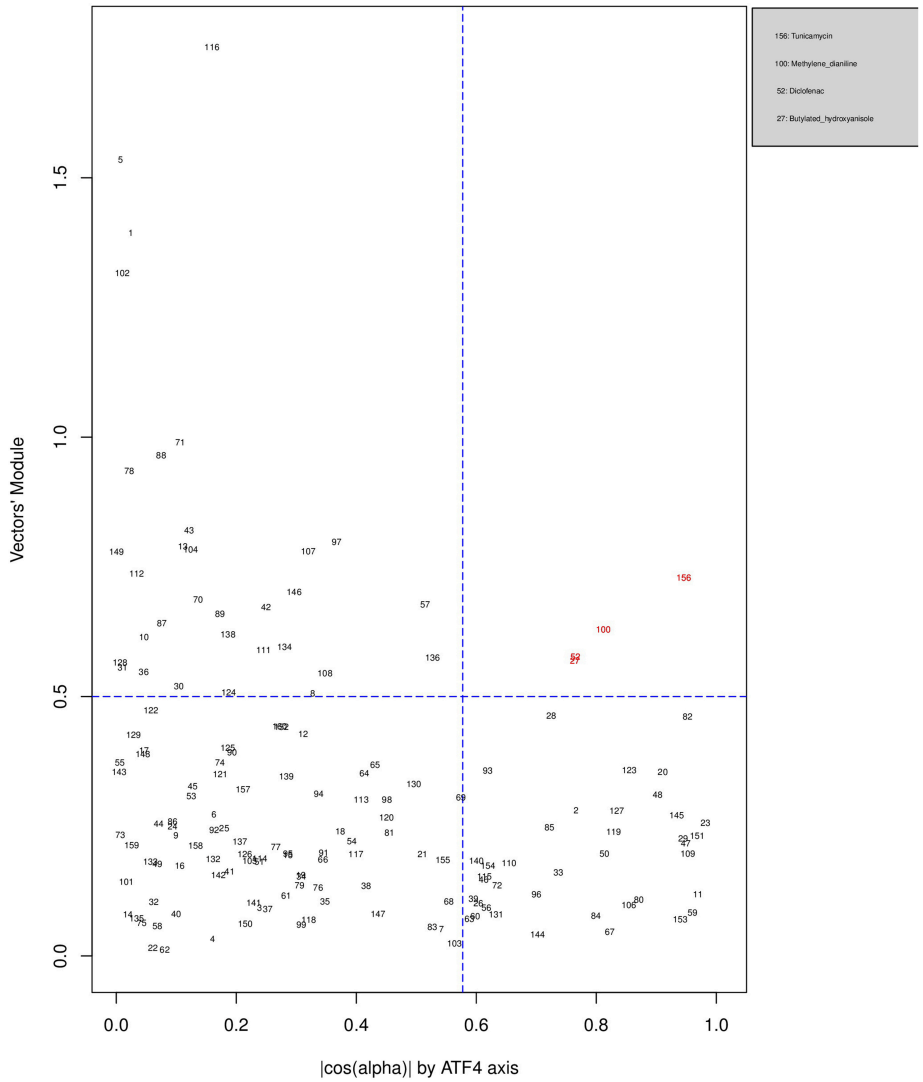


FIGURE 8 | Distribution of chemicals by potency (Y-axis: module $\left\|\overrightarrow{OK}\right\|$ of the vector linking the origin O(0,0) to the chemical's point in a 3D space) and specificity to the ATF4 pathway (X-axis: the absolute value of the $|\cos(\alpha)|$ of the angle between $\overrightarrow{OK}$ and the AhR axis in a 3D space). Chemicals are represented by their rank in the alphabetically ordered list. Chemicals that are both strong (horizontal blue dashed line: $\left\|\overrightarrow{OK}\right\| > 0.5$) and AhR specific (vertical blue dashed line: $\cos(\alpha) = \frac{1}{\sqrt{3}}$) are in red and their names are listed in the legend on the right.

Using multiple toxicogenomic databases we investigated the most appropriate activators of these three pathways, where it is expected that the chemical does not directly activate the other two pathways. These compounds were, Benzo(a)pyrene and Omeprazole for AhR, Potassium Bromate and Phorone for Nrf2 and Tunicamycin A for ATF4. All conditions up to and including 24 h were pooled to generate a list of genes allocated to the three pathways (**Table 4**). This list confirmed the majority of a *priori* literature based information of "Activated genes" (i.e., upregulated). Although some genes were now reallocated to different pathways. The overlap with "Inhibited genes" (i.e., down regulated), was much poorer. This is too be expected as TF activated gene down regulation is much more complex and is often due to competition for auxiliary transcription facilitating proteins. Cytochrome P450 1A was the central element of the AhR pathway: *CYP1A1* is the most prominent gene of this pathway, regardless of the experimental category, followed by *CYP1A2*. These findings are similar to previous investigations and have been implemented in a systems biology model (Hamon et al., 2014). For the Nrf2 pathway, the prototypical Nrf2 genes (*HMOX1*, *SRXN1,* and *GCLM*) appear in the Nrf2 signature of all datasets, but also in the AhR-Nrf2 overlapping signature for most liver categories. This may reflect the fact that several AhR agonists are themselves metabolized to reactive chemicals via AhR dependent CYP expression. For example benzo(a)pyrene is a substrate of the CYP1 sub family of cytochrome P450 enzymes,

and it promotes its own metabolism to reactive epoxide and quinone products (Gelboin, 1980). These metabolic products can lead to oxidative stress and to an activation of the Nrf2 pathway as part of a second line of responses (Burchiel and Luster, 2001). The only activated gene that appears in the ATF4 signature of each of the three studied categories is HERPUD1. In most cases, HERPUD1 also had the highest log2(FC) averages. Overlapping zones show an interaction between AhR and Nrf2, between Nrf2 and ATF4, but a very limited or non-existent interaction between AhR and ATF4 pathways.

We have used the exclusive pathway genes to create pathway chemical activation capacity (CAC) scores. The CAC reflects both specificity for the pathway (**cos** (**α**)) and the activation potency $\left\|\overrightarrow{OK}\right\|$. CAC scores were generated for 160 chemicals using the TG-GATEs liver data. For ATF4, tunicamycin, methylene dianiline, diclofenac, and butylated hydroxyanisole were ranked highest, in that order. Tunicamycin was used as a specific ATF4 specific activator. Both diclofenac and butylated hydroxyanisole have previously been demonstrated to positive modulate the ATF4 pathway (Afonyushkin et al., 2010; Fredriksson et al., 2014). The molecular mechanism for methylene dianiline has not been fully elucidated and this evidence would suggest an ER disturbance and/or proteotoxic mechanism. For AhR, 34 chemicals were considered positive by CAC scores. Omeprazole was ranked highest, followed by acetamidofluorene, 2-Nitrofluorene, mexiletine, flutamide, isoniazid, and hexachlorobenzene. Many of the 34 chemicals have not been previously linked with AhR, but several are. These include, hexachlorobenzene (Randi et al., 2008; de Tomaso Portaz et al., 2015), ketoconazole (Novotna et al., 2014), clozapine (Donohoe et al., 2008), and doxorubicin (Volkova et al., 2011). Fluphenazine has not been established as a ligand for the AhR, its structure—a halogenated aromatic ring system—closely matches the motif involved in binding to this receptor (Donohoe et al., 2008). In a recent study we have demonstrated that isoniazid induced CYP1A1 in HepaRG cells, which is a potential indicator of AhR activation (Limonciel et al., 2018). Only Sulindac from the 160 was ranked as active using the CAC selection criteria, which may seem surprising given the frequency of oxidative injury in liver toxicities. Although butylated hydroxyanisole was marginal. The reason for a lack of Nrf2 activation prediction might be simply due to the fact that none of the 160 compounds, including the positive compound phorone cause an Nrf2 response in the liver within the first 24 h. Another possibility is that removing the overlapping genes has weakened the ability to pick up this pathway. Indeed, this is a weakness in the overall strategy as it is difficult to determine in such data sets if the pathways themselves are co-regulated since there are several gene overlaps amongst the pathways.

SUMMARY AND CONCLUSION

The size of the data set, its multiple sources, abundancy of compounds, concentrations and time of exposures, *in vitro* and *in vivo*, different organs are both a blessing and a curse. On the one hand, it is generally an advantage to have as broad as data set as possible, but the different sizes and focuses of the individual data sets/studies meant we needed to reduce the data to the lowest denomination. Another major issue was the low abundance of well described pathway activators. Despite these issues we have made some interesting observations and have developed a method to quantify a chemical's capacity to activate one three pathways.

We uncovered variations in AhR, ATF4 and Nrf2 signatures across tissues, compounds, species and *in vivo* vs. *in vitro*. Some of these alterations are likely to be linked to pharmacokinetics, including distribution and metabolism, others may be linked to tissue specific regulation of these pathways. While some genes were very variable across experimental conditions, some were extremely robust, for example CYP1A1 in the AhR pathway and HERPUD1 in the ATF4 pathway. Some genes swing between a pathway's specific signature and overlapping zones for example *GCLC* between Nrf2 and AhR-Nrf2. Others are regularly on overlapping signatures for example *TPX2* and *TRIB3*. However, it is not possible with this type of analysis to delineate whether these overlaps are solely on a gene level or also on the pathway level.

The CAC score system developed, based on $\cos(\alpha) \times \left\|\overrightarrow{OK}\right\|$, can be used to quantify a chemical's specificity and potency to selectively activate one of these pathways. However, future work will be required to validate and optimize the gene signatures utilized.

DATA AVAILABILITY STATEMENT

The dataset used for this analysis can be found in an Excel document in the **Supplementary Material**.

AUTHOR CONTRIBUTIONS

FB and PJ conceived of the original idea and supervised the findings of this work. EZ, FB, PJ, AL, and AW planned the simulations. FB and PJ developed the theoretical formalism and verified the analytical methods. Supervised by PJ and FB, EZ performed the analytic computations and the numerical simulations. SW and BvdW generated target gene lists. XJ and AK-S designed the figures. EZ wrote the manuscript with support from FB, PJ, and AL and then all authors reviewed and commented the final manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

ACKNOWLEDGMENTS

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking (IMIJU) under grant agreement number 115439, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. This work was supported by the 2015 CEFIC-LRI award (AL) and partly supported by the EU-ToxRisk project (An Integrated European "Flagship" Program Driving Mechanism-Based Toxicity Testing and Risk Assessment

for the 21st Century) funded by the European Commission under the Horizon 2020 programme (Grant Agreement No. 681002). This publication reflects only the author's views and neither the IMI JU nor EFPIA nor the European Commission are liable for any use that may be made of the information contained therein.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00429/full#supplementary-material

REFERENCES

Afonyushkin, T., Oskolkova, O. V., Philippova, M., Resink, T. J., Erne, P., Binder, B. R., et al. (2010). Oxidized phospholipids regulate expression of ATF4 and VEGF in endothelial cells via NRF2-dependent mechanism: novel point of convergence between electrophilic and unfolded protein stress pathways. *Arterioscler. Thromb. Vasc. Biol.* 30, 1007–1013. doi: 10.1161/ATVBAHA.110.204354

Aschauer, L., Limonciel, A., Wilmes, A., Stanzel, S., Kopp-Schneider, A., Hewitt, P., et al. (2015). Application of RPTEC/TERT1 cells for investigation of repeat dose nephrotoxicity: a transcriptomic study. *Toxicol. in vitro* 30, 106–116. doi: 10.1016/j.tiv.2014.10.005

Ballmaier, D., and Epe, B. (1995). Oxidative DNA damage induced by potassium bromate under cell-free conditions and in mammalian cells. *Carcinogenesis* 16, 335–342. doi: 10.1093/carcin/16.2.335

Bassik, M. C., and Kampmann, M. (2011). Knocking out the door to tunicamycin entry. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11731–11732. doi: 10.1073/pnas.1109035108

Burchiel, S. W., and Luster, M. I. (2001). Signaling by environmental polycyclic aromatic hydrocarbons in human lymphocytes. *Clin. Immunol.* 98, 2–10. doi: 10.1006/clim.2000.4934

Crean, D., Bellwon, P., Aschauer, L., Limonciel, A., Moenks, K., Hewitt, P., et al. (2015). Development of an *in vitro* renal epithelial disease state model for xenobiotic toxicity testing. *Toxicol. in vitro* 30, 128–137. doi: 10.1016/j.tiv.2014.11.015

de Tomaso Portaz, A. C., Caimi, G. R., Sánchez, M., Chiappini, F., Randi, A. S., Kleiman de Pisarev, D. L., et al. (2015). Hexachlorobenzene induces cell proliferation, and aryl hydrocarbon receptor expression (AhR) in rat liver preneoplastic foci, and in the human hepatoma cell line HepG2. AhR is a mediator of ERK1/2 signaling, and cell cycle regulation in HCB-treated HepG2 cells. *Toxicology* 336, 36–47. doi: 10.1016/j.tox.2015.07.013

Donohoe, D. R., Weeks, K., Aamodt, E. J., and Dwyer, D. S. (2008). Antipsychotic drugs alter neuronal development including ALM neuroblast migration and PLM axonal outgrowth in *Caenorhabditis elegans*. *Int. J. Dev. Neurosci.* 26, 371–380. doi: 10.1016/j.ijdevneu.2007.08.021

Fredriksson, L., Wink, S., Herpers, B., Benedetti, G., Hadi, M., de Bont, H., et al. (2014). Drug-induced endoplasmic reticulum and oxidative stress responses independently sensitize toward TNFα-mediated hepatotoxicity. *Toxicol. Sci.* 140, 144–159. doi: 10.1093/toxsci/kfu072

Gelboin, H. V. (1980). Benzo[alpha]pyrene metabolism, activation and carcinogenesis: role and regulation of mixed-function oxidases and related enzymes. *Physiol. Rev.* 60, 1107–1166. doi: 10.1152/physrev.1980.60.4.1107

Haarmann-Stemmann, T., Aarbakke, J., Fritsche, E., and Krutmann, J. (2012). The AhR–Nrf2 pathway in keratinocytes: on the road to chemoprevention? *J. Invest. Dermatol.* 132, 7–9. doi: 10.1038/jid.2011.359

Hamon, J., Jennings, P., and Bois, F. Y. (2014). Systems biology modeling of omics data: effect of cyclosporine a on the Nrf2 pathway in human renal cells. *BMC Syst. Biol.* 8:76. doi: 10.1186/1752-0509-8-76

Hochberg, Y., and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat. Med.* 9, 811–818. doi: 10.1002/sim.4780090710

Howden, C. W. (1991). Clinical pharmacology of omeprazole. *Clin. Pharmacokinet.* 20, 38–49. doi: 10.2165/00003088-199120010-00003

Iannone, A., Tomasi, A., Vannini, V., and Swartz, H. M. (1990). Metabolism of nitroxide spin labels in subcellular fractions of rat liver. II. Reduction in the cytosol. *Biochim. Biophys. Acta* 1034, 290–293. doi: 10.1016/0304-4165(90)90053-Y

Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., et al. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43, D921–D927. doi: 10.1093/nar/gku955

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatatics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Jennings, P., Limonciel, A., Felice, L., and Leonard, M. (2013). An overview of transcriptional regulation in response to toxicological insult. *Arch. Toxicol.* 87, 49–72. doi: 10.1007/s00204-012-0919-y

Jin, U.-H., Lee, S., and Safe, S. (2012). Aryl hydrocarbon receptor (AHR)-active pharmaceuticals are selective AHR modulators in MDA-MB-468 and BT474 breast cancer cells. *J. Pharmacol. Exp. Ther.* 343, 333–341. doi: 10.1124/jpet.112.195339

Jin, U.-H., Lee, S.-O., Pfent, C., and Safe, S. (2014). The aryl hydrocarbon receptor ligand omeprazole inhibits breast cancer cell invasion and metastasis. *BMC Cancer* 14:498. doi: 10.1186/1471-2407-14-498

Kong, Y., Trabucco, S. E., and Zhang, H. (2014). "Oxidative Stress, Mitochondrial Dysfunction and the Mitochondria Theory of Aging," in *Interdisciplinary Topics in Gerontology*, eds. L. Robert and T. Fulop (Basel: S. KARGER AG), 86–107.

Leonard, M. O., Limonciel, A., and Jennings, P. (2014). "Stress response pathways," in *In Vitro Models for Ototoxic Research* (New York, NY: Springer New York), 433–458.

Limonciel, A., Ates, G., Carta, G., Wilmes, A., Watzele, M., Shepard, P. J., et al. (2018). Comparison of base-line and chemical-induced transcriptomic responses in HepaRG and RPTEC/TERT1 cells using TempO-Seq. *Arch. Toxicol.* 92, 2517–2531. doi: 10.1007/s00204-018-2256-2

Limonciel, A., Moenks, K., Stanzel, S., Truisi, G. L., Parmentier, C., Aschauer, L., et al. (2015). Transcriptomics hit the target: Monitoring of ligand-activated and stress response pathways for chemical testing. *Toxicol. in vitro* 30, 7–18. doi: 10.1016/j.tiv.2014.12.011

Limonciel, A., Wilmes, A., Aschauer, L., Radford, R., Bloch, K. M., McMorrow, T., et al. (2012). Oxidative stress induced by potassium bromate exposure results in altered tight junction protein expression in renal proximal tubule cells. *Arch. Toxicol.* 86, 1741–1751. doi: 10.1007/s00204-012-0897-0

Mueller, S. O., Dekant, W., Jennings, P., Testai, E., and Bois, F. (2015). Comprehensive summary–predict-IV: a systems toxicology approach to improve pharmaceutical drug safety testing. *Toxicol. in vitro* 30, 4–6. doi: 10.1016/j.tiv.2014.09.016

Nebert, D. W., Dalton, T. P., Okey, A. B., and Gonzalez, F. J. (2004). Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer. *J. Biol. Chem.* 279, 23847–23850. doi: 10.1074/jbc.R400004200

Novotna, A., Korhonova, M., Bartonkova, I., Soshilov, A. A., Denison, M. S., Bogdanova, K., et al. (2014). Enantiospecific effects of ketoconazole on aryl hydrocarbon receptor. *PLoS ONE* 9:e101832. doi: 10.1371/journal.pone.0101832

Oguro, T., Hayashi, M., Numazawa, S., Asakawa, K., and Yoshida, T. (1996). Heme oxygenase-1 gene expression by a glutathione depletor, phorone, mediated through AP-1 activation in rats. *Biochem. Biophys. Res. Commun.* 221, 259–265. doi: 10.1006/bbrc.1996.0583

Oslowski, C. M., and Urano, F. (2011). Measuring ER stress and the unfolded protein response using mammalian tissue culture system. *Meth. Enzymol.* 490, 71–92. doi: 10.1016/B978-0-12-385114-7.00004-0

Randi, A. S., Sanchez, M. S., Alvarez, L., Cardozo, J., Pontillo, C., and Kleiman de Pisarev, D. L. (2008). Hexachlorobenzene triggers AhR translocation to the nucleus, c-Src activation and EGFR transactivation in rat liver. *Toxicol. Lett.* 177, 116–122. doi: 10.1016/j.toxlet.2008.01.003

Smyth, G. K., Michaud, J., and Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* 21, 2067–2075. doi: 10.1093/bioinformatics/bti270

Taguchi, K., Motohashi, H., and Yamamoto, M. (2011). Molecular mechanisms of the Keap1-Nrf2 pathway in stress response and cancer evolution: molecular mechanisms of the Keap1-Nrf2 pathway. *Genes Cells* 16, 123–140. doi: 10.1111/j.1365-2443.2010.01473.x

Vinken, M., Doktorova, T., Ellinger-Ziegelbauer, H., Ahr, H.-J., Lock, E., Carmichael, P., et al. (2008). The carcinoGENOMICS project: critical selection of model compounds for the development of omics-based *in vitro* carcinogenicity screening assays. *Mutat. Res.* 659, 202–210. doi: 10.1016/j.mrrev.2008.04.006

Volkova, M., Palmeri, M., Russell, K. S., and Russell, R. R. (2011). Activation of the aryl hydrocarbon receptor by doxorubicin mediates cytoprotective effects in the heart. *Cardiovasc. Res.* 90, 305–314. doi: 10.1093/cvr/cvr007

Wilmes, A., Bielow, C., Ranninger, C., Bellwon, P., Aschauer, L., Limonciel, A., et al. (2014). Mechanism of cisplatin proximal tubule toxicity revealed by integrating transcriptomics, proteomics, metabolomics and biokinetics. *Toxicol. in vitro* 30, 117–127. doi: 10.1016/j.tiv.2014.10.006

Wilmes, A., Limonciel, A., Aschauer, L., Moenks, K., Bielow, C., Leonard, M. O., et al. (2013). Application of integrated transcriptomic, proteomic and metabolomic profiling for the delineation of mechanisms of drug induced cell stress. *J. Proteomics* 79, 180–194. doi: 10.1016/j.jprot.2012.11.022

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., et al. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 8, 625–637. doi: 10.1089/106652701753307520

Younes, M., Sharma, S. C., and Siegers, C. P. (1986). Glutathione depletion by phorone organ specificity and effect on hepatic microsomal mixed-function oxidase system. *Drug Chem. Toxicol.* 9, 67–73. doi: 10.3109/01480548609042831

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 October 2018
doi: 10.3389/fgene.2018.00484

Network and Pathway Analysis of Toxicogenomics Data

*Gal Barel and Ralf Herwig**

Department Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

Toxicogenomics is the study of the molecular effects of chemical, biological and physical agents in biological systems, with the aim of elucidating toxicological mechanisms, building predictive models and improving diagnostics. The vast majority of toxicogenomics data has been generated at the transcriptome level, including RNA-seq and microarrays, and large quantities of drug-treatment data have been made publicly available through databases and repositories. Besides the identification of differentially expressed genes (DEGs) from case-control studies or drug treatment time series studies, bioinformatics methods have emerged that infer gene expression data at the molecular network and pathway level in order to reveal mechanistic information. In this work we describe different resources and tools that have been developed by us and others that relate gene expression measurements with known pathway information such as over-representation and gene set enrichment analyses. Furthermore, we highlight approaches that integrate gene expression data with molecular interaction networks in order to derive network modules related to drug toxicity. We describe the two main parts of the approach, i.e., the construction of a suitable molecular interaction network as well as the conduction of network propagation of the experimental data through the interaction network. In all cases we apply methods and tools to publicly available rat *in vivo* data on anthracyclines, an important class of anti-cancer drugs that are known to induce severe cardiotoxicity in patients. We report the results and functional implications achieved for four anthracyclines (doxorubicin, epirubicin, idarubicin, and daunorubicin) and compare the information content inherent in the different computational approaches.

Keywords: network analysis, protein–protein interaction network, pathways, drug toxicity, toxicogenomics, transcriptomics, anthracyclines

OPEN ACCESS

Edited by:
Danyel Jennen,
Maastricht University, Netherlands

Reviewed by:
Mohamed Diwan M. AbdulHameed,
Independent Researcher, Frederick, MD, United States
Richard John Brennan,
Sanofi, United States

***Correspondence:**
Ralf Herwig
herwig@molgen.mpg.de

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Genetics

Received: *19 June 2018*
Accepted: *28 September 2018*
Published: *22 October 2018*

Citation:
Barel G and Herwig R (2018) Network and Pathway Analysis of Toxicogenomics Data. Front. Genet. 9:484. doi: 10.3389/fgene.2018.00484

INTRODUCTION

To thoroughly study the mechanisms behind drug induced toxicity a robust analysis by means of computational methods is crucial (Liebler and Guengerich, 2005). Understanding the influence of the compounds on different biological processes is complex and requires sophisticated interpretations of the data. In the field of toxicogenomics transcriptome data, that were collected upon drug treatment and that reflect gene expression levels in response to it, is in the focus of the analysis. Various studies, both *in vitro* and *in vivo*, focusing on different compounds and organs, have been already carried out (Hartung, 2009). Most of the studies were based on microarray technology (Mei et al., 2010), even though newer technologies, such as high-throughput sequencing

(RNA-seq), are already in use in other research areas. Such transcriptomic profiles have previously been used for predicting toxic drug effects (Gusenleitner et al., 2014; Kohonen and Parkkinen, 2017; Nystrom-Persson et al., 2017; Rueda-Zarate et al., 2017), but further analysis for identifying the functional and molecular mechanisms behind the toxic effects is still much needed.

Here, we describe different approaches for the analysis of toxicogenomics data at the molecular network and pathway levels. We use publicly available data from microarray experiments and perform a differential expression analysis in order to identify the genes that are up- or down-regulated due to the administration of the drug (DEGs: differentially expressed genes). Suggested methods and tools include (i) over-representation analysis, (ii) gene set enrichment (pathway) analysis, and (iii) network propagation. All methods are complementary and deliver different and complementary mechanistic views on drug action and drug effects derived from the underlying gene expression data. Over-representation analysis provides a first impression on which pathways and biological functions are involved in the cell's response to the drug. This kind of analysis is typically done with statistical tests that evaluate the list of DEGs interrogating pre-defined gene sets that represent pathways or Gene Ontology (GO) functions. Gene set enrichment (pathway) analysis is a complementary approach in the sense that not only DEGs are investigated but rather the entire gene expression response. This ensures that not only those pathways appear interesting that agglomerate many DEGs but also those that agglomerate subtle but consistent transcriptome changes of many of their members (not necessarily DEGs). A third, more unsupervised approach is network propagation, a mathematical concept that traces the effects of perturbations (e.g., gene expression changes) simultaneously across a molecular network according to a specified rule. It assumes that a perturbation in a certain gene is not only affecting that particular gene but rather its entire network neighborhood. The signal induced by a perturbation is then propagated to the neighbors and the neighbors of the neighbors until a steady-state (or convergence state) is achieved. The result of the network propagation is a final state in which each node is assigned a final weight which can be used to identify highly affected nodes as well as specific interconnected parts of the network (network modules) that are mostly affected by the induced perturbations.

A previous effort to infer functional effects of drug treatments from gene expression data was done on a pathway level in the work of Hardt et al. (2016). They assembled the ToxDb database, which contains gene expression data for more than 400 drugs and 2000 pathway concepts. This includes the association of drugs with specific molecular pathways, which can indicate to which mechanisms of action lead to toxicity. Here, we make use of this resource and the proposed implementation to extract pathways that are relevant for the toxic effect of different drugs. Additionally, we apply the same scoring scheme for measuring gene and pathway responses from gene expression data and enact a network analysis in order to identify functional modules that can also be associated with the toxic effect (see Methods).

Biological interactions are often described using molecular interaction networks (Barabasi and Oltvai, 2004), where each node represents a biological player, i.e., gene or protein, and each edge describes an interaction between a pair of nodes. Analyzing these networks can help to better elucidate the functional mechanisms that are being studied (McGillivray et al., 2018). There are numerous types of biological interaction networks (Vidal et al., 2011), as they can be based on different types of interactions and represent various biological actions. They vary between depicting gene regulatory interactions, viral-host interactions, metabolic reactions, protein–protein interactions, and more. Many of these interactions have already been made publically available through various specific databases, such as Reactome (Matthews et al., 2009), PID (Schaefer et al., 2009), KEGG (Kanehisa et al., 2012), and many others. Furthermore, there have been several attempts to combine and integrate different resources into one meta-resource, such as the work by Martha et al. (2011), IntNetDB (Xia et al., 2006), and ConsensusPathDB. In this work we make use of ConsensusPathDB (Kamburov et al., 2009), which currently integrates more than 600,000 interactions of different types which are collected from 32 public resources (Kamburov et al., 2013). Furthermore, we restrict our analysis to protein–protein interactions which are generally based on various experimental technologies (Walhout and Vidal, 2001).

One possible use of biological networks is for the identification of smaller subnetworks (subgraphs within the network), also referred to as modules, which depict an area that is more relevant for a specific biological function (Gustafsson et al., 2014). By integrating experimental data with interaction networks we can compute subnetworks that better represent the biological mechanisms which lead to a specific phenotype. There are several existing algorithms for module detection in biological networks. For a comprehensive overview of the different methods, see the recent review by Cowen et al. (2017). Many of these algorithms are based on a random walk process, where the weights of the nodes are propagated through the network, until a steady state is reached. The weighing of the nodes is dependent on the specific context and can be extracted for example from gene expression values or genetic mutation data. In this work we make use of the HotNet2 algorithm (Leiserson et al., 2015) that was originally developed for identifying subnetworks that result from somatic mutations. We apply the algorithm to toxicogenomics data and identify the most significant subnetworks for a drug treatment based on the gene expression response scoring.

We exemplify our approach on anthracycline drugs. Anthracyclines are a family of drugs that induce cardiotoxicity upon cancer treatment, and their use can result in cardiomyopathy and heart failure in many cases after a long period of time after treatment (Geisberg and Sawyer, 2010). These compounds are vastly used as chemotherapy agents, and have been shown to be extremely effective, but also to cause a major morbidity in cancer patients due to their toxic effects (Lenneman and Sawyer, 2016). Every exposure to anthracyclines carries some risk of resulting in cardiac dysfunction. The symptoms could present early on as well as at later times, in up to 23% of the patients (Steinherz et al., 1991). Although it is known that

anthracyclines disrupt the synthesis of DNA and RNA, mainly by inhibiting topoisomerase II (Geisberg and Sawyer, 2010) and that they lead to mitochondrial dysfunction (McGowan et al., 2017), the mechanisms that cause the cardiotoxic effects still remain largely unclear (Truong et al., 2014). Previous studies have tried to elucidate this problem, however, there is still need for further investigation so that detection and prevention could be improved (Raschi et al., 2010). We focus our analysis on the four most widely used compounds: daunorubicin, doxorubicin, epirubicin, and idarubicin. In addition, we compare the results to other chemotherapy agents from different drug families, which are also known to cause cardiotoxic effects.

MATERIALS

DrugMatrix

The toxicogenomics DrugMatrix (Ganter et al., 2005) database includes gene expression experiments from different rat tissue types at different time points and drug dosages. Data were downloaded via the diXa data collection (Hendrickx et al., 2015) that is available at http://wwwdev.ebi.ac.uk/fg/dixa/index.html. This data collection includes toxicogenomics profiles for 372 different compounds that were collected using the Affymetrix whole genome 230 2.0 rat GeneChip array. Data are available for heart, kidney, liver and muscle tissues, as well as for hepatocytes. The experiments were conducted for up to five times (after 0.25, 1, 3, 5 and 7 days) with only one dose concentration. In some cases, more than one dose was tested, and in others only one or two time points were measured.

Anthracycline Expression Data

The analysis was focused on four different anthracyclines compounds: daunorubicin (CHEMBL178), doxorubicin (CHEMBL53463), idarubicin (CHEMBL1117), and 4-epidoxorubicin (CHEMBL1237042). We downloaded the CEL files from the DrugMatrix database via the diXa data collection for these compounds in heart tissue. A full description of the treatments is given by **Table 1**. Daunorubicin is the only compound for which data are available at two doses, a higher "toxic" dose and a lower "pharmacological" dose (Ganter et al., 2005). Thus, for the analysis of daunorubicin we used only the higher dose.

Expression Data From Other Cardiotoxic Drugs

In order to evaluate the molecular effects that were identified in this work for the anthracyclines we also applied our workflow to three other chemotherapy agents that are known to induce cardiotoxicities (Truong et al., 2014). Out of 41 drugs that are mentioned in the review by Truong et al., only these three had data available in the DrugMatrix database. Cyclophosphamide (CHEMBL88) and ifosfamide (CHEMBL1024) are both alkylating agents, and imatinib (CHEMBL941) is from a family of small-molecule targeted therapy drugs. We downloaded the CEL files from the DrugMatrix database for these compounds in heart tissue. A full description of the treatments is given by **Table 2**. Data for imatinib were available in two different doses, and so we applied our analysis for the higher dose only.

ConsensusPathDB – A Molecular Interaction Network Resource

ConsensusPathDB (Kamburov et al., 2009) is a meta-database for molecular interactions and pathways that currently integrates 32 public resources (Kamburov et al., 2013) and is composed of more than 600,000 unique interactions of different types and holds more than 5,000 human pathway concepts. The database is available through a web server[1] where queries of genes, proteins, drugs and other types of biomolecules can be made, along with gene and metabolites analysis, such as enrichment and over-representation analysis (Herwig et al., 2016).

ConsensusPathDB holds an integrated network which is comprised of more than 300,000 binary protein–protein interactions (PPIs) representing a comprehensive model of the human interactome. These interactions were scored with a mixture of topology-based and annotation-based measures, such as the ones described in Goldberg and Roth (2003), Kuchaiev et al. (2009), Yu et al. (2010), and Kamburov et al. (2012a). These measures were aggregated into a meta-score using the IntScore (Kamburov et al., 2012b) approach, which combines the individual confidence scores, and provides a final score that better indicates how plausible the interaction is. The PPI network, along with the quality assessment scores, can be downloaded via http://cpdb.molgen.mpg.de/download/ConsensusPathDB_human_PPI.gz.

ToxDB

ToxDB (Hardt et al., 2016) integrates toxicogenomics data from two large-scale studies, Open TG-GATEs (Uehara et al., 2010) and DrugMatrix (Ganter et al., 2005), with pathway concepts from ConsensusPathDB (Kamburov et al., 2009). It contains a total of 7,464 different treatment data sets, covering 437 drugs, and 2,694 molecular pathway concepts with response scores. Its web interface is available at http://toxdb.molgen.mpg.de/and

[1] http://consensuspathdb.org

TABLE 1 | Anthracyclines drug treatment experiments from the DrugMatrix database.

Drug	Time points (days)	Dosage (mg/kg)
Daunorubicin	1, 3, and 5	2/3.25
Doxorubicin	1, 3, and 5	3
Idarubicin	1, 3, and 5	0.625
4-Epidoxorubicin	1, 3, and 5	2.7

TABLE 2 | Other chemotherapy drugs and their experiments information from the DrugMatrix database.

Drug	Time points (days)	Dosage (mg/kg)
Cyclophosphamide	3 and 5	25
Ifosfamide	3 and 5	143
Imatinib	1, 3, and 5	15/150

allows browsing for the effect of a drug treatment on cellular pathway response. The user can also browse for a specific pathway and retrieve the treatments that affect it the most.

METHODS

Microarray Data Processing

We processed the microarray data sets of the heart tissues that were treated with anthracyclines. The oligonucleotide sequences (oligoprobes) that were downloaded from DrugMatrix were mapped to the rat genome-build and probe sets were redefined using the resource at http://brainarray.mbni.med.umich.edu/CustomCDF such that each probe is assigned to a unique gene, and each gene is associated with a varying number of probes. It has been shown that re-mapping of oligoprobes unambiguously to the latest genome-build increases performance of Affymetrix Gene Chip transcriptomics platforms (Dai et al., 2005). The replicates of the different drug treatments were grouped together, according to their corresponding dosage and time point. The raw data were normalized using the GC Robust Multi-Array method in the R package *gcrma* (Gentry, 2017).

Orthology Mapping

Rat genes had to be mapped to human genes by orthology, in order to use the human pathway concepts and PPIs from ConsensusPathDB. This was done via the orthology mapping of the Ensembl BioMart repository (Yates and Akanni, 2016). We used only "One2one" and "One2many" homology relationships: if the rat gene has exactly one orthologous human gene, the corresponding rat microarray value is assigned to that human gene. Otherwise, if the rat gene has multiple orthologs in the human genome, the corresponding rat microarray value is assigned to all human paralogs.

Differential Gene Expression Analysis

The normalized microarray data were analyzed with the R package *limma* (Ritchie et al., 2015) in order to calculate differentially expressed genes (DEGs), i.e., genes that are up- or downregulated significantly when comparing compound treatment against control experiments. It estimates fold-changes and standard errors by fitting a linear model to each gene profile and uses an empirical Bayesian approach to smoothen these errors.

We applied *limma* for every pair of case-control normalized microarray values. Therefore, for every gene, given any drug, dosage and time point combination, we can calculate its fold change value and a corresponding *P*-value. Fold change is computed as the ratio of the mean expression values of treatment and control. *P*-value is the significance of the fold change given the null hypothesis that there is no change in expression between treatment and control.

Gene Scoring

In order to measure the response of a gene to a drug treatment experiment we use the following scoring scheme:

for every gene i, every drug j and every time-dosage treatment k:

$$S_{ijk} = \left|\log_2 r_{ijk}\right| \left|\log_{10} P_{ijk}\right| \quad (1)$$

Here r_{ijk} is the fold change between the treatment and the control experiments, and P_{ijk} is the *P*-value from the differential expression analysis. This score describes a weighted fold change of the gene, such that the more significant the change is, the higher the weight is. Using this scoring scheme allows us taking into consideration the rather low sample size of the experiments, as well as to avoid a pre-selection of the genes based on their *P*-values only. The score serves as a measure of how much the gene was affected by the treatment, regardless of the change in expression (higher or lower expressed in comparison to the control).

Pathway Scoring

In previous works (Yildirimman et al., 2011; Hardt et al., 2016) we have also defined a pathway scoring scheme, which is based on the scoring of the genes that the pathway is comprised of. Here, we take all available human pathways from ConsensusPathDB, and their associated genes. We compute for each pathway a relative pathway response (RPR) score which serves as a measure for the response of the pathway to the drug, given gene expression data. The higher the RPR score is, the more significant is the response of the pathway to the treatment. A pathway M_l is defined as a set of m genes: $M_l=\{g_1,...,g_m\}$. Given a treatment of drug j at a time point and dosage k, we can calculate the pathway score:

$$M_{l,j,k} = \frac{1}{m} \sum_{g_i \in M} S_{ijk} \quad (2)$$

Where S_{ijk} is the gene score of gene i, as defined in Equation 1. The RPR score of the pathway M_l with respect to the drug j and the time-dosage k is calculated by dividing the pathway score by $\overline{M}_{j,k}$ the median of all pathway scores, given drug j and time-dosage k:

$$RPR_{l,j,k} = \log_2 \left(\frac{\overline{M}_{l,j,k}}{\overline{M}_{j,k}} \right) \quad (3)$$

In addition, we computed RPR scores for all pathways in all the different experimental conditions and derived a background distribution. This background distribution is used to judge the significance of a given RPR score and reflects the response of the pathway to the experimental condition.

Network Module Analysis

A network module analysis was carried out by applying the HotNet2 (Leiserson et al., 2015) algorithm, which was originally developed to identify significantly mutated subnetworks in cancer in PPI networks based on somatic mutations data. The algorithm takes as input a score vector $S=(S_1,...,S_n)$, where n is the number of genes, and a graph $G=(V, E)$. The gene scores are computed context dependent (see below), and the graph represents a PPI network, where each node corresponds to a protein coding gene, and each edge to an interaction between

their respective proteins. HotNet2 then applies an insulated heat diffusion process that includes the following steps:

1. Heat diffusion – at each time step heat is diffused from every node *i* to every one of its neighbors *j*. The amount of heat that will be placed on node *j* given the initial heat on node *i* is given by the entry (*i,j*) of the diffusion matrix *F*, which is defined by:

$$F = \beta\,(\mathrm{I} - (1-\beta)\,W)^{-1} \tag{4}$$

$W_{\mathrm{i,j}} = \frac{1}{deg(j)}$ if *(i,j)* are neighbors, otherwise 0.

The parameter β is an insulating parameter and *W* is the normalized adjacency matrix of the input graph *G* such that deg(*j*) is the degree (number of neighbors) of node *j*. I is the identity matrix.

2. Exchanged heat − the amount of heat that diffuses from node *j* to node *i* when heat S_{j} is placed on node *j* is given by the exchanged heat matrix *E* which is defined by:

$$E = FD_{\mathrm{s}} \tag{5}$$

Where D_{s} is a diagonal matrix with the entries of *S*.

3. Identification of subnetworks − a new weighted directed graph *H* is created using the nodes *V*. Node *i* will be connected to node *j* in this graph if $E(i,j) > \delta$, where δ is a minimum edge weight parameter, and their respective edge will have a weight equal to *E*(*i,j*). Then, the strongly connected components of *H* are identified and are selected to be the final subnetworks.

4. Statistical test for the subnetworks – a two-stage statistical test, that is described in the original HotNet algorithm (Vandin et al., 2011, 2012), is applied to determine the significance of the number and the sizes of the subnetworks.

To identify functional modules that are associated with the different drug treatments we used the HotNet2 algorithm (Leiserson et al., 2015) that is available at http://compbio.cs.brown.edu/projects/hotnet2/. Since the first step of the algorithm depends only on the graph G and the chosen parameter β, we calculated the diffusion matrix F for the high-confidence ConsensusPathDB PPI network, while choosing $\beta = 0.5$. For the scoring of the genes, we used our own data-derived scores: for each drug and treatment, we used as input their gene scores, as described in Equation 1. The output of the HotNet2 algorithm depends largely on δ, the minimum edge weight parameter. The lower its value, the larger are the subnetworks. HotNet2 outputs four different results, for four different δ values, which are chosen based on a permutation test in their algorithm [for further details see (Leiserson et al., 2015)]. In this work we chose for further analysis the subnetworks which are resulted when taking the smallest δ parameter from the output of the HotNet2 algorithm.

Over-Representation Analysis (ORA)

ConsensusPathDB allows performing over-representation analysis (ORA) with different functionally relevant gene sets (Herwig et al., 2016). Given a set of genes, proteins or metabolites over-represented sets are searched among three pre-defined categories: (1) network neighborhood-based sets, (2) pathway-based sets, and (3) Gene Ontology (GO)-based sets. According to the hypergeometric test, a *P*-value is calculated based on the number of identifiers that are present in the given set and in the pre-defined sets. As background, the user can choose another set of identifiers, for example all genes that were measured in the experiment, or simply use all entities that are annotated in ConsensusPathDB. In our work, ORA was used to identify only the pathway based enriched sets, choosing all possible pathways from ConsensusPathDB and applying a *P*-value cutoff of 0.01. As background, we used the full list of genes that were measured in the corresponding experiment.

RESULTS

Workflow for Analyzing Toxicogenomics Data in the Context of Networks and Pathways

We established a computational workflow for analyzing toxicogenomics data by incorporating pathway and network information using different complementary approaches in order to gain functional information from gene expression data (**Figure 1**). We exemplify the results on the four anthracyclines drugs: daunorubicin (DAU), doxorubicin (DOX), idarubicin (IDA), 4-epidoxorubicin (EPI). We also applied our analysis to three other anti-cancer drugs that are known to cause cardiotoxicity: cyclophosphamide (CYC), ifosfamide (IFO), and imatinib (IMA). We compare our results for the anthracyclines with our results for these drugs in order to identify differences and commonalities and distinguish the effects that are explicit to anthracyclines. The workflow is based on the results of a differential expression analysis, and combines pathway and network information from both ConsensusPathDB and ToxDB. It begins with an over-representation analysis for the DEGs, using pathway concepts that are collected in ConsensusPathDB, in order to assign a biological function to the most significantly changed genes. Next, it continues with a pathway analysis using ToxDB, extrapolating from DEGs to the entire gene expression response and from gene lists to pathway concepts. Using molecular interaction information from ConsensusPathDB, the workflow also includes a PPI network construction and an analysis that applies a network propagation algorithm which combines the DEGs with the PPI network. Finally, it is able to identify subnetworks that we define as drug toxicity modules.

Assigning Biological Function to Gene Lists With Over-Representation Analysis

A first step in functional interpretation of toxicogenomics results is to interrogate the lists of DEGs (see Methods) for known annotation sets such as pathways or GO terms using Fisher's test or similar statistics (see Methods). Summarizing the different experiments (time points and dosages) for the four anthracyclines (DOX, DAU, EPI and IDA) results in 1,883 DEGs for EPI, 1,555 for DOX and 1,062 for DAU whereas IDA shows a much weaker response with 388 genes (**Figure 2A**). In all cases, human genes were inferred based on homology mapping of the corresponding rat microarray probes. All anthracyclines were administered at maximum tolerated doses (MTDs) and, thus should be of comparable toxicity (DOX 3 mg/kg; DAU 3.25 mg/kg; EPI

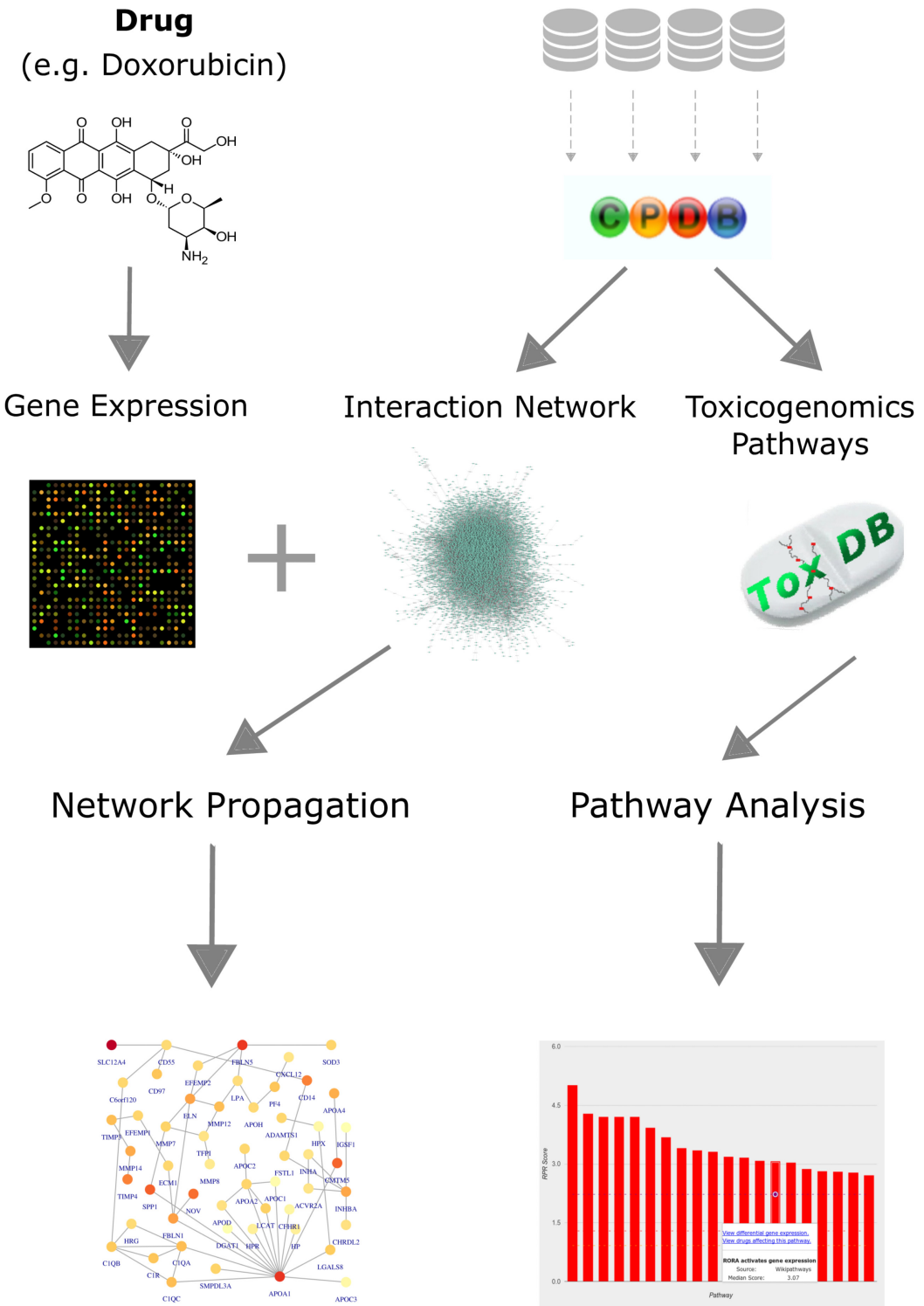


FIGURE 1 | Workflow for analyzing toxicogenomics data at the network and pathway level. Gene expression data, following a drug perturbation, is collected and analyzed to identify differentially expressed genes. Pathway concepts and interaction network are extracted from ConsensusPathDB. Identification of relevant pathways is done via ToxDB, while functional modules are detected by applying a network propagation algorithm that combines both the gene expression data and the PPI network. The workflow allows us to identify: (i) differentially expressed genes that could be candidates for further experiments, (ii) relevant pathways that are disrupted in response to the treatment, and (iii) network toxicity modules that hold functional information about the mechanisms of action.

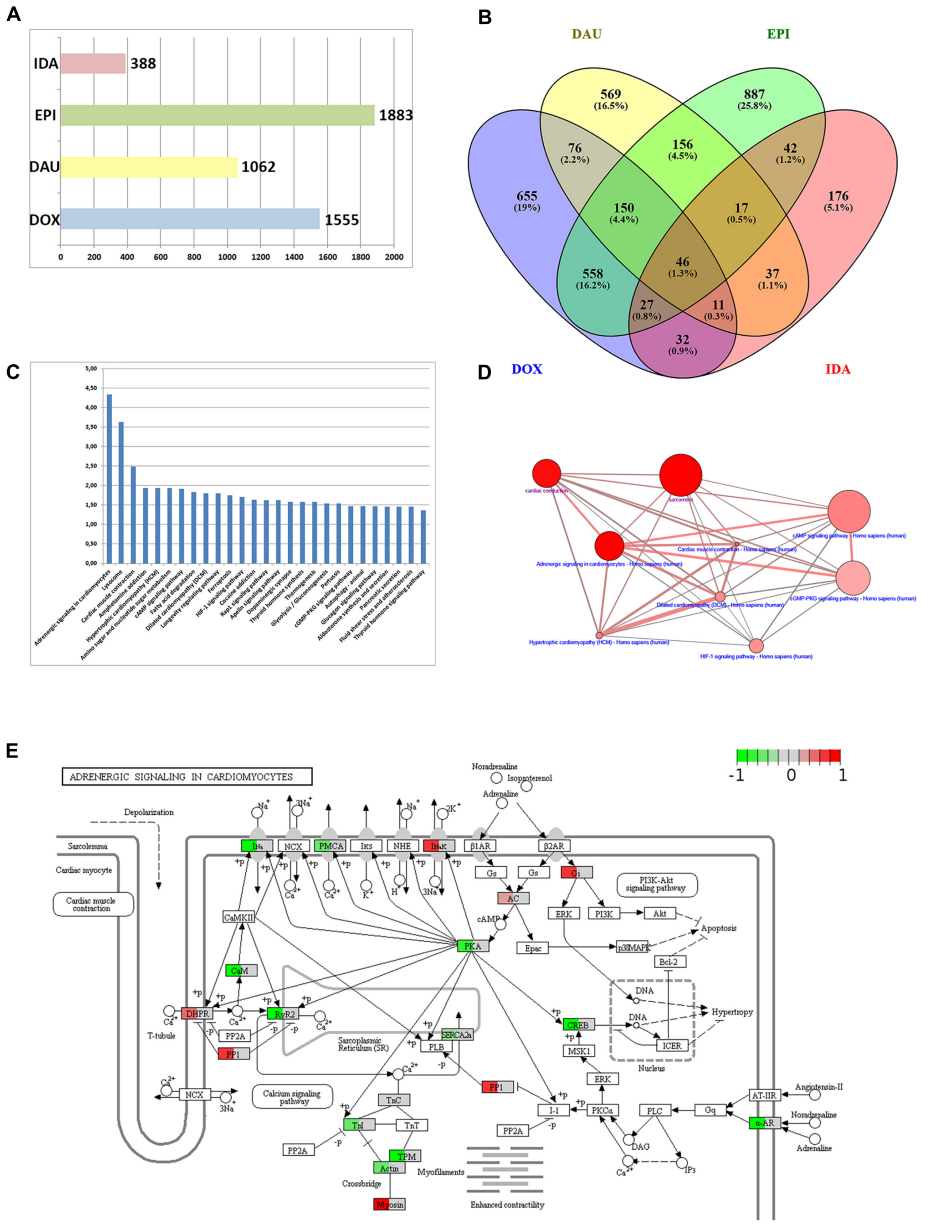


FIGURE 2 | Anthracycline over-representation analysis. **(A)** Summary of the number of DEGs from the different experimental conditions: DOX (3 mg/kg at 1, 3, and 5 days), DAU (3.25 and 2 mg/kg at 1, 3, and 5 days), EPI (2.7 mg/kg at 1, 3, and 5 days), and IDA (0.625 mg/kg at 1, 3, and 5 days). **(B)** VENN diagram of DEGs with respect to the four compound treatments. **(C)** 27 KEGG pathways that were found significantly over-represented with respect to the 1555 DEGs after DOX treatment using the Fisher test statistic with the ConsensusPathDB. *Y*-axis = –log10(*P*-value). **(D)** Interdependency of significant pathways from **(C)** (blue label) and GO categories (magenta label) computed with ConsensusPathDB. Size of balls indicates pathway size, shade of balls indicate overlap with DEG list. **(E)** "Adrenergic signaling in cardiomyocytes" pathway found significantly over-represented (P = 3.49E−07) and expression data of 32 DEGs overlaid with the pathway. Mapping of gene expression fold-changes to pathway has been done with Pathview (Luo et al., 2017).

2.7 mg/kg and IDA 0.625 mg/kg). Also, it has been shown that gene expression signatures are predictive of toxicity and that number of DEGs is indicative of phenotypically observed injury of the organ (Paules, 2003; Andersen et al., 2008; Holmgren et al., 2015) which has given rise to the concept of phenotypic anchoring, i.e., the association of gene expression signatures to toxic phenotypes. The difference in DEGs between DOX and IDA is in line with previous findings: For example, Platel et al. (1999) showed that in rat the MTDs for DOX and IDA were 3 mg/kg and 0.75 mg/kg, i.e., comparable to the levels used in the DrugMatrix screen, and that at these MTDs IDA showed significantly lower cardiotoxicity than DOX. Anthracyclines show highly specific response at the gene expression level (**Figure 2B**) with 40–50% of all DEGs specific for a certain compound. The strongest relative

agreement in gene expression response was observed between DOX and EPI (45%), whereas the relative agreement between IDA and the other three compounds is much lower (12–15%). DOX response can be characterized at the pathway level using ORA analysis (see Methods). **Figure 2C** exemplifies the results for DOX using gene sets that represent KEGG (Kanehisa et al., 2012) pathways and that might reflect the cell's response to the drug treatment. In total, 27 KEGG pathways are significantly over-represented (Q-value < 0.05 and at least 10 DEGs overlapping with the pathway gene set). A number of disease gene sets have been identified such as "Hypertrophic cardiomyopathy" (Q = 0.0115), "Dilated cardiomyopathy"(Q = 0.0158) or "Cardiac muscle contraction" (Q = 0.0032). Interestingly, these pathways were also found in a recent study investigating cardio-toxicity in human pluripotent stem cell derived-cardiomyocytes (Maillet et al., 2016) and thus seem to extrapolate from rat *in vivo* to human *in vitro* studies. The top-enriched pathway in our setting is "Adrenergic signaling in cardiomyocytes" (*Q* = 4.54E-05). 32 genes of that pathway are differentially expressed including troponins (*TNNC1* and *TNNI3*), tropomyosins (*TPM1* and *TPM2*), and other well-known toxicity-associated genes such as *RYR2* (ryanodine receptor 2). **Figure 2D** displays the interdependencies of these and other disease-related gene sets.

An important feature in the analysis is the visualization of the gene expression changes in the pathway map. Pathway maps can be retrieved by pathway resources such as KEGG. There are many tools that allow visualizing gene expression fold-changes on these pathways which is exemplified in **Figure 2E** with the "Adrenergic signaling in cardiomyocytes" and the expression fold changes of the DOX treatment.

From Genes to Pathways – Pathway Analysis Using the ToxDB

The next level of analysis is to extrapolate the gene expression values from single genes to entire pathways. We have built a tool, ToxDB that combines gene expression data and pathway concepts. ToxDB builds on three components: (i) a comprehensive collection of pathway concepts along with drug treatment microarray data, (ii) a numerical method to compute pathway responses from genome-scale expression data, and (iii) a web interface that allows user interaction. By this procedure each pathway is assigned a numerical value that reflects its response to the treatment (see Methods). ToxDB contains pre-calculated pathway scores for ca. 2,700 different pathways and ca. 7,500 experimental conditions mainly extracted from two large toxicogenomics studies, TG-GATES and DrugMatrix (see Methods). A background distribution of pathway scores is used to infer statistical significance. ToxDB can be used in different views. The drug view allows drug centric analysis: by selecting a compound, for example DOX, and a specific experiment all responding pathways can be viewed (**Figure 3A**). By further clicking on a specific response pathway [here "Hypertrophic cardiomyopathy (HCM)"] the expression results of all genes can be inspected that are associated with this pathway (**Figure 3B**). DEGs of this pathway are known cardiac-relevant genes such as *MYH7* (myosin, heavy chain 7, cardiac muscle, beta; log2−FC = 3.65, *P* = 9.05E−06), *DES* (desmin; log2-FC = 0.264, P = 7.55E−02), *TPM4* (tropomyosin 4; log2-FC = -1.07, P = 9.14E−04), or *RYR2* (ryanodine receptor 2; log2-FC = −1.98, *P* = 5.43E−07). A second view is the pathway view: the user can select a single pathway (here "Cardiac muscle contraction") and as a result all experiments are shown in which this pathway responded significantly (**Figure 3C**). Pathways can be selected from ten different resources which comprise most widely used pathway resources such as KEGG, Reactome

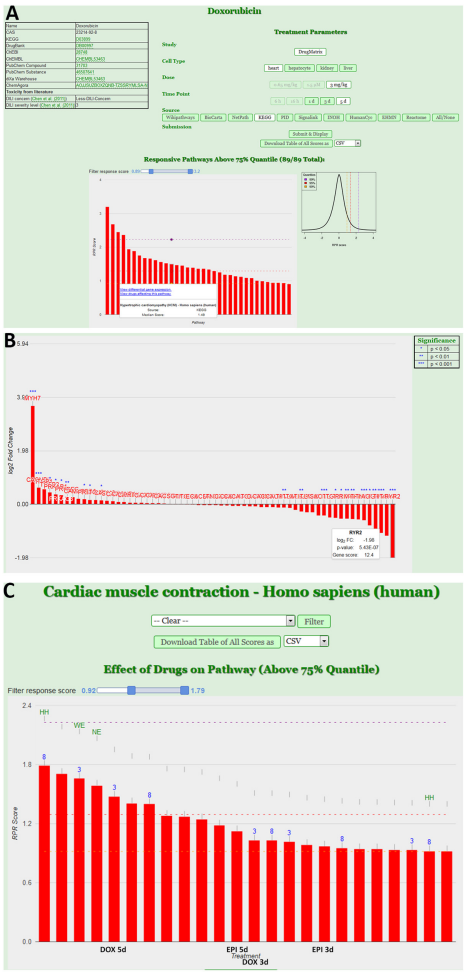


FIGURE 3 | Pathway analysis using the ToxDB. **(A)** Drug view contains links to chemical information (top left), specification of the experimental data (top right) and the display of the response pathways for that experiment in form of an interactive bar plot. Bars indicate the strength of the response, with mouse over the user can display further information of the pathway. The background distribution of the response scores is displayed as density plot next to the plot. **(B)** Gene view. Once a pathway is selected, user can inspect the experimental results of the genes. In this case the log2 fold-changes of the genes associated with the pathway "Hypertrophic cardiomyopathy (HCM)" from the KEGG database are shown. Stars indicate differentially expressed genes (DEGs). **(C)** Pathway view. Users can infer specific pathways (here "Cardiac muscle contraction"). The interactive bar plots represents the response of that pathway in different experiments (e.g., DOX and EPI at 3 and 5 days experiments).

or BioCarta. It can be seen from the view that anthracycline experiments (DOX and EPI at different time points) are among the compounds that induce the most significant responses of cardiac muscle contraction.

Protein–Protein Interaction Network Construction

Protein–protein interaction networks are typically used as scaffolds for drawing network propagation of gene expression data on. The underlying argument is "guilt-by-association," i.e., the assumption that genes/proteins that interact with each other usually share function and, thus, that network modules computed from these PPI networks amplify functional information. Thus, the PPI network needs to be properly selected in the sense that it should be sufficiently comprehensive and that the false-positive rate of interactions should be low.

In this work we make use of the PPI network from ConsensusPathDB (release 32) and reduce it to a high-confidence network by taking only the interactions with a confidence score of 0.95 or higher (**Figure 4C**). This network is comprised of 10,707 proteins and 114,516 unique interactions (**Figure 4A**). Biological networks are normally characterized with a power law distribution of the node degree (Barabasi and Oltvai, 2004). This means that most of the nodes in the network are only connected to a few other nodes, while a small majority is very highly connected, with more than 400 neighbors (**Figure 4B**). We make use of this high-confidence interaction network in our workflow in order to identify subnetworks that are highly relevant to the drug treatments.

Toxicity Network Modules Are Identified by Applying Network Propagation

Toxicity modules were calculated using the HotNet2 algorithm for each drug and time point independently. A detailed list of all the toxicity modules is provided in **Supplementary Table 1**. Here we discuss the results when using the gene expression values for DOX only. The modules for the other anthracyclines are provided in **Supplementary Figures 1–3**. Since the drug treatments of DOX were measured three times over the course of 5 days, we derived one module for each one of the time points (**Figure 5A**). Looking at each one of the modules, and at all of them together, allows us to analyze the changes over time. We identified that the effect becomes stronger after 3 days, as the size of the module grows, but also that it is again much lower after 5 days. This could be due to the toxic effect of the drugs on the cells, i.e., the cells might already be dying. We confirmed this by looking at the over-represented pathways for the genes in the "5 days" module (**Figure 5B**). We observed two pathways that indicate cell death: "Apoptosis" and "Apoptotic Signaling Pathway." In addition, we identified another pathway that might be involved in cardiotoxicity: "Cardiac Progenitor Differentiation" (from the WikiPathways database). This pathway includes several factors that are involved in cardiac differentiation, such as *TNNI3* that we also detected as differentially expressed, and is based on two recent reviews (Burridge et al., 2012; Stillitano et al., 2012). The module also includes the genes *IGF1* and *IGF2* that are involved in the differentiation of immature cardiomyocytes and have been associated with cardiac hypertrophy (Wang et al., 2012). Other genes that might be involved in cardiotoxicity and are present in the "1 day" and "3 days" modules are *APOA1*, that have been previously associated with hereditary amyloid cardiomyopathy (Hamidi Asl et al., 1999), and *ELN* that has been involved in both progressive aortic valve malformation and latent valve disease in mice (Hinton et al., 2010).

Network Modules Amplify Functional Information

We compared the over-represented pathways when using only the high scoring genes (genes with a score above the 99th quantile of the background distribution of all scores), and when using the genes from the network modules (**Figure 6**). In 8 out of 12 drug-treatment conditions the enrichment scores when using the genes from the network modules, were higher than the scores when using the high scoring genes only. Furthermore, when comparing the significance of the enrichment, by looking at the means of the Q-values (FDR corrected *P*-values), in all but one case we observed a higher enrichment when using the genes from the modules. This suggests that the network modules are enriched in more functional information, and therefore they serve as a powerful mean for studying systemic processes, such as drug induced toxicity.

Differences and Commonalities Between Anthracyclines and Other Chemotherapeutic Drugs

To assess the specificity of our results for anthracycline-induced cardiotoxicity, we applied the same workflow (**Figure 1**) to three other anti-cancer drugs which are known to cause cardiotoxic phenotypes: cyclophosphamide, ifosfamide and imatinib (see Materials). Looking only at the high scoring genes (genes with a score above the 99th quantile), with gene scores computed according to section 3.4, we observe hardly any common genes between the three drugs (**Figure 7A**). Interestingly the 17 genes that are common between these three drugs are also common with all the other anthracyclines drugs. However, when we compare the genes that are present in the toxicity modules of these drugs and the toxicity modules of all anthracyclines (**Figure 7B**) we detect 214 common genes. This highlights the fact that the network propagation approach amplifies gene expression responses toward relevant cardiotoxic mechanisms and phenotypes that are shared by the different drugs so that different gene expression responses can result in similar pathway responses. Evidently the number of genes in the anthracyclines modules is much higher as they are derived from more drugs and experiments, but nonetheless the percentage of number of genes that are shared is much higher. This could again indicate to the functional information that is inherent within the toxicity modules, which might suggest to the mechanisms that are involved in causing the toxic effect.

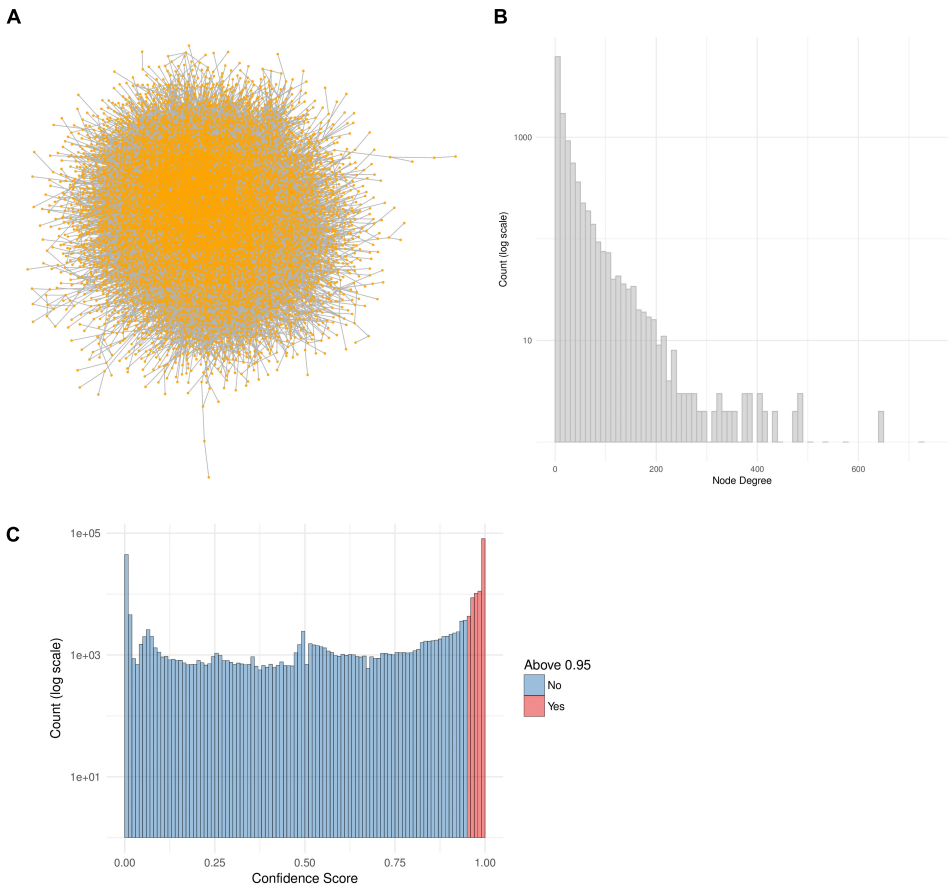


FIGURE 4 | PPI Network construction from ConsensusPathDB. **(A)** High-confidence PPI network extracted from the ConsensusPathDB database with 10,707 nodes and 114,516 undirected edges. **(B)** Node degree distribution of the PPI network. **(C)** Distribution of all IntScore confidence scores from all ConsensusPathDB unique interactions. In blue are the confidence scores below 0.95 and in red are those above it. The high-confidence network includes only the red interactions.

In order to identify biological functions that are specific for anthracyclines we performed enrichment analysis with the set of 330 genes that are solely part of the anthracycline toxicity modules (**Figure 7C**). We observe enrichment of cardiac disease pathways such as "Viral myocarditis," "Hypertrophic cardiomyopathy," and "Dilated cardiomyopathy," mainly through the inclusion of ITGB and TGFB gene family members and RYR2. Another strong signal is the presence of immune response pathways. It is well-known that anthracycline treatment can induce systemic inflammation mediated through interleukins (Mills et al., 2008; Sauter et al., 2011). Interestingly, many inflammatory and immune response pathways are enriched with the anthracycline toxicity modules, in particular through interleukins (IL1A, IL12A, IL12B, IL23A, IL33, and IL27RA) that are not included in the modules of the other drugs.

DISCUSSION

Combining the information from ConsensusPathDB and ToxDB, including pathway concepts and a PPI network, together with experimental data, allows for a more comprehensive view of the effects of the drug treatments. Firstly, by using ToxDB we are able to identify pathway concepts and by that suggest specific mechanisms that may be either the cause or the consequences of the toxic effects. In addition, by using the information from the PPI network and a propagation algorithm, we can also identify specific interactions that could be highly relevant for further experiments. These network modules carry out more functional information, since their genes and interactions represent parts of different pathways, and thus they are enriched in more information about specific biological mechanisms. Indeed, by propagating perturbation data across a network it is possible to gain information not only for the genes that were actually measured by the experiment but in addition also for the genes that haven't been measured experimentally but that are connected with many measured neighbors in the network.

When looking only at DEGs, it is very difficult to describe the toxic effects of a drug given a specific treatment. Usually, this list of genes is comprised of hundreds of possible candidates, and it can be very challenging to distinguish which ones are involved in causing toxicity. Other works have tried to reduce the number of genes by looking at a smaller toxicogenomics

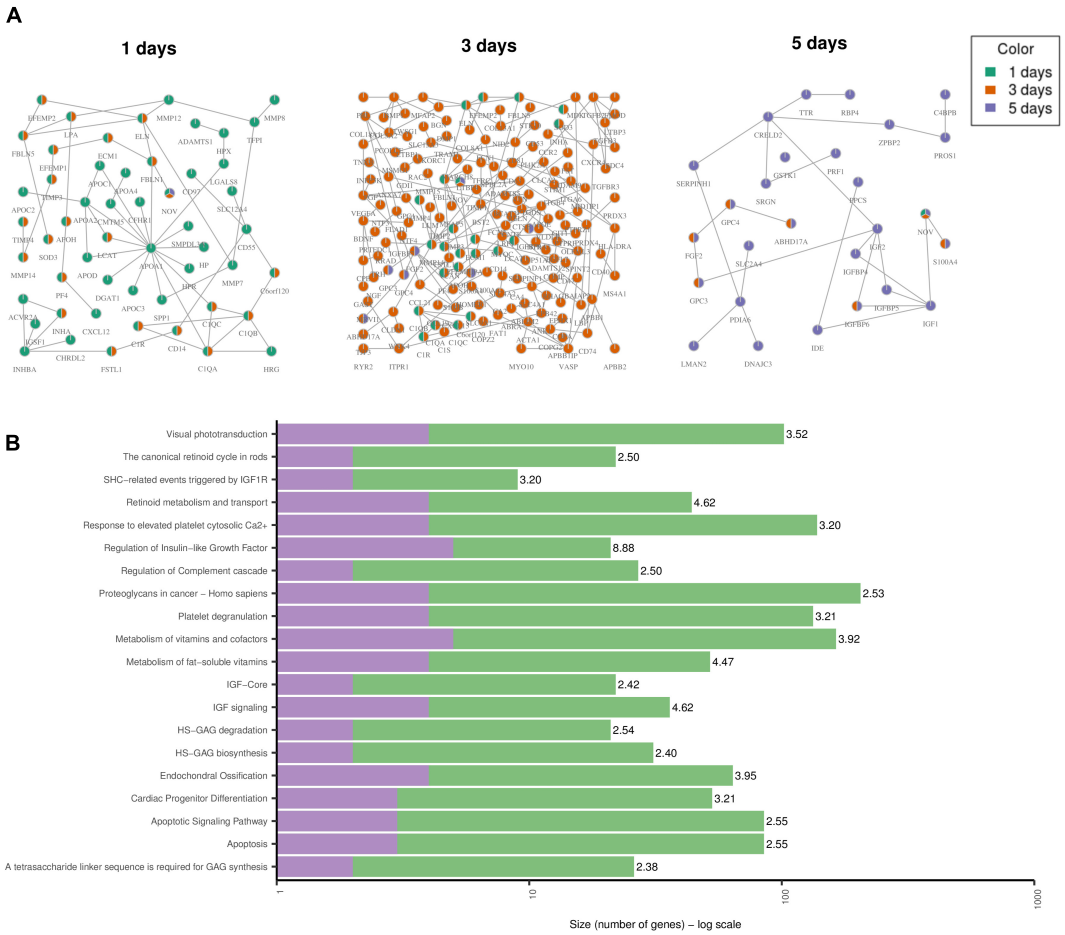


FIGURE 5 | Toxicity Modules. **(A)** Toxicity modules were identified using the HotNet2 propagation algorithm for DOX drug treatments after 1, 3, and 5 days. Each node corresponds to a protein coding gene (the nodes are named using their HGNC symbol) and each edge is an interaction as defined by the PPI network of ConsensusPathDB (see Methods). The colors of the nodes indicate the time point. Nodes that are colored in green only are present in the "1 day" module only. In the same way for orange and purple. Nodes that are colored in two colors are present in the two corresponding modules. Nodes that are colored in three colors are present in all modules. **(B)** The top 20 over-represented pathways for the "5 days" module, based on the ORA of the genes in the module with ConsensusPathDB (see Methods). The purple color represents the overlap of genes from the pathway and the module. The green are the rest of the genes from the pathway (that are not in the module). The number next to each bar displays the significance of the over-representation [–log10(*Q*-value) of the corresponding pathway].

space (Kohonen and Parkkinen, 2017). By defining a more complex gene score, we were able to reduce the number of genes such that it becomes easier to extract plausible candidates for further studies. Furthermore, by applying a network propagation scheme to the gene scores and the high-confidence PPI network, we were able to both reduce the list even further, and also identify functional modules within PPI networks. These functional modules can better reflect the mechanisms that lead to toxicity, as they contain not only the obvious candidate genes based on the differential expression analysis, but also other genes that might be associated with the toxic effect, and are also connected to the more significantly changed genes.

ConsensusPathDB is a meta-database that agglomerates information from multiple resources and therefore includes different kinds of interactions: protein–protein, genetic, metabolic, signaling, gene regulatory and drug-target. In addition, it also holds information about biochemical molecules and pathways. The high confidence PPI network that we have constructed is comprised solely of highly scored protein–protein interactions that are extracted from several resources, such as BIND, INTACT, HPRD. However, the ORA that is provided within ConsensusPathDB, searches for over representation of genes within cellular pathways. These pathways are derived from other resources, such as KEGG, Reactome, WikiPathways, etc. The different resources are completely independent data sets and the ConsensusPathDB simply serves as a common analysis platform. Therefore, when we apply the ORA to the extracted network modules, we can identify how enriched they are not only with protein–protein interactions, but also with pathway information. As we have illustrated in **Figure 6**, network modules contain not only protein–protein interaction information (that is inherent within its structure) but also are enriched in

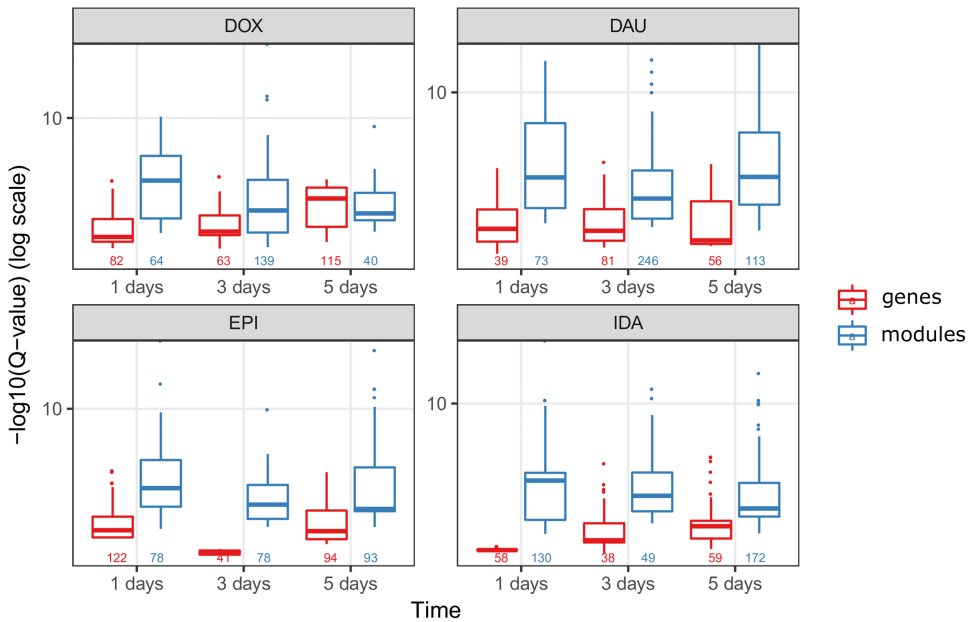


FIGURE 6 | Network modules amplify functional information. We compared the scores of the over-represented pathways when using the highly scoring genes (score >99th quantile) (in red) and when using the genes from the HotNet2 modules (in blue), for all four anthracyclines and in the three time points of the experiments. The scores of the pathways are the –log10(*Q*-value) of the *Q*-values from the ORA that was done via ConsensusPathDB (*Q*-values are the FDR corrected *P*-values from the hypergeometric test). Below the boxplots, the numbers indicate the number of the significantly (*P*-value < 0.01) over-represented pathways, for each one of the conditions.

other functional information that is represented in various pathways.

It should be noted that besides the described publicly available tools for pathway annotation and analysis, there are commercially available tools that hold functionality for pathway and network analysis such as IPA (Ingenuity/Qiagen), TransPath (geneXplain), or MetaCore (Thomson Reuters). These and other commercial and publicly available tools can be used to construct suitable molecular networks and perform enrichment analysis and module computation. A survey of databases and resources is given by Pathguide (Bader et al., 2006), a recent review and comparison of pathway tools has been published for example for metabolomics data (Marco-Ramell et al., 2018).

Toxicology studies often explore the effects of compounds over time and varying dosages (Hartung, 2009). Here, we analyzed gene expression levels for three different time points: after 1, 3, and 5 days. Every time point experiment was independently compared to the control experiment, such that a network module was constructed for every time point. To discern the effect over time, we compared between the modules, and determined the possible changes due to time. We were able to identify the toxic effect over time by looking at the different modules and also the genes within the module that could implicate the pathways that are leading to toxicity. In the future, one could try to first integrate the experimental data from the different time points, such that the change in expression levels over time is taken under consideration. For example by applying a mathematical model to detect differential expression over time, like the one suggested by Conesa et al. (2006). We could further use the results of such model and incorporate them into the network propagation algorithm in order to identify a module that encompasses data from all the time points together.

The approach we applied in this work consists of three main components: gene expression analysis, a PPI network and a network propagation algorithm. All of these have several alternatives, and could be further incorporated in future analysis. Firstly, the PPI network can be replaced with other genetic interaction networks, for example a gene regulatory network that is derived from experimental data (Zheng and Huang, 2018). Secondly, different types of experimental data can be used for ranking the genes and using their ranks as scores for the chosen propagation algorithm. Gene expression values from RNA-seq experiments could easily be investigated in the same manner, along with protein abundance data, mutation data or epigenetic data. Finally, there already exist different approaches for applying propagation algorithms to detect network modules. Here we have chosen to use the HotNet2 algorithm, but several others, like the ones in the review by Cowen et al. (2017), might also be considered.

In our work we focus on anthracyclines, a group of commonly used chemotherapy drugs. We used the data that are available in the DrugMatrix (Ganter et al., 2005) database and applied our workflow (**Figure 1**). This workflow could easily be applied to other data resources as well as other groups of drugs. Some previous works have already been developed to analyze toxicogenomics data from

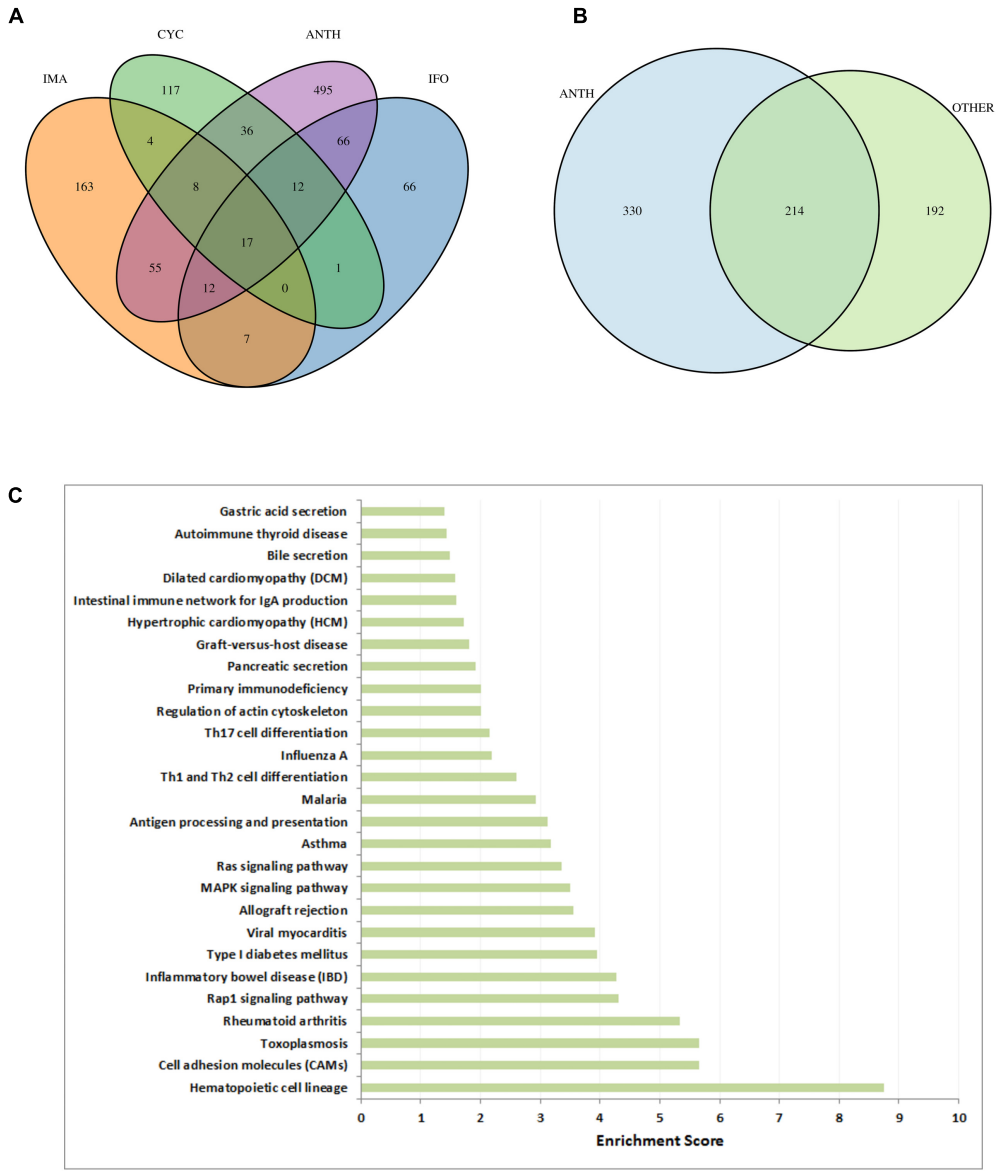


FIGURE 7 | Cardiotoxic effects of anthracyclines in comparison to other drugs. **(A)** VENN diagram of high-scoring genes (genes with a score above the 99th percentile) with respect to the three other compounds (CYC, IFO, and IMA) and all anthracyclines together (ANTH). **(B)** VENN diagram of genes in the toxicity modules, with respect to all anthracyclines (ANTH) together and all other drugs together (OTHER). **(C)** Significantly enriched KEGG pathways ($P < 0.01$) with the 330 genes that are contained in the computed anthracycline toxicity modules and not contained in the toxicity modules of the other cardiotoxic drugs. Bars indicate an enrichment score computed as –log10(*Q*-value), where *Q*-value is the FRD-corrected *P*-value of the enrichment.

DrugMatrix and were applied for identifying different types of drug induced toxicities. For example, Tawa et al. (2014) characterized liver induced drug toxicity by identifying gene co-expression modules that are associated with a toxic response. They defined these gene modules using six different methods, including Pearson correlations and PPI information. A similar approach was also applied to identify gene co-expression modules for kidney induced drug toxicity (AbdulHameed et al., 2016). In another work, AbdulHameed et al. (2014) also tried to identify liver induced drug toxicity by integrating toxicogenomics data with pathway and PPI network information. They performed a differential expression analysis and identified relevant gene modules by applying the KeyPathwayMiner (Alcaraz et al., 2012) algorithm. Other network based approaches have also been suggested for the analysis of toxicogenomics data from the DrugMatrix database. For instance, Sutherland et al. (2017) have constructed gene co-expression networks using WGCNA (Zhang and Horvath, 2005) and associated modules with different drug toxicity phenotypes. Mulas et al. (2017) compiled a pipeline for network comparison and used it to identify drugs with similar

toxicity profiles. In our workflow, we chose to apply a network propagation algorithm that is based on a random walk model. We showed that this approach allows for the identification of drug toxicity modules that are highly enriched in functional information and provide new insights into the toxic causing mechanisms.

Gene expression signatures have been associated with toxicity phenotypes with the concept of phenotypic anchoring (Paules, 2003). Here, the idea is that specific signatures emerge over time and dose that can be related to distinguishable phenotypes. We have observed that, for example the number of DEGs in DOX and IDA at MTDs reflect previously observed differences in the toxicity of both compounds. Additionally, when comparing enrichment scores in heart-related diseases pathways, DOX appears as the most toxic compound followed by EPI, while IDA and DAU show basically no enrichment in these pathways (**Supplementary Figure 4**).

Associating genotype with phenotype, and specifically predicting a toxic phenotype that rises due to drug treatment, still remains an intricate challenge. Integrating experimental data with prior knowledge in the form of biological networks, as suggested in our work, is a suitable step when trying to describe the molecular effects of drug treatments. However, there is still much to be improved. The PPI networks still hold a high bias in interactions due to annotation (Schramm et al., 2013; Luecken et al., 2018) and will keep getting refined as our understanding of the biological systems increases. Better experimental techniques become more and more available, and data from those will need to be integrated for an even more comprehensive analysis (Hasin et al., 2017; Yan et al., 2017; Karczewski and Snyder, 2018). And finally, better computational approaches for differentiating between cases and controls, as well as for analyzing big networks such as PPIs, are still to be developed.

AUTHOR CONTRIBUTIONS

GB developed the workflow, performed parts of the data analysis, and wrote the manuscript. RH conceived the study, performed parts of the data analysis, and wrote the manuscript.

FUNDING

The study was in part funded by the European Commission under its 7th Framework Program (Grant HeCaToS, 602156).

ACKNOWLEDGMENTS

We thank Christopher Hardt for assistance in data analysis and Atanas Kamburov for work with ConsensusPathDB.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00484/full#supplementary-material

REFERENCES

AbdulHameed, M. D. M., Ippolito, D. L., Stallings, J. D., and Wallqvist, A. (2016). Mining kidney toxicogenomic data by using gene co-expression modules. *BMC Genomics* 17:790. doi: 10.1186/s12864-016-3143-y

AbdulHameed, M. D. M., Tawa, G. J., Kumar, K., Ippolito, D. L., Lewis, J. A., Stallings, J. D., et al. (2014). Systems level analysis and identification of pathways and networks associated with liver fibrosis. *PLoS One* 9:e112193. doi: 10.1371/journal.pone.0112193

Alcaraz, N., Friedrich, T., Kötzing, T., Krohmer, A., Müller, J., Pauling, J., et al. (2012). Efficient key pathway mining: combining networks and OMICS data. *Integr. Biol.* 4, 756–764. doi: 10.1039/c2ib00133k

Andersen, M. E., Clewell, H. J. III, Bermudez, E., Willson, G. A., and Thomas, R. S. (2008). Genomic signatures and dose-dependent transitions in nasal epithelial responses to inhaled formaldehyde in the rat. *Toxicol. Sci.* 105, 368–383. doi: 10.1093/toxsci/kfn097

Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Res.* 34, D504–D506. doi: 10.1093/nar/gkj126

Barabasi, A. L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Burridge, P. W., Keller, G., Gold, J. D., and Wu, J. C. (2012). Production of de novo cardiomyocytes: human pluripotent stem cell differentiation and direct reprogramming. *Cell Stem Cell* 10, 16–28. doi: 10.1016/j.stem.2011.12.013

Conesa, A., Nueda, M. J., Ferrer, A., and Talon, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* 22, 1096–1102. doi: 10.1093/bioinformatics/btl056

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* 18, 551–562. doi: 10.1038/nrg.2017.38

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* 33:e175. doi: 10.1093/nar/gni179

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

Geisberg, C. A., and Sawyer, D. B. (2010). Mechanisms of anthracycline cardiotoxicity and strategies to decrease cardiac damage. *Curr. Hypertens. Rep.* 12, 404–410. doi: 10.1007/s11906-010-0146-y

Gentry, J. (2017). *"gcrma: Background Adjustment Using Sequence Information".)*.

Goldberg, D. S., and Roth, F. P. (2003). Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. U.S.A.* 100, 4372–4376. doi: 10.1073/pnas.0735871100

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

Gustafsson, M., Nestor, C. E., Zhang, H., Barabasi, A. L., Baranzini, S., Brunak, S., et al. (2014). Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome Med.* 6:82. doi: 10.1186/s13073-014-0082-6

Hamidi Asl, L., Liepnieks, J. J., Hamidi Asl, K., Uemichi, T., Moulin, G., Desjoyaux, E., et al. (1999). Hereditary amyloid cardiomyopathy caused by a variant apolipoprotein A1. *Am. J. Pathol.* 154, 221–227. doi: 10.1016/S0002-9440(10)65268-6

Hardt, C., Beber, M. E., Rasche, A., Kamburov, A., Hebels, D. G., Kleinjans, J. C., et al. (2016). ToxDB: pathway-level interpretation of drug-treatment data. *Database* 2016:baw052. doi: 10.1093/database/baw052

Hartung, T. (2009). Toxicology for the twenty-first century. *Nature* 460, 208–212. doi: 10.1038/460208a

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18:83. doi: 10.1186/s13059-017-1215-1

Hendrickx, D. M., Aerts, H. J., Caiment, F., Clark, D., Ebbels, T. M., Evelo, C. T., et al. (2015). diXa: a data infrastructure for chemical safety assessment. *Bioinformatics* 31, 1505–1507. doi: 10.1093/bioinformatics/btu827

Herwig, R., Hardt, C., Lienhard, M., and Kamburov, A. (2016). Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat. Protoc.* 11, 1889–1907. doi: 10.1038/nprot.2016.117

Hinton, R. B., Adelman-Brown, J., Witt, S., Krishnamurthy, V. K., Osinska, H., Sakthivel, B., et al. (2010). Elastin haploinsufficiency results in progressive aortic valve malformation and latent valve disease in a mouse model. *Circ. Res.* 107, 549–557. doi: 10.1161/CIRCRESAHA.110.221358

Holmgren, G., Synnergren, J., Bogestal, Y., Ameen, C., Akesson, K., Holmgren, S., et al. (2015). Identification of novel biomarkers for doxorubicin-induced toxicity in human cardiomyocytes derived from pluripotent stem cells. *Toxicology* 328, 102–111. doi: 10.1016/j.tox.2014.12.018

Kamburov, A., Grossmann, A., Herwig, R., and Stelzl, U. (2012a). Cluster-based assessment of protein-protein interaction confidence. *BMC Bioinformatics* 13:262. doi: 10.1186/1471-2105-13-262

Kamburov, A., Stelzl, U., and Herwig, R. (2012b). IntScore: a web tool for confidence scoring of biological interactions. *Nucleic Acids Res.* 40, W140–W146. doi: 10.1093/nar/gks492

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* 41, D793–D800. doi: 10.1093/nar/gks1055

Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res.* 37, D623–D628. doi: 10.1093/nar/gkn698

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

Karczewski, K. J., and Snyder, M. P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* 19, 299–310. doi: 10.1038/nrg.2018.4

Kohonen, P., and Parkkinen, J. A. (2017). A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury. *Nat. Commun.* 8:15932. doi: 10.1038/ncomms15932

Kuchaiev, O., Rasajski, M., Higham, D. J., and Przulj, N. (2009). Geometric de-noising of protein-protein interaction networks. *PLoS Comput. Biol.* 5:e1000454. doi: 10.1371/journal.pcbi.1000454

Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168

Lenneman, C. G., and Sawyer, D. B. (2016). Cardio-oncology: an update on cardiotoxicity of cancer-related treatment. *Circ. Res.* 118, 1008–1020. doi: 10.1161/CIRCRESAHA.115.303633

Liebler, D. C., and Guengerich, F. P. (2005). Elucidating mechanisms of drug-induced toxicity. *Nat. Rev. Drug Discov.* 4, 410–420. doi: 10.1038/nrd1720

Luecken, M. D., Page, M. J. T., Crosby, A. J., Mason, S., Reinert, G., and Deane, C. M. (2018). CommWalker: correctly evaluating modules in molecular networks in light of annotation bias. *Bioinformatics* 34, 994–1000. doi: 10.1093/bioinformatics/btx706

Luo, W., Pant, G., Bhavnasi, Y. K., Blanchard, S. G. Jr., and Brouwer, C. (2017). Pathview Web: user friendly pathway visualization and data integration. *Nucleic Acids Res.* 45, W501–W508. doi: 10.1093/nar/gkx372

Maillet, A., Tan, K., Chai, X., Sadananda, S. N., Mehta, A., Ooi, J., et al. (2016). Modeling doxorubicin-induced cardiotoxicity in human pluripotent stem cell derived-cardiomyocytes. *Sci. Rep.* 6:25333. doi: 10.1038/srep25333

Marco-Ramell, A., Palau-Rodriguez, M., Alay, A., Tulipani, S., Urpi-Sarda, M., Sanchez-Pla, A., et al. (2018). Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC Bioinformatics* 19:1. doi: 10.1186/s12859-017-2006-0

Martha, V. S., Liu, Z., Guo, L., Su, Z., Ye, Y., Fang, H., et al. (2011). Constructing a robust protein-protein interaction network by integrating multiple public databases. *BMC Bioinformatics* 12(Suppl. 10):S7. doi: 10.1186/1471-2105-12-S10-S7

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., De Bono, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 37, D619–D622. doi: 10.1093/nar/gkn863

McGillivray, P., Clarke, D., Meyerson, W., Zhang, J., Lee, D., Gu, M., et al. (2018). Network analysis as a grand unifier in biomedical data science. *Annu. Rev. Biomed. Data Sci.* 1, 153–180. doi: 10.1038/ncomms10031

McGowan, J. V., Chung, R., Maulik, A., Piotrowska, I., Walker, J. M., and Yellon, D. M. (2017). Anthracycline chemotherapy and cardiotoxicity. *Cardiovasc. Drugs Ther.* 31, 63–75. doi: 10.1007/s10557-016-6711-0

Mei, N., Fuscoe, J. C., Lobenhofer, E. K., and Guo, L. (2010). Application of microarray-based analysis of gene expression in the field of toxicogenomics. *Methods Mol. Biol.* 597, 227–241. doi: 10.1007/978-1-60327-389-3_16

Mills, P. J., Ancoli-Israel, S., Parker, B., Natarajan, L., Hong, S., Jain, S., et al. (2008). Predictors of inflammation in response to anthracycline-based chemotherapy for breast cancer. *Brain Behav. Immun.* 22, 98–104. doi: 10.1016/j.bbi.2007.07.001

Mulas, F., Li, A., Sherr, D. H., and Monti, S. (2017). Network-based analysis of transcriptional profiles from chemical perturbations experiments. *BMC Bioinformatics* 18:130. doi: 10.1186/s12859-017-1536-9

Nystrom-Persson, J., Natsume-Kitatani, Y., Igarashi, Y., Satoh, D., and Mizuguchi, K. (2017). Interactive Toxicogenomics: gene set discovery, clustering and analysis in Toxygates. *Sci. Rep.* 7:1390. doi: 10.1038/s41598-017-01500-1

Paules, R. (2003). Phenotypic anchoring: linking cause and effect. *Environ. Health Perspect.* 111, A338–A339. doi: 10.1289/ehp.111-a338

Platel, D., Pouna, P., Bonoron-Adele, S., and Robert, J. (1999). Comparative cardiotoxicity of idarubicin and doxorubicin using the isolated perfused rat heart model. *Anticancer Drugs* 10, 671–676. doi: 10.1097/00001813-199908000-00007

Raschi, E., Vasina, V., Ursino, M. G., Boriani, G., Martoni, A., and De Ponti, F. (2010). Anticancer drugs and cardiotoxicity: insights and perspectives in the era of targeted therapy. *Pharmacol. Ther.* 125, 196–218. doi: 10.1016/j.pharmthera.2009.10.002

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Rueda-Zarate, H. A., Imaz-Rosshandler, I., Cardenas-Ovando, R. A., Castillo-Fernandez, J. E., Noguez-Monroy, J., and Rangel-Escareno, C. (2017). A computational toxicogenomics approach identifies a list of highly hepatotoxic compounds from a large microarray database. *PLoS One* 12:e0176284. doi: 10.1371/journal.pone.0176284

Sauter, K. A., Wood, L. J., Wong, J., Iordanov, M., and Magun, B. E. (2011). Doxorubicin and daunorubicin induce processing and release of interleukin-1beta through activation of the NLRP3 inflammasome. *Cancer Biol. Ther.* 11, 1008–1016. doi: 10.4161/cbt.11.12.15540

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Res.* 37, D674–D679. doi: 10.1093/nar/gkn653

Schramm, S. J., Jayaswal, V., Goel, A., Li, S. S., Yang, Y. H., Mann, G. J., et al. (2013). Molecular interaction networks for the analysis of human disease: utility, limitations, and considerations. *Proteomics* 13, 3393–3405. doi: 10.1002/pmic.201200570

Steinherz, L. J., Steinherz, P. G., Tan, C. T., Heller, G., and Murphy, M. L. (1991). Cardiac toxicity 4 to 20 years after completing anthracycline therapy. *JAMA* 266, 1672–1677. doi: 10.1001/jama.1991.03470120074036

Stillitano, F., Karakikes, I., Costa, K. D., Fish, K., Hajjar, R. J., and Hulot, J.-S. (2012). Preclinical animal models for testing iPSC/ESC-based heart therapy. *Drug Discov. Today* 9, e229–e236.

Sutherland, J. J., Webster, Y. W., Willy, J. A., Searfoss, G. H., Goldstein, K. M., Irizarry, A. R., et al. (2017). Toxicogenomic module associations with pathogenesis: a network-based approach to understanding drug toxicity. *Pharmacogenomics J.* 18:377. doi: 10.1038/tpj.2017.17

Tawa, G. J., Abdulhameed, M. D. M., Yu, X., Kumar, K., Ippolito, D. L., Lewis, J. A., et al. (2014). Characterization of chemically induced liver injuries using gene co-expression modules. *PLoS One* 9:e107230. doi: 10.1371/journal.pone.0107230

Truong, J., Yan, A. T., Cramarossa, G., and Chan, K. K. (2014). Chemotherapy-induced cardiotoxicity: detection, prevention, and management. *Can. J. Cardiol.* 30, 869–878. doi: 10.1016/j.cjca.2014.04.029

Uehara, T., Ono, A., Maruyama, T., Kato, I., Yamada, H., Ohno, Y., et al. (2010). The Japanese toxicogenomics project: application of toxicogenomics. *Mol. Nutr. Food Res.* 54, 218–227. doi: 10.1002/mnfr.200900169

Vandin, F., Clay, P., Upfal, E., and Raphael, B. J. (2012). Discovery of mutated subnetworks associated with clinical data in cancer. *Pac. Symp. Biocomput.* 17, 55–66. doi: 10.1142/9789814366496_0006

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522. doi: 10.1089/cmb.2010.0265

Vidal, M., Cusick, M. E., and Barabasi, A. L. (2011). Interactome networks and human disease. *Cell* 144, 986–998. doi: 10.1016/j.cell.2011.02.016

Walhout, A. J., and Vidal, M. (2001). Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell Biol.* 2, 55–62. doi: 10.1038/35048107

Wang, K. C., Botting, K. J., Padhee, M., Zhang, S., Mcmillen, I. C., Suter, C. M., et al. (2012). Early origins of heart disease: low birth weight and the role of the insulin-like growth factor system in cardiac hypertrophy. *Clin. Exp. Pharmacol. Physiol.* 39, 958–964. doi: 10.1111/j.1440-1681.2012.05743.x

Xia, K., Dong, D., and Han, J. D. (2006). IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics* 7:508.

Yan, J., Risacher, S. L., Shen, L., and Saykin, A. J. (2017). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform.* [Epub ahead of print]. doi: 10.1093/bib/bbx066

Yates, A., and Akanni, W. (2016). Ensembl 2016. *Nucleic Acid Res.* 44, D710–D716. doi: 10.1093/nar/gkv1157

Yildirimman, R., Brolen, G., Vilardell, M., Eriksson, G., Synnergren, J., Gmuender, H., et al. (2011). Human embryonic stem cell derived hepatocyte-like cells as a tool for in vitro hazard assessment of chemical carcinogenicity. *Toxicol. Sci.* 124, 278–290. doi: 10.1093/toxsci/kfr225

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26, 976–978. doi: 10.1093/bioinformatics/btq064

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:Article17. doi: 10.2202/1544-6115.1128

Zheng, G., and Huang, T. (2018). The Reconstruction and analysis of gene regulatory networks. *Methods Mol. Biol.* 1754, 137–154. doi: 10.1007/978-1-4939-7717-8_8

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 October 2018
doi: 10.3389/fgene.2018.00485

A Comparison of the TempO-Seq S1500+ Platform to RNA-Seq and Microarray Using Rat Liver Mode of Action Samples

Pierre R. Bushel[1]*, *Richard S. Paules*[2] and *Scott S. Auerbach*[2]*

[1] Biostatistics and Computational Biology Branch, NIEHS, Research Triangle Park, Durham, NC, United States,
[2] Biomolecular Screening Branch, National Toxicology Program, NIEHS, Research Triangle Park, Durham, NC, United States

The TempO-Seq™ platform allows for targeted transcriptomic analysis and is currently used by many groups to perform high-throughput gene expression analysis. Herein we performed a comparison of gene expression characteristics measured using 45 purified RNA samples from the livers of rats exposed to chemicals that fall into one of five modes of action (MOAs). These samples have been previously evaluated using Affymetrix™ rat genome 230 2.0 microarrays and Illumina® whole transcriptome RNA-Seq. Comparison of these data with TempO-Seq analysis using the rat S1500+ beta gene set identified clear differences in the platforms related to signal to noise, root mean squared error, and/or sources of variability. Microarray and TempO-Seq captured the most variability in terms of MOA and chemical treatment whereas RNA-Seq had higher noise and larger differences between samples within a MOA. However, analysis of the data by hierarchical clustering, gene subnetwork connectivity and biological process representation of MOA-varying genes revealed that the samples clearly grouped by treatment as opposed to gene expression platform. Overall these findings demonstrate that the results from the TempO-Seq platform are consistent with findings on other more established approaches for measuring the genome-wide transcriptome.

Keywords: TempO-Seq, S1500+, microarray, RNA-Seq, mode of action, chemicals, toxicants, toxicogenomics

OPEN ACCESS

Edited by:
Danyel Jennen,
Maastricht University, Netherlands

Reviewed by:
Richard John Brennan,
Sanofi, United States
Channa Keshava,
Environmental Protection Agency (EPA), United States

***Correspondence:**
Pierre R. Bushel
bushel@niehs.nih.gov
Scott S. Auerbach
auerbachs@niehs.nih.gov

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Genetics

Received: *30 July 2018*
Accepted: *28 September 2018*
Published: *30 October 2018*

Citation:
Bushel PR, Paules RS and Auerbach SS (2018) A Comparison of the TempO-Seq S1500+ Platform to RNA-Seq and Microarray Using Rat Liver Mode of Action Samples. Front. Genet. 9:485. doi: 10.3389/fgene.2018.00485

INTRODUCTION

High-throughput transcriptomics (HTT) is increasingly being adopted for screening in chemical and toxicological genomics in part due to advances in technological (i.e., direct from lysate transcriptomics) and greater efficiency (e.g., target screening using sentinel genes; Subramanian et al., 2017). The National Toxicology Program has pursued the development of the S1500+ gene set (Mav et al., 2018) screening platform utilizing the TempO-Seq™ technology from BioSpyder™ (Yeakley et al., 2017). Before there is widespread adoption of a new transcriptomic technology such as the TempO-Seq S1500+ platform, it will be important to establish its performance and degree of reproducibility compared to other more established techniques for gene expression assessment including microarray and whole transcriptome RNA-Seq. In addition to baseline performance

issues such as signal to noise and identification of appropriate normalization procedures (Su et al., 2014), it is also critical to determine reproducibility of findings from established legacy platforms particularly in the case where large compendium data such as the Connectivity Map (Lamb et al., 2006) or the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Igarashi et al., 2015) have been generated and serve as means to interpret new findings derived from newer technologies such as TempO-Seq. In addition, it is important for biologists that stand-alone assessments of gene set enrichment yield valid findings consistent with established modes or mechanisms of action and scalability of machine learning classifiers established using older technology (Waters et al., 2010).

To address the absolute and relative performance metrics of the rat S1500+ beta gene set TempO-Seq platform we have measured the transcriptome of identical liver RNA samples from the DrugMatrix database that were used to evaluate the performance of whole transcriptome RNA-Seq compared to microarray toward the SEquence Quality Control (SEQC)/MicroArray Quality Control III (MAQC3) toxicogenomics study in which the transcripts from the latter two platforms were matched for a fair comparison (Gong et al., 2014). The training data set consists of 63 samples measured using TempO-Seq S1500+, Illumina® whole transcriptome RNA-Seq, and Affymetrix™ Rat 230 2.0 microarrays. From the exposures of the rats to the chemicals, five different modes of action (MOAs) in the liver are represented in the samples including orphan nuclear hormone receptors (CAR/PXR) activation, aryl hydrocarbon receptor (AhR) activation, peroxisome proliferator-activated receptor alpha (PPARA) activation, cytotoxicity, and DNA Damage (**Table 1**). The treatments used vary considerably in their elicited transcriptomic signal (i.e., number of MOA-varying genes) and reveal degrees of distinctiveness in the altered gene sets which is ideal for establishing the level of granularity/resolution by which the technologies produce similarity in their resultant findings. Using the DrugMatrix samples we provide here a systematic comparison of the TempO-Seq technology relative to microarray and whole transcriptome RNA-Seq.

MATERIALS AND METHODS

Samples and Exposures

Mode of action (MOA) samples, preparation of them, RNA extraction and microarray and RNA-Seq analyses are as previously described (Wang et al., 2014). Briefly, male Sprague-Dawley rats (aged 6–8 weeks and weighing 200–260 g) were dosed once daily in triplicate for 3, 5, or 7 days, depending on the test chemical, and livers were harvested 24 h after the last dose. Animals were handled in accordance with the United States Department of Agriculture and Code of Federal Regulations Animal Welfare Act (9 CFR Parts 1, 2, and 3). Details on the design and in life portion of these studies can be found elsewhere. For each of the five MOAs there were three test chemicals (**Table 1**). RNAs from the treated rats were extracted and stored in the National Toxicology Program (NTP) DrugMatrix Frozen Tissue Library.

Microarray Analysis

cRNA was labeled and hybridized to the Affymetrix (Santa Clara, CA, United States) whole genome GeneChip® Rat Genome 230 2.0 Array as previously described (Wang et al., 2014). The arrays were scanned using the GeneChip Scanner 3000 7G and CEL files generated using the GeneChip Operating Software (GCOS). The data was then $\log_2$ transformed and normalized using the robust multichip average (RMA) algorithm (Irizarry et al., 2003a,b). The transformed/normalized data is available at the DrugMatrix ftp site (ftp://anonftp.niehs.nih.gov/drugmatrix/Affymetrix_data/Normalized_data_by_organ/). Raw data files and processed data in various file formats are available in the Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2013) under accession number GSE47875.

RNA-Seq Analysis

Poly-A RNA was extracted from each RNA sample, fragmented, adapter ligated and enriched by 15 polymerase chain reaction (PCR) cycles for library generation. The library size distribution was validated on the Agilent Bioanalyzer (Santa Clara, CA, United States) using a DNA 1000 kit. The final library was generated from a band between 200 and 500 bp with a peak at ∼260 bp. Using Illumina TruSeq RNA Sample Preparation Kit and SBS Kit v3 (San Diego, CA, United States), samples were prepared for sequencing. Paired-end RNA-Seq cluster generation and sequencing by synthesis was performed using Illumina HiScan or HiSeq 2000 sequencers according to the manufacture's protocol. Depths of 30–130 million of paired 100 bp reads were generated for each sample. Details of the methods are as previously described (Wang et al., 2014). The raw data fastq files are available in the National Center for Biotechnology Information Sequence Read Archive (SRA; Leinonen et al., 2011) under accession number SRP039021.

Preprocessing of RNA-Seq Data

Alignment, quantification and normalization of the RNA-Seq data are as previously described (Wang et al., 2014). Briefly, RNA-Seq reads in fastq files were mapped using the Magic aligner (ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic) to the following references:

- The *Rattus norvegicus* genome build RGSC v3.4
- The RefSeq and AceView 2008 (Thierry-Mieg and Thierry-Mieg, 2006) gene and transcript models, respectively
- Mitochondrial genes
- rRNA genes (manually constructed from multiple GenBank accessions, in the absence of RefSeq)
- External RNA Control Consortium (ERCC) RNA spike in control sequences (National Institute of Standards and Technology, Gaithersburg, MD, United States)
- A control genome constructed by complementing the *R. norvegicus* genome bases (i.e., exchange A:T and G:C), but not reversing the order. As such, the control genome has exactly the same composition as the reference genome but alignments to it are false positives and removed

TABLE 1 | Chemicals, modes of action, and exposures.

MOA	Chemical	Dose (mg/kg body weight)	Duration (days)	Agent type
Aryl hydrocarbon receptor (AhR)	3-Methylcholanthrene (3ME)	300	5	Carcinogen
	Leflunomide (LEF)	60	5	Antirheumatic drug
	beta-Naphthoflavone (NAP)	1,500	5	Putative chemopreventive agent
Orphan nuclear hormone receptors (CAR/PXR)	Phenobarbital (PHE)	54	5	Barbiturate drug
	Methimazole (MET)	100	3	Antithyroid drug
	Econazole (ECO)	334	5	Antifungal medication
Cytotoxicity (Cytotox)	Chloroform (CHO)	600	5	Organic compound
	Thioacetamide (THI)	200	5	Carcinogen
	Carbon tetrachloride (CAR)	1,175	7	Solvent for cleaning products, refrigerant
DNA Damage (DNA_Damage)	Aflatoxin B1 (AFL)	0.3	5	Mycotoxin
	Ifosfamide (IFO)	143	3	Chemotherapy drug
	N-Nitrosodimethylamine (NIT)	10	5	Organic compound
Peroxisome proliferator-activated receptor alpha (PPARA)	Pirinixic acid (PIR)	364	5	Hypolipidemic drug
	Bezafibrate (BEZ)	617	7	Hypolipidemic drug
	Nafenopin (NAF)	338	5	Hypolipidemic drug

No mismatch is reported closer than 8 bases to the edge of the aligned segment. Reads mapping to several alternative transcripts of the same gene are retained but counted only once. The read count for each transcript per sample was transformed and normalized as follows:

$$\text{Index} = log_2\left(Z + \sqrt{(4 + Z^2)}\right) - 1$$

Where $Z = 10^{12}(\frac{n}{NL})$, n is the read count of the transcript, N is the read depth for the sample and L is the length of the transcript. For transcripts that are not highly expressed (<3 read counts) the Index was imputed with 5.0. The preprocessed data (not imputed) is available in GEO under accession number GSE55347.

To match AceView transcripts from the RNA-Seq platform to probe sets on the Affymetrix microarray, each transcript sequence was mapped against the Affymetrix probes from each probe set using the Magic Aligner and allowing for a single-mismatch. Transcripts ($n = 28,975$) mapping to at least 8 probes within a probe set unambiguously (meaning not mapping to any other probes from other probe sets) are considered a one-to-one match in terms of them being representative of the same transcript probe set. These were then mapped to UniGene (Pontius et al., 2002) cluster IDs (March 30, 2016) for Gene Ontology (GO) biological process (BP) enrichment analysis.

TempO-Seq Analysis

The sequencing library for the rat liver RNA samples (identical samples employed for RNA-Seq in the previously published SEQC toxicogenomics study Wang et al., 2014) was prepared by BioSpyder Technologies, Inc. (Carlsbad, CA, United States) according to their protocol guidelines. One microliter of each RNA sample (500–660 ng/uL) was hybridized with the S1500+ beta detector oligo pool mix (2 µl per sample) using the following thermocycler settings: 10 min at 70°C, followed by gradual decrease to 45°C over 49 min, and ending with 45°C for 1 min. Hybridization was followed by nuclease digestion (24 µl nuclease mix addition followed by 90 min at 37°C), ligation (24 µl ligation mix addition followed by 60 min at 37°C), then heat denaturation (at 80°C for 30 min). Ten microliters of each ligation product were then transferred to a 96-well PCR amplification microplate that also contained 10 µl of PCR mix per well. Through amplification well-specific, "barcoded" primer pairs were introduced to templates. Five microliters of the PCR amplification products from each well were then pooled into a single sequencing library. The TempO-Seq library was then processed with a PCR clean-up kit (Machery-Nagel, Mountain View, CA, United States) prior to sequencing. Sequencing was performed using a 50 cycle single-end read flow cell on a NextSeq 550 Sequencing System (Illumina, San Diego, CA, United States). Processing of sequencing data was conducted using Illumina's BCL2FASTQ software employing default parameter settings. Sequencing data were demultiplexed to generate fastq files and passed through internal quality controls. fastq files were analyzed using the TempO-SeqR software package (BioSpyder Technologies, Inc., Carlsbad, CA, United States). The raw data fastq files are available in the SRA under accession number SRP158667. The TempO-SeqR package maps reads from the fastq file using the Bowtie2-2.1.0 algorithm (Langmead et al., 2009) to a subset of the rat transcriptome (Refseq release 70 downloaded July 23rd 2015) reflecting the 50 nt sequences targeted by the detector oligos. Indels were not allowed, up to 2 base pair mismatches were allowed and multimapping of sequence reads was not allowed. The output of the TempO-SeqR package was a table of counts with each column representing a sample and each row representing a gene generated using the QuasR v1.8.4 Bioconductor package (Gaidatzis et al., 2015). The count data matrix is available in GEO under accession number GSE118956.

Gene symbols were mapped to UniGene cluster IDs (June 6, 2015).

Preprocessing of TempO-Seq Data

Of the 2,284 genes targeted in the rat S1500+ beta gene set (NTP Tox21 S1500 Webpage: https://ntp.niehs.nih.gov/results/tox21/researchphases/index.html), those with a total read count $\leq$214 across all the samples were removed leaving 2,055 genes. The counts per gene were normalized to counts per million (CPM) by dividing it by the total read count per sample and multiply by 10^6. The CPM normalized data was then transformed with $\log_2$ using an offset of 1.

Log_2 Ratio Values Generation

For each gene/transcript in a data set, the average of the $\log_2$ normalized data for the control samples were subtracted from the $\log_2$ normalized data of each gene/transcript within a sample matched according to nutritional status of the vehicle (i.e., corn oil vs. other non-nutritive vehicles).

Principal Variance Component Analysis

Principal Variance Components Analysis (PVCA; Li et al., 2009) combines the use of principal component analysis (PCA) with variance components analysis (VCA) through mixed linear modeling of gene expression data with random effect terms that account for variation related to factors in the experimental design. The variance of each random effect is called a variance component. Briefly, given a general linear model where $y = X\beta + e$ and y denotes gene expression observations, X is the design matrix, β is the known fixed effects parameter vector and e is the unexplained variation. However, if the experimental design contains random factor levels, the model becomes a mixed effect linear model $y = X\beta + Zu + e$, where in addition to the terms denoted in a fixed effect model, Z is the design matrix for random effects, u is the vector of unknown random-effect parameters, and e is the unobserved vector of independent and identically distributed (iid) Gaussian random errors.

Given that the variance of y is $\mathbf{V=ZGZ' + R}$, **V** can be modeled by setting up the random effects design matrix **Z** and by specifying the variance-covariance structure for **G** and **R**. In usual variance component models, **G** is a diagonal matrix with variance components on the diagonal, each replicated along the diagonal corresponding to the design matrix **Z**. **R** is simply the residual variance component times the n x n identity matrix. Thus, the goal becomes finding a reasonable estimate of **G** and **R**. The method of restricted maximum likelihood (REML) is the standard procedure to accomplish this and was specified in the **lmer** function of the lme4 R package (R Development Core Team, 2012) for fitting linear mixed effects models (Bates et al., 2015).

The following steps comprise of PVCA:

- From a PxN (genes by samples) matrix of $\log_2$ ratio values, obtain the NxN correlation matrix
- Perform PCA on the correlation matrix to obtain eigenvalues
- Determine the first K principal components (PCs) to explain $\geq$58.76% of the variation in the data
- Fit all factors as random effects in a mixed linear model using the K PCs and REML to obtain unbiased estimates of variance
- Standardize the variance component estimates from the model
- Compute weighted proportions of the standardized variance component estimates. Here the weights are the proportions of variation explained by the PCs
- Compute weighted average proportions of the standardize variance component estimates by averaging model effects according to the proportion of total variance across all estimates including the residual

Root Mean Squared Distance

Root mean squared distance (RMSD) is a measure of the gene expression distance between pairs of biological replicates (Wang et al., 2014). The gene expression distance between biological replicates *x* and y is

$$RMSD_{xy} = \sqrt{\frac{\sum_{i=1}^{N} \left(I_{ix} - I_{iy}\right)^2}{N}}$$

where I is the $\log_2$ gene expression ratio of ith gene/transcript in the corresponding biological replicate, and N is the number of genes/transcripts on the gene expression platform. For each chemical, there are three biological replicates and for each MOA there are three chemicals. The pairwise RMSD measures (n =

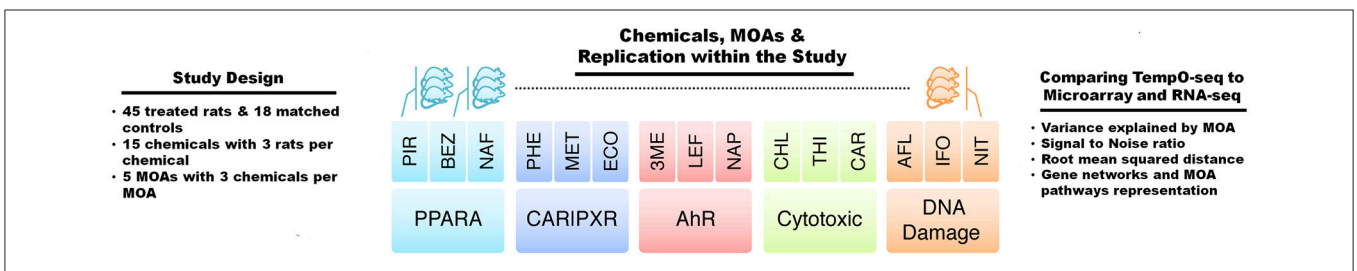


FIGURE 1 | Study design. The study comprised of gene expression data acquired from male Sprague-Dawley rats dosed once daily in triplicate for 3, 5, or 7 days depending on the test chemical or matched control, and livers were harvested 24 h after the last dose. The abbreviations for the names of the chemicals are listed in **Table 1**. There were five modes of action (MOAs) with three chemicals per MOA. The MOAs are PPARA, peroxisome proliferator-activated receptor alpha; CAR/PXR, orphan nuclear hormone receptors; AhR, aryl hydrocarbon receptor; Cytotoxic, cytotoxicity, and DNA Damage. Comparisons between the data from TempO-Seq to microarray and RNA-Seq were performed by statistical and bioinformatics methodologies.

36) between treated rats within a MOA were averaged. This MOA-RMSD can be interpreted as a measure of the difference among the chemicals within a MOA. To compare the three gene expression platforms, the five MOA-RMSD measures were averaged to give a Platform-RMSD.

Mode of Action ANOVA

To obtain genes from each platform that vary significantly by MOA, we modeled the gene expression data with a MOA analysis of variance (MOA-ANOVA)

$$Y_{ijkl} = \mu + M_i + R_j + C(M^*R)_{ijk} + \varepsilon_{ijkl}$$

where Y_{ijkl} represents the lth $\log_2$ ratio gene expression observation on the ith MOA (M), jth route (R) and kth chemical (C). μ is the grand mean for the whole experiment and ε_{ijkl} represents the random error. The errors are assumed to be normally and independently distributed with mean 0 and standard deviation δ for all measurements. Chemical is a random effect. Multiple testing correction was controlled at a false discovery rate (FDR) of 0.05 (Benjamini and Hochberg, 1995).

TABLE 2 | Platforms used for comparison.

Gene expression type	Microarray	RNA-Seq	TempO-Seq
Platform	Affymetrix whole genome GeneChip Rat Genome 230 2.0	Illumina HiScan & HiSeq 2000	BioSpyder S1500+ Beta
Technology	*In situ* oligonucleotide array	Next generation nucleotide chain termination sequencing by synthesis	Templated oligonucleotide detection
Gene content/gene model	~31,000 gene probe sets	~38,100 AceView transcripts	~2,200 Refseq genes
Normalization	RMA	Magic normalized index	TPM
Transformation	Log_2	Log_2	Log_2

Gene Expression Profile Signal to Noise

Let us denote each gene expression $\log_2$ ratio as g_{ij} where i indicates a MOA inter-group index from 1 to m, j is the MOA intra-group index from 1 to n_i, m is the number of MOAs and n_i is the number of chemicals in i^{th} MOA inter-group. To evaluate a gene expression profile within a MOA, we calculate each MOA intra-group average $\bar{g}_i$ and sample variance s_i^2. We define a gene expression profile's signal as

$$S = \begin{cases} \max\left\{\bar{g}_i\right\}, \textit{if } \min\left\{\bar{g}_i\right\} & > 0 \\ -\min\left\{\bar{g}_i\right\}, \textit{elseif } \max\left\{\bar{g}_i\right\} & < 0 \\ \max\left\{\bar{g}_i\right\} - \min\left\{\bar{g}_i\right\} & \textit{otherwise} \end{cases}$$

where $1 \leq i \leq m$.

We then define a gene expression profile's noise as the square-root of the pooled variance

$$N = \sqrt{\frac{\sum_i^m \left[(n_i - 1) \cdot s_i^2\right]}{\sum_i^m (n_i - 1)} \sum_i^m \frac{1}{n_i}}$$

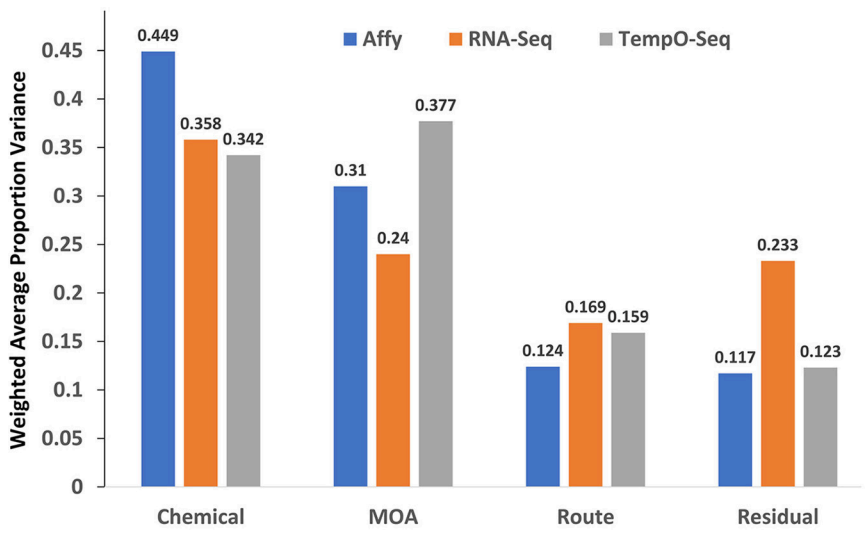


FIGURE 2 | Variance components explained. Shown on the y-axis is the weighted average of the proportion of variance explained by platform for each of the mixed effect linear model terms denoted in the x-axis.

where the sample variance

$$s_i^2 = \frac{\sum_j^{n_i} (g_{ij} - \bar{g}_i)^2}{n_i - 1}.$$

From S and N, we define a gene expression profile's signal-to-noise ratio as $SNR = S/N$. We use Extracting Patterns and Identifying co-Expressed Genes (EPIG; Chou et al., 2007) to (1) obtain a gene expression profile's SNR statistics and (2) cluster gene expression profiles into significant ($p < E10^{-4}$) co-expression patterns.

Gene Ontology Subtrees to Tag and Annotate Genes Within a Set

To compare each platform in terms of enrichment of the genes that vary by MOA, we used GO subtrees to tag and annotate genes (goSTAG) within a set (Bennett and Bushel, 2017). Briefly, for each list of genes that vary by MOA at FDR < 0.01, the gene symbols were mapped to the GO BPs of the genes they represent using version 3.4 of the GO database and the rat2302 database. The 1.01 version of the "goSTAG" Bioconductor package in R was used to perform enrichment of GO BP terms, clustering and subtree generation. The union of the enriched GO BP terms from all of the DEGs lists yielded 203 terms. BP terms which were not significant had missing p-values and were imputed with 1.0. – Log_{10} p-values, a min of 5 genes per GO BP, FDR < 0.05, Pearson correlation similarity metric and Ward algorithm for clustering, cluster slicing using correlation (r) of 0.1 and a minimum of 5 GO BP terms per cluster for subtree generation were used as input and parameters. Clusters (those with a $1 - r \geq 0.9$) of GO BP terms ($n \geq 5$) were labeled according to the node having the maximum number of paths to it within the GO BP subtree directed acyclic graph derived from the terms in the cluster.

RESULTS

Study Design and Exposures

Gene expression analysis has advanced over the past 20+ years. Two main platforms for surveying genome-wide gene expression are microarray and RNA-Seq. Each of these platforms has its advantages and disadvantages (Lowe et al., 2017). The SEQC/MAQC3 consortium evaluated the concordance between Affymetrix microarray and Illumina RNA-Seq using

TABLE 3 | Replication agreement and signal to noise within platform.

Measure	Microarray	RNA-Seq	TempO-Seq
Ave. Chemical-RMSD	0.33	0.98	0.71
Platform-RMSD	0.42	1.11	0.93
Average SNR	6.6	6.9	9.14

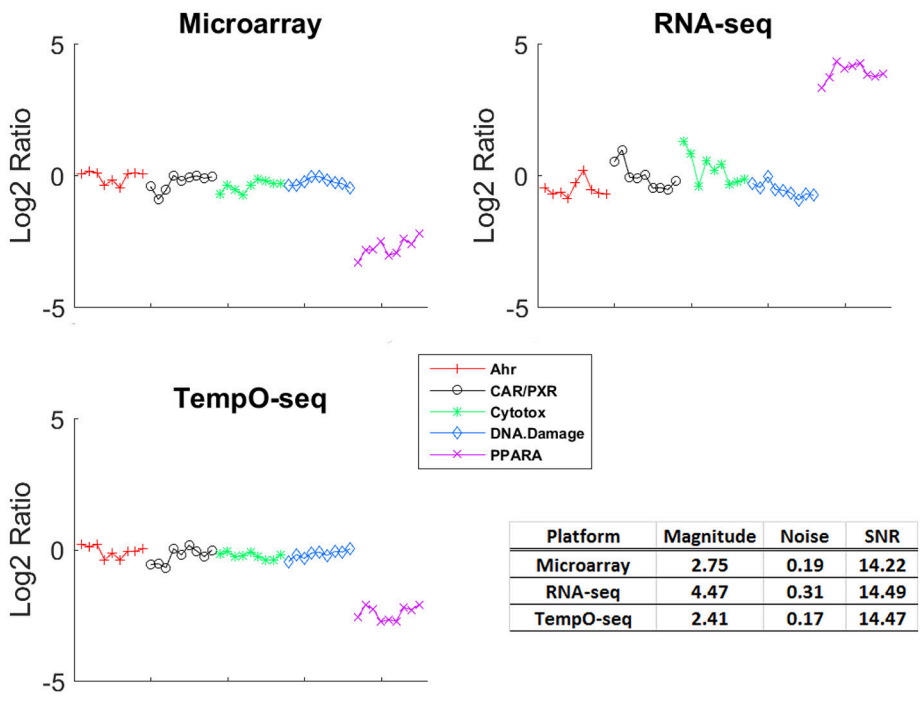


FIGURE 3 | Gene expression patterns with maximal signal to noise. For each platform, the EPIG pattern with the maximal signal to noise ratio (SNR) is shown. The y-axis is the $\log_2$ ratio of gene expression (treated to the average of the control [matched according to nutritional status of the vehicle]), the x-axis is the samples grouped by MOA (represented by the colors and symbols in the legend). The table inset displays the magnitude of fold change, the noise and the SNR for each of the patterns shown.

toxicogenomics gene expression data (Wang et al., 2014). We used the SEQC/MAQC3 study design to compare the two aforementioned platforms (with transcripts matched between the two) with the TempO-Seq platform targeting the rat S1500+ beta gene set. As shown in **Figure 1**, the study design consists of rats exposed in triplicate to 45 chemical or controls whereby three of the chemicals share one of five MOAs: PPARA, CAR/PXR, AhR, cytotoxicity, and DNA damage. The chemicals, the MOA that each one represents, exposure doses and durations and the types of agents are listed in **Table 1**. The doses and durations of the exposures were selected to ensure a maximal transcriptional response. Animals were dosed once daily for 3, 5, or 7 days, depending on the chemical. Livers were harvested 24 h after the last dose, RNA samples extracted and then prepared for gene expression analysis. We used three statistical strategies and bioinformatics tools to examine GO BPs, metabolic pathways and BP subnetworks for comparison of the three platforms.

Specifications of the Platforms

Table 2 details some general specifications of the three gene expression platforms. The Affymetrix rat whole genome microarray with >31,000 gene probe sets uses *in situ* hybridization for interrogation of gene expression. The *de facto* normalization procedure is RMA. The Illumina RNA-Seq next generation HiScan or HiSeq 2000 platforms were used. They measure gene expression by nucleotide chain termination sequencing by synthesis. Although at this time there is no standard approach for bioinformatics analysis of RNA-Seq data, we used the AceView transcriptome gene model and Magic normalization index that performed the best among several bioinformatics pipelines in the SEQC/MAQC3 consortium evaluation (Wang et al., 2014). In addition, 28,975 transcripts from the two aforementioned platforms were matched bioinformatically (see the Materials and Methods section) to assure a one-to-one mapping. Finally, BioSpyder's rat S1500+ beta TempO-Seq platform differs from RNA-Seq in that it uses templated oligonucleotides representative of >2,200 Refseq genes to sequence captured RNA templates. Filtering by total read counts retained 2,055 genes (see the Materials and Methods section). We used CPM for normalization. The data from all three platforms were $\log_2$ transformed to make the data more normally distributed.

Variance Components of the Study Design Captured by the Platforms

The study design contained factors that represents the chemical used for exposure, the MOA of the chemical and the route of the exposure. We performed PVCA on the normalized and $\log_2$ transformed data from each platform to determine which

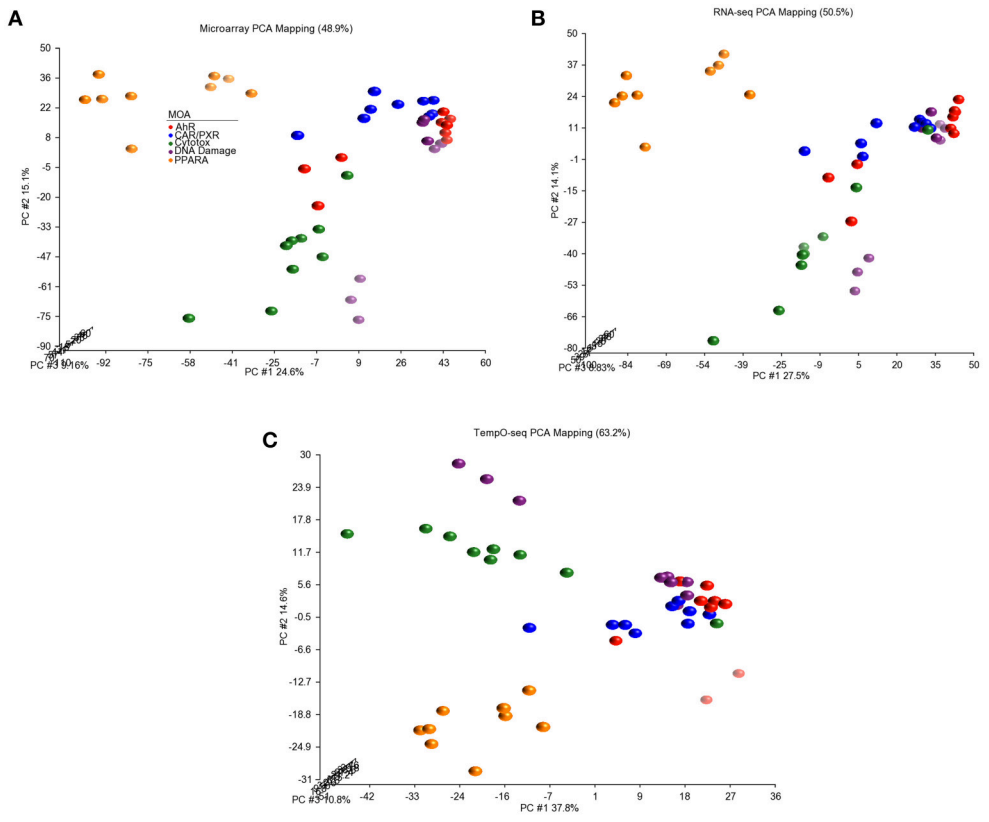


FIGURE 4 | Principal component analysis of the data. **(A)** Microarray. **(B)** RNA-Seq. **(C)** TempO-Seq. PCA performed using the $\log_2$ ratio expression data (treated to matched control according to nutritional status) of the genes that vary by MOA at FDR < 0.01.

captured the most variation in gene expression. As shown in **Figure 2**, the microarray platform captured slightly more variance related to the chemical used for treatment (0.449), but the TempO-Seq platform captured variation related to the MOA (0.377) slightly more than the other two platforms. It seems that the RNA-Seq platform had more unexplained variation captured as residuals. This was not related to the two different Illumina sequencers used (data not shown). Route showed no difference in the variation captured by the three platforms.

Expression Pattern Magnitude of Change and Signal to Noise Revealed by Each Platform

One of the more informative ways to compare gene expression data is to assess the magnitude of change and the SNR of a response. To compare the gene expression from the three platforms, we analyzed the data using EPIG which used magnitude of fold change, correlation and SNR to categorize gene expression profiles into co-expressed patterns (Chou et al., 2007). Shown in **Figure 3** is the pattern of gene expression from each platform that had the maximal magnitude of fold change relative to control. The samples were grouped by MOA. Although RNA-Seq had the highest magnitude of fold change (4.47), the noise of the expression profiles that made up the pattern is higher (0.31) than the other two platforms. When all the patterns for each platform were taken into consideration, the average SNR was substantially higher for TempO-Seq than the other two platforms (**Table 3**). This may be related to the EPIG analysis of the TempO-Seq data yielding only four patterns whereas microarray yielded 17 and RNA-Seq yielded 11 (data not shown).

Cohesiveness of Replicate Gene Expression by Platform

A unique design of the study is that there is replication at the animal level, the chemical level, and the MOA level (**Figure 1**). We harnessed this feature to assess how well each platform captured similar gene expression between replicates. We used RMSD to assess the gene expression distance between pairs of biological replicates. A smaller measure means the replicates are closer to each other in terms of gene expression. The platform-RMSD is an aggregate (overall average) of the distance between animals treated with a chemical, the chemicals within a MOA and the five MOAs. The average chemical-RMSD is the mean of the RMSDs for each chemical by platform. As shown in **Table 3**, the Platform-RMSD and average chemical-RMSD were more than 2 times lower for microarray than for the other

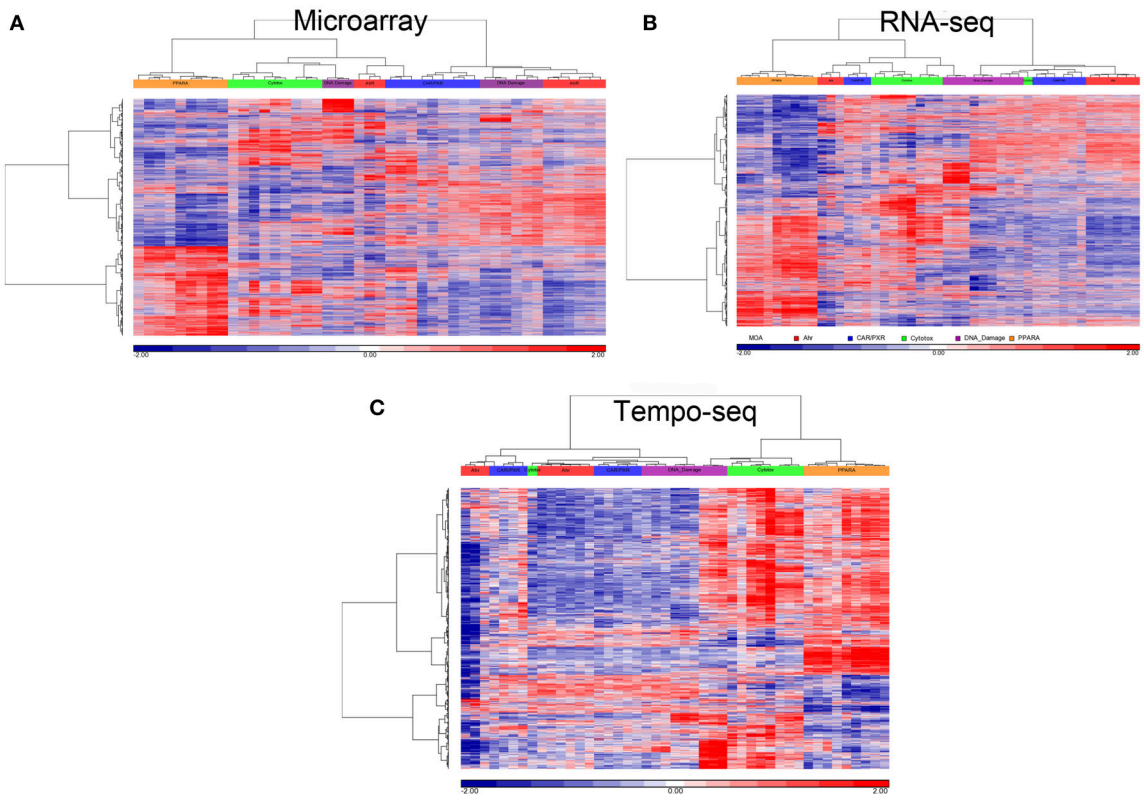


FIGURE 5 | Clustering of data. **(A)** Microarray. **(B)** RNA-Seq. **(C)** TempO-Seq. Clustering performed using the $\log_2$ ratio expression data (treated to matched control according to nutritional status) of the genes that vary by MOA at FDR < 0.01 with cosine correlation as the similarity metric and the Ward clustering criterion. The data for clustering was standardized to a mean of 0 and standard deviation of 1. Samples' MOA colored as in the legend to **Figure 4A**.

two platforms. RNA-Seq had the highest RMSD (1.11 at the platform level and 0.98 at the chemical level) which may be related to the higher noise level seen in the expression pattern for this platform (**Figure 3**). Despite the relatively noisy RNA-Seq platform, PCA of the gene expression data revealed that RNA-Seq captured a higher percent of the variability (55.5%) in the data than the other two platforms and also projected the samples in 3-dimensional space closer to each other in terms of MOA (Data not shown).

Biological Responsiveness by Platform

Since three chemicals share a MOA, for each platform we used an ANOVA model with MOA as a main factor to identify genes that vary significantly at an FDR < 0.01. For microarray, RNA-Seq and TempO-Seq, 9,499 probe sets, 7,217 transcripts and 1,366 genes were detected as varying, respectively (**Supplemental Table 1**). These genes should drive the clustering of the gene expression data by MOA. As shown in **Figures 4**, **5**, respectively, PCA and 2-dimensional hierarchical clustering

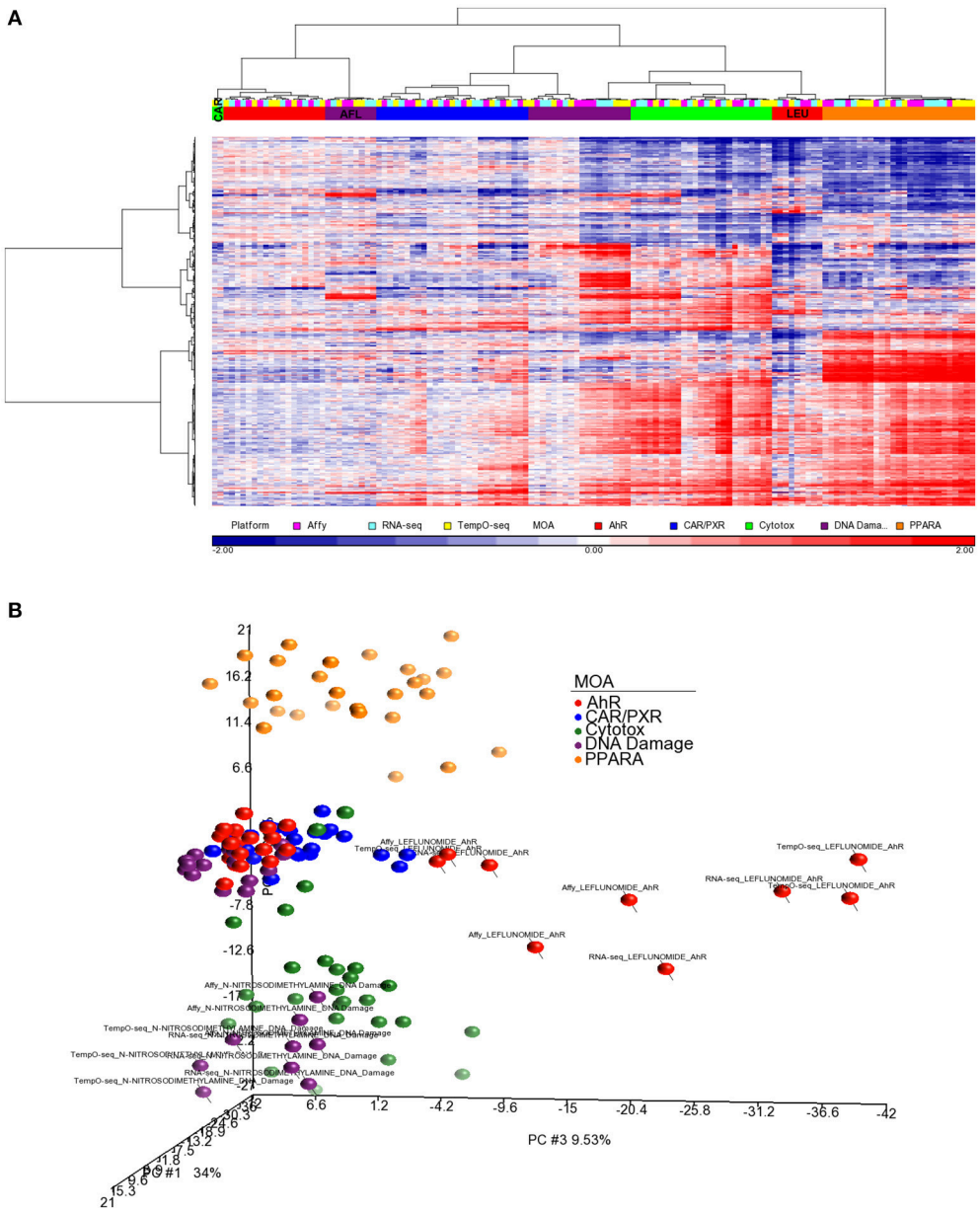


FIGURE 6 | Clustering of the data using a common gene set. **(A)** Hierarchical clustering performed using the $\log_2$ ratio expression data (treated to matched control according to nutritional status) of the genes that vary by MOA at FDR < 0.01 and map to 731 UniGene cluster IDs that overlap between the three platforms. Genes that were mapped to the same UniGene cluster ID were averaged. The cosine correlation was used as the similarity metric and the Ward clustering criterion for merging clusters. Samples' MOA colored as in the legend to **Figure 4A**. Platforms are represented by the following colors: pink, Affymetrix; light blue, RNA-Seq; yellow, TempO-Seq. **(B)** PCA of the data used in **(A)**. Principal component (PC) #1 = 34%, PC #2 = 16.6 %, and PC #3 = 9.53.

of the data from each platform by their MOA varying genes was reasonably good. However, each platform had at least two MOAs with a chemical that didn't cluster with its respective MOA chemicals. For microarray NIT and LEF didn't cluster with DNA damage and AhR MOA chemicals, respectively. For RNA-Seq LEF, ECO and one biological replicate of CAR didn't cluster with AhR, CAR/PXR, and cytotoxicity MOA chemicals, respectively. For TempO-Seq LEF and ECO didn't cluster with Ahr, and CAR/PXR MOA chemicals, respectively. In addition, one biological replicate of MET and CAR and didn't cluster with their biological replicates in the CAR/PXR and cytotoxicity MOA chemicals, respectively.

For better cluster resolution, we mapped the MOA varying genes from each platform to UniGene cluster IDs and then compiled the $\log_2$ ratio data from all three platforms using the 731 UniGene cluster IDs that overlapped (**Supplemental Table 2**). Genes that were mapped to the same UniGene cluster ID were averaged. As shown in **Figure 6A**, the clustering of the samples was mostly by MOA except for LEF, AFL and one biological replicate from CAR. PCA of the data captures ∼60% of the variation in the data and projected the samples in 3-dimensional space closer to each other in terms of MOA except for LEF and NIT samples from all three platforms (**Figure 6B**).

Enrichment of GO biological processes by the platforms' varying genes yielded 49 significant categories (FDR $<$ 5%) that overlapped (**Figure 7A**). Microarray had the most enriched categories (n = 173), followed by RNA-Seq (n = 141), and then TempO-Seq (n = 99). Some of the enriched GO biological processes that overlapped related to fatty acid metabolism, apoptosis, liver development, and lipid metabolism (**Table 4**). As shown in **Figure 7B**, the correlation of the 49 GO biological processes fold enrichment between the three platforms was very high ($r > +0.9$).

Comparing and contrasting gene set enrichments can be challenging when there are many categories to consider. To more formally compare the three platforms in terms of biology, we used goSTAG to identify subtrees of enriched GO BPs from the MOA varying genes and then find the categories that are shared or differ between platforms. As shown in **Figure 8**, all three platforms enriched for subtrees that map to fatty acid beta-oxidation and glycine metabolic process. However, TempO-Seq enriched for subtrees that map to negative regulation of ERK1 and ERK2 cascade. RNA-Seq uniquely enriched subtrees that map to ATP metabolic process and microarray exclusively enriched for subtrees that map to positive regulation of glycolytic process.

DISCUSSION

Over the last two decades gene expression analysis has advanced to permit genome-wide transcriptomics. Affymetrix microarray and Illumina RNA-Seq are two platforms that have gained popularity for gene expression analysis. Each has its own advantages and disadvantages but currently, the platform of choice for gene expression analysis seems to be RNA-Seq.

Comparison of the two platforms was performed using liver RNA samples from rats exposed to chemicals that have particular modes of action (MOAs; Wang et al., 2014; **Figure 1** and **Table 1**). We used the microarray and RNA-Seq training data from these samples to compare with the data generated from the samples using the TempO-Seq platform. TempO-Seq is unique in that the platform's gene content (∼2,200) consists of bioinformatically curated (Mav et al., 2018) and expert domain-nominated rat genes that represent the totality of biological perturbation space (**Table 2**). This makes the TempO-Seq platform very appealing for transcriptomics in that (1) sequencing of the RNA is from

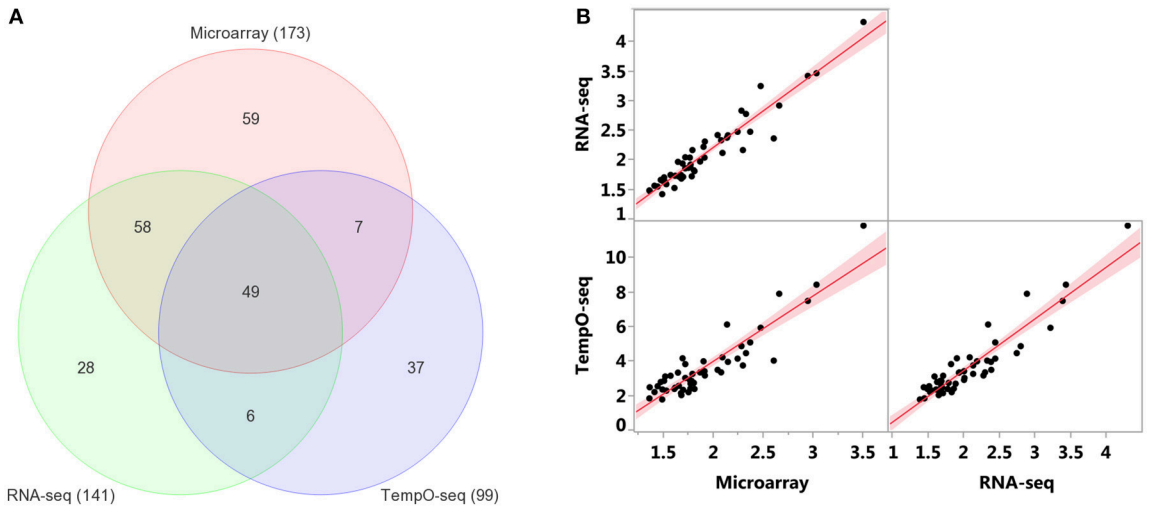


FIGURE 7 | Comparison of enriched GO biological processes (BPs). **(A)** Overlap of enriched GO BPs $<$ FDR 5%. Minimum number of genes = 3 for TempO-Seq and 5 for the other two. **(B)** Pairwise comparison of GO BPs fold enrichment from the 49 categories in common between the three platforms. Red line is the linear fit (regression line) with 95% level confidence boundaries.

TABLE 4 | Enriched GO BPs (FDR < 5%) that overlap between platforms®.

GOID	GO BP Term	Microarray				RNA-seq				TempO-seq			
		Count	%	Pop Hits	FE	Count	%	Pop Hits	FE	Count	%	Pop Hits	FE
GO:0001666	Response to hypoxia	119	2.11	270	1.55	99	2.18	270	1.59	39	3.32	270	2.28
GO:0001731	Formation of translation preinitiation complex	17	0.30	24	2.50	18	0.40	24	3.24	9	0.77	24	5.92
GO:0001889	Liver development	85	1.51	145	2.07	81	1.78	145	2.42	32	2.72	145	3.48
GO:0006413	Translational initiation	32	0.57	52	2.17	29	0.64	52	2.41	13	1.11	52	3.95
GO:0006446	Regulation of translational initiation	19	0.34	31	2.16	17	0.37	31	2.37	12	1.02	31	6.11
GO:0006457	Protein folding	57	1.01	112	1.79	49	1.08	112	1.89	17	1.45	112	2.40
GO:0006629	Lipid metabolic process	44	0.78	89	1.74	42	0.93	89	2.04	17	1.45	89	3.01
GO:0006631	Fatty acid metabolic process	33	0.58	60	1.94	32	0.71	60	2.31	12	1.02	60	3.16
GO:0006635	Fatty acid beta-oxidation	35	0.62	46	2.68	31	0.68	46	2.92	23	1.96	46	7.89
GO:0006637	Acyl-coa metabolic process	19	0.34	28	2.39	16	0.35	28	2.47	9	0.77	28	5.07
GO:0006695	Cholesterol biosynthetic process	17	0.30	26	2.30	17	0.37	26	2.83	8	0.68	26	4.86
GO:0006749	Glutathione metabolic process	31	0.55	52	2.10	28	0.62	52	2.33	11	0.94	52	3.34
GO:0006915	Apoptotic process	157	2.78	366	1.51	120	2.64	366	1.42	41	3.49	366	1.77
GO:0006953	Acute-phase response	25	0.44	38	2.32	19	0.42	38	2.16	9	0.77	38	3.74
GO:0006979	Response to oxidative stress	76	1.35	146	1.83	61	1.34	146	1.81	22	1.87	146	2.38
GO:0007568	Aging	150	2.66	315	1.68	125	2.75	315	1.72	51	4.34	315	2.56
GO:0007584	Response to nutrient	69	1.22	137	1.77	59	1.30	137	1.86	19	1.62	137	2.19
GO:0007623	Circadian rhythm	52	0.92	121	1.51	46	1.01	121	1.64	18	1.53	121	2.35
GO:0009636	Response to toxic substance	65	1.15	119	1.92	61	1.34	119	2.22	30	2.55	119	3.98
GO:0009749	Response to glucose	54	0.96	106	1.80	47	1.04	106	1.92	18	1.53	106	2.68
GO:0010033	Response to organic substance	72	1.28	152	1.67	69	1.52	152	1.96	32	2.72	152	3.32
GO:0010243	Response to organonitrogen compound	35	0.62	68	1.81	34	0.75	68	2.16	14	1.19	68	3.25
GO:0014070	Response to organic cyclic compound	138	2.45	272	1.79	128	2.82	272	2.04	50	4.26	272	2.90
GO:0031100	Organ regeneration	45	0.80	91	1.74	39	0.86	91	1.85	22	1.87	91	3.82
GO:0031667	Response to nutrient levels	49	0.87	112	1.54	42	0.93	112	1.62	22	1.87	112	3.10
GO:0032355	Response to estradiol	91	1.61	201	1.60	81	1.78	201	1.74	40	3.40	201	3.14
GO:0032496	Response to lipopolysaccharide	114	2.02	280	1.43	101	2.23	280	1.56	39	3.32	280	2.20
GO:0032869	Cellular response to insulin stimulus	66	1.17	123	1.89	56	1.23	123	1.97	26	2.21	123	3.34
GO:0033539	Fatty acid beta-oxidation using acyl-coa dehydrogenase	16	0.28	19	2.97	15	0.33	19	3.42	9	0.77	19	7.48
GO:0042493	Response to drug	246	4.36	528	1.64	211	4.65	528	1.73	82	6.98	528	2.45
GO:0042542	Response to hydrogen peroxide	40	0.71	78	1.81	31	0.68	78	1.72	14	1.19	78	2.83
GO:0043065	Positive regulation of apoptotic process	133	2.36	338	1.39	115	2.53	338	1.47	53	4.51	338	2.47
GO:0043066	Negative regulation of apoptotic process	203	3.60	517	1.38	177	3.90	517	1.48	60	5.11	517	1.83
GO:0043434	Response to peptide hormone	63	1.12	129	1.72	51	1.12	129	1.71	19	1.62	129	2.32

(Continued)

TABLE 4 | Continued

GOID	GO BP Term	Microarray				RNA-seq				TempO-seq			
		Count	%	Pop Hits	FE	Count	%	Pop Hits	FE	Count	%	Pop Hits	FE
GO:0045471	Response to ethanol	82	1.45	193	1.50	74	1.63	193	1.66	34	2.89	193	2.78
GO:0046686	Response to cadmium ion	27	0.48	45	2.11	22	0.48	45	2.12	12	1.02	45	4.21
GO:0051289	Protein homotetramerization	37	0.66	76	1.72	34	0.75	76	1.94	20	1.70	76	4.15
GO:0051301	Cell division	83	1.47	179	1.63	63	1.39	179	1.52	27	2.30	179	2.38
GO:0051384	Response to glucocorticoid	69	1.22	133	1.83	56	1.23	133	1.82	23	1.96	133	2.73
GO:0051603	Proteolysis involved in cellular protein catabolic process	28	0.50	51	1.93	24	0.53	51	2.04	11	0.94	51	3.40
GO:0055088	Lipid homeostasis	27	0.48	42	2.27	24	0.53	42	2.47	11	0.94	42	4.13
GO:0055114	Oxidation-reduction process	314	5.56	651	1.70	262	5.77	651	1.74	88	7.49	651	2.13
GO:0070542	Response to fatty acid	26	0.46	39	2.35	25	0.55	39	2.77	11	0.94	39	4.45
GO:0071407	Cellular response to organic cyclic compound	53	0.94	122	1.53	48	1.06	122	1.70	22	1.87	122	2.85
GO:0071456	Cellular response to hypoxia	57	1.01	137	1.47	49	1.08	137	1.55	22	1.87	137	2.53
GO:0097421	Liver regeneration	41	0.73	55	2.63	30	0.66	55	2.36	14	1.19	55	4.02
GO:0098609	Cell-cell adhesion	102	1.81	211	1.70	82	1.81	211	1.68	27	2.30	211	2.02
GO:1904871	Positive regulation of protein localization to Cajal body	8	0.14	8	3.52	8	0.18	8	4.33	6	0.51	8	11.84
GO:1904874	Positive regulation of telomerase RNA localization to Cajal body	13	0.23	15	3.05	12	0.26	15	3.46	8	0.68	15	8.42

®*Universe is 17,535. List totals: Microarray: 4,976; RNA-Seq: 4,053; TempO-Seq: 1,111. Enrichment performed using UniGene cluster IDs and the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8.*

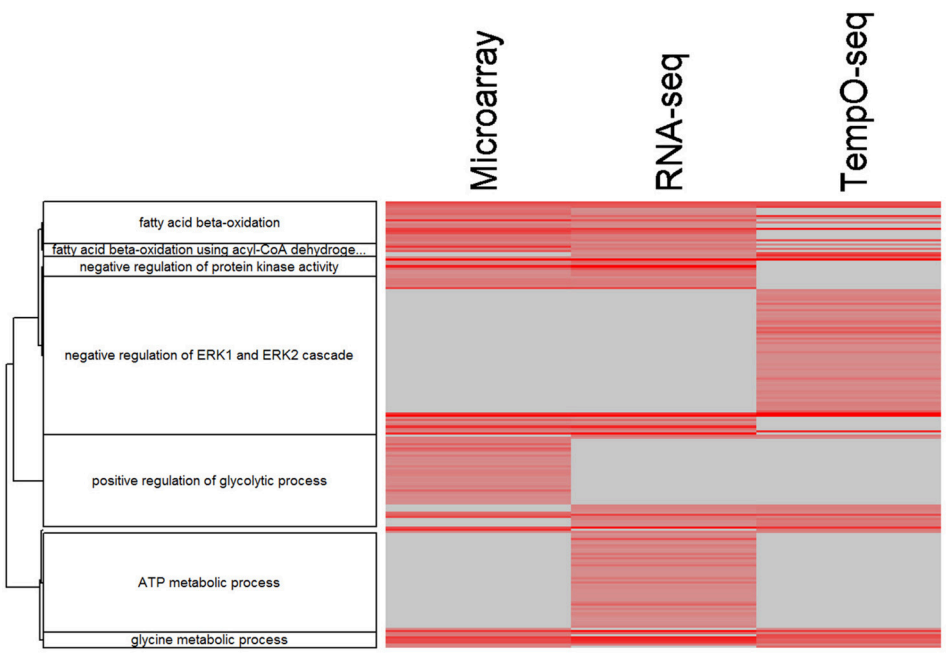


FIGURE 8 | Clustering of enriched GO BPs. goSTAG clustering of 203 enriched GO BPs using 5 genes per category, BH FDR < 0.05, correlation distance (1-Pearson correlation) and Ward clustering, dendrogram threshold = 0.9 and minimum number of GO BP terms per cluster = 5. Data is the $-\log_{10}$ *p*-value. The more red the intensity, the more significant the enrichment. Gray indicates that the GO BP term was not enriched significant and thus the *p*-value was imputed with 1.0.

sample lysates negating the need for library construction which are large sources of variability in RNA-Seq (Su et al., 2014), (2) the sequencing cost is much less than RNA-Seq given the number of targeted templates and pooling of samples in a multiplexed sequencing run allowing for more resources to expand experimental designs, and (3) the data storage is reasonable and the bioinformatics more simplified leading to a quicker turn-around in results and data analysis manageable by a wider group of analysts.

An obvious disadvantage to the TempO-Seq platform is that the gene content is predefined requiring extensive template fabrication and careful probe sequence curation. Furthermore, the targeted sequencing design renders TempO-Seq incapable of discerning novel transcripts. Despite these shortcomings, TempO-Seq performed similarly to microarray and RNA-Seq with respect to analysis of the SEQC/MAQC3 MOA toxicogenomics data. When statistical parameters were used to evaluate the three platforms, TempO-Seq had comparable variance structure related to chemical treatment, MOA and route of administration (**Figure 2**). Not surprisingly, we observed that RNA-Seq had higher unexplained variance (**Figure 2**), a larger noise component in expression patterns (**Figure 3**), and greater error between biological replicates (**Table 3**). This might be due to the large variation in lowly expressed genes that RNA-Seq detects at high sequencing depths. When the top percentile of expressed genes from RNA-Seq were used to evaluate expression differences between biological replicates, the error was much lower than when additional lower expressed genes were used (Wang et al., 2014). TempO-Seq variation, noise, and error in gene expression was moderate, falling between microarray and RNA-Seq.

Since each transcript profiling platform has different numbers of gene content and annotation, we explored the ability of each to cluster the samples by using the set of genes that vary statistically by MOA. We used an ANOVA model for each data set with chemical, MOA and route as the main effects. For microarray, RNA-Seq and TempO-Seq, 9,499 probe sets, 7,217 transcripts, and 1,366 genes were detected as significantly (FDR $<$ 0.01) varying, respectively (**Supplemental Table 1**). These MOA-varying genes and those mapped to 731 UniGene cluster IDs (**Supplemental Table 2**) as a common set were used for cluster analysis. In both cases the clustering of the samples by MOA for each platform was similar in that at most two chemicals from two MOAs were not clustered with their respective MOA chemicals (**Figures 5**, **6A**). In addition, the clustering of the samples by PCA with platform-specific MOA-varying genes mapped to the common UniGene set projected the samples into 3-dimensional space (**Figure 6B**) representative of the MOAs similar to the outcome when just MOA-varying genes from each platform were used (**Figure 4**). Hence, it is plausible that the TempO-Seq platform with the reduced gene content set is sufficient to resolve gene expression space elicited by a wide variety of chemical stressors with distinct MOAs. Utilization of the TempO-Seq platform for evaluation of chemicals using gene expression suggests that the platform may gain popularity in biomolecular screening efforts in the near future (Grimm et al., 2016; House et al., 2017).

It has been proven that reproducibility between gene expression is higher when the data are compared on the pathway level than the gene level (Guo et al., 2006; Fan et al., 2010; Wang et al., 2014). We enriched the MOA-varying UniGenes according to GO BPs and revealed that the reduced representation of genes on the TempO-Seq platform had a negligible effect on the overrepresentation (**Figure 7** and **Table 4**). This is in line with the bioinformatics process to select the S1500+ sentinel gene content on the platform using diversity and co-expression importance scores (Mav et al., 2018). These genes were selected to cover $>$90% of the biological pathway space represented by MSigDB (Subramanian et al., 2005). Yet each platform does appear to have enrichment of unique BPs as depicted in GO subtrees of overrepresented biological categories (**Figure 8**).

Having another tool for biologists to survey genome-wide gene expression is a luxury for scientific experimentation. With microarray fully matured and easy to analyze, and RNA-Seq flexible to interrogate complex transcriptional machinery, scientists have diverse platforms to investigate biological consequences that regulate gene expression genome-wide. The emerging TempO-Seq platform adds to the genomics tool chest and with comparable performance capabilities to its predecessors, will undoubtedly play a pivotal role in high-throughput screening efforts.

AUTHOR CONTRIBUTIONS

PB conceptualized the analysis strategy, performed the analyses, interpreted the results, and wrote parts of the paper. SA provided the samples that the data were generated from, interpreted the results, provided biological, toxicological context, and wrote parts of the paper. RP helped to interpret the results and provided biological, toxicological context.

ACKNOWLEDGMENTS

We thank Drs. Danielle Thierry-Mieg and Jean Thierry-Mieg for providing the aligned RNA-Seq data and for mapping the AceView transcripts to the Affymetrix microarray probe sets. The authors would like to thank Dr. B. Alex Merrick and Dr. Raja Jothi for their helpful suggestions and comments to improve the manuscript. This research was supported by the National Institute of Environmental Health Sciences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00485/full#supplementary-material

REFERENCES

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bates, D., Machler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. SerB 57*, 289–300.

Bennett, B. D., and Bushel, P. R. (2017). goSTAG: gene ontology subtrees to tag and annotate genes within a set. *Source Code Biol. Med.* 12:6. doi: 10.1186/s13029-017-0066-1

Chou, J. W., Zhou, T., Kaufmann, W. K., Paules, R. S., and Bushel, P. R. (2007). Extracting gene expression patterns and identifying co-expressed genes from microarray data reveals biologically responsive processes. *BMC Bioinformatics* 8:427. doi: 10.1186/1471-2105-8-427

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Fan, X., Lobenhofer, E. K., Chen, M., Shi, W., Huang, J., Luo, J., et al. (2010). Consistency of predictive signature genes and classifiers generated using different microarray platforms. *Pharmacog. J.* 10, 247–257. doi: 10.1038/tpj.2010.34

Gaidatzis, D., Lerch, A., Hahne, F., and Stadler, M. B. (2015). QuasR: quantification and annotation of short reads in R. *Bioinformatics* 31, 1130–1132. doi: 10.1093/bioinformatics/btu781

Gong, B., Wang, C., Su, Z., Hong, H., Thierry-Mieg, J., Thierry-Mieg, D., et al. (2014). Transcriptomic profiling of rat liver samples in a comprehensive study design by RNA-Seq. *Sci. Data* 1:140021. doi: 10.1038/sdata.2014.21

Grimm, F. A., Iwata, Y., Sirenko, O., Chappell, G. A., Wright, F. A., Reif, D. M., et al. (2016). A chemical-biological similarity-based grouping of complex substances as a prototype approach for evaluating chemical alternatives. *Green Chem.* 18, 4407–4419. doi: 10.1039/C6GC01147K

Guo, L., Lobenhofer, E. K., Wang, C., Shippy, R., Harris, S. C., Zhang, L., et al. (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.* 24, 1162–1169. doi: 10.1038/nbt1238

House, J. S., Grimm, F. A., Jima, D. D., Zhou, Y. H., Rusyn, I., and Wright, F. A. (2017). A pipeline for high-throughput concentration response modeling of gene expression for toxicogenomics. *Front. Genet.* 8:168. doi: 10.3389/fgene.2017.00168

Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., et al. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43, D921–D927. doi: 10.1093/nar/gku955

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31:e15. doi: 10.1093/nar/gng015

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25. doi: 10.1186/gb-2009-10-3-r25

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database, C. (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Li, J., Bushel, P. R., Chu, T. M., and Wolfinger, R. D. (2009). "Principal variance components analysis: estimating batch effects in microarray gene expression data," in *Batch Effects and Noise in Microarray Experiments: Sources and Solutions,* ed A. Scherer (West Sussex: John Wiley & Sons, Ltd.), 141–154.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLoS Comput. Biol.* 13:e1005457. doi: 10.1371/journal.pcbi.1005457

Mav, D., Shah, R. R., Howard, B. E., Auerbach, S. S., Bushel, P. R., Collins, J. B., et al. (2018). A hybrid gene selection approach to create the S1500+ targeted gene sets for use in high-throughput transcriptomics. *PLoS ONE* 13:e0191105. doi: 10.1371/journal.pone.0191105

Pontius, J. U., Wagner, L., and Schuler, G. D. (2002). "UniGene: a unified view of the transcriptome," in *The NCBI Handbook,* eds J. Mcentyre and L. Ostell (Bethesda, MD: NIH National Center for Biotechnology Information U.S.) 363–376.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Available online at: http://www.R-project.org

Su, Z., Łabaj, P. P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., et al. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* 32, 903–914. doi: 10.1038/nbt.2957

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A Next Generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 1437.e17–1452.e17. doi: 10.1016/j.cell.2017.10.049

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Thierry-Mieg, D., and Thierry-Mieg, J. (2006). AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* 7(Suppl. 1), S12.11–S12.14. doi: 10.1186/gb-2006-7-s1-s12

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., et al. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* 32, 926–932. doi: 10.1038/nbt.3001

Waters, M. D., Jackson, M., and Lea, I. (2010). Characterizing and predicting carcinogenicity and mode of action using conventional and toxicogenomics methods. *Mutat. Res.* 705, 184–200. doi: 10.1016/j.mrrev.2010.04.005

Yeakley, J. M., Shepard, P. J., Goyena, D. E., Vansteenhouse, H. C., Mccomb, J. D., and Seligmann, B. E. (2017). A trichostatin a expression signature identified by TempO-Seq targeted whole transcriptome profiling. *PLoS ONE* 12:e0178302. doi: 10.1371/journal.pone.0178302

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 November 2018
doi: 10.3389/fgene.2018.00516

Robust Co-clustering to Discover Toxicogenomic Biomarkers and Their Regulatory Doses of Chemical Compounds Using Logistic Probabilistic Hidden Variable Model

Mohammad Nazmol Hasan[1,2], *Md. Masud Rana*[1], *Anjuman Ara Begum*[1], *Moizur Rahman*[3] *and Md. Nurul Haque Mollah*[1]*

[1] *Bioinformatics Laboratory, Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh,* [2] *Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Bangladesh,* [3] *Department of Veterinary and Animal Sciences, University of Rajshahi, Rajshahi, Bangladesh*

OPEN ACCESS

Edited by:
Paul Jennings,
VU University Amsterdam, Netherlands

Reviewed by:
Olivier Taboureau,
Paris Diderot University, France
Concetta Ambrosino,
University of Sannio, Italy

***Correspondence:**
Md. Nurul Haque Mollah
mollah.stat.bio@ru.ac.bd

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Genetics

Received: *02 July 2018*
Accepted: *12 October 2018*
Published: *01 November 2018*

Citation:
Hasan MN, Rana MM, Begum AA, Rahman M and Mollah MNH (2018) Robust Co-clustering to Discover Toxicogenomic Biomarkers and Their Regulatory Doses of Chemical Compounds Using Logistic Probabilistic Hidden Variable Model. Front. Genet. 9:516. doi: 10.3389/fgene.2018.00516

Detection of biomarker genes and their regulatory doses of chemical compounds (DCCs) is one of the most important tasks in toxicogenomic studies as well as in drug design and development. There is an online computational platform "Toxygates" to identify biomarker genes and their regulatory DCCs by co-clustering approach. Nevertheless, the algorithm of that platform based on hierarchical clustering (HC) does not share gene-DCC two-way information simultaneously during co-clustering between genes and DCCs. Also it is sensitive to outlying observations. Thus, this platform may produce misleading results in some cases. The probabilistic hidden variable model (PHVM) is a more effective co-clustering approach that share two-way information simultaneously, but it is also sensitive to outlying observations. Therefore, in this paper we have proposed logistic probabilistic hidden variable model (LPHVM) for robust co-clustering between genes and DCCs, since gene expression data are often contaminated by outlying observations. We have investigated the performance of the proposed LPHVM co-clustering approach in a comparison with the conventional PHVM and Toxygates co-clustering approaches using simulated and real life TGP gene expression datasets, respectively. Simulation results show that the proposed method improved the performance over the conventional PHVM in presence of outliers; otherwise, it keeps equal performance. In the case of real life TGP data analysis, three DCCs (glibenclamide-low, perhexilline-low, and hexachlorobenzene-medium) for glutathione metabolism pathway dataset as well as two DCCs (acetaminophen-medium and methapyrilene-low) for PPAR signaling pathway dataset were incorrectly co-clustered by the Toxygates online platform, while only one DCC (hexachlorobenzene-low) for glutathione metabolism pathway was incorrectly co-clustered by the proposed LPHVM approach. Our findings from the real data analysis are also supported by the other findings in the literature.

Keywords: toxicogenomic biomarker, doses of chemical compounds (DCCs), co-clustering, outlying observations, logistic transformation, probabilistic hidden variable model (PHVM), logistic probabilistic hidden variable model (LPHVM)

INTRODUCTION

Toxicogenomics studies combines toxicology with several *omics* technologies (genomics, transcriptomics, proteomics, and metabolomics) to assess the risk of toxins (small molecules, peptides, or proteins) and chemical agents (drugs, gasoline, alcohol, pesticides, fuel oil, and cosmetics) in organism (NRC, 2007; Afshari et al., 2011). Through integration of these *omics* technologies with bioinformatics, toxicogenomics can be used to suggest the molecular mechanism of toxicity. This can reduce the cost in terms of time, labor, compound synthesis, and animal use which are main limitations of traditional toxicology work (Nuwaysir et al., 1999; Chen et al., 2012). In drug discovery and development, it is also necessary to assess the doses of chemical compounds (DCCs) toxicity administering these DCCs on individuals for measuring drugs' safety. This assessment can be done by toxicogenomic biomarkers those are upregulated or downregulated by the influence of a set of DCCs on individuals. These toxicogenomic biomarkers can be identified from the extensive gene-treatment expression dataset of target organs of individuals (Fielden et al., 2007; Uehara et al., 2008; Igarashi et al., 2015).

An online toxicogenomic data analysis platform "ToxDB" increases its predictive power based on the pathway level gene expression data (Hardt et al., 2016). It calculates the pathway scores for a chemical compound to identify significant biomarker genes using t-statistic from different pathways. Nevertheless, there is no facility in this platform to study another interesting problem of relationship between gene groups and DCCs groups asserted by Afshari et al. (2011). To address this problem another online platform "Toxygates" produces co-clusters between genes and DCCs using hierarchical clustering (HC) (Nyström-Persson et al., 2017). But HC does not use two-way (gene-DCC) information simultaneously for co-clustering and it is sensitive to outlying observations (García-Escudero et al., 2010). Probabilistic hidden variable model (PHVM) has been developed for co-clustering between words and documents in a text mining problem (Hofmann, 2001). It uses two-way (row-column) information simultaneously during co-clustering. It was also successfully used in detecting hidden patters of biological profiling datasets (Joung et al., 2006; Bicego et al., 2010). Therefore, PHVM would be more effective approach than HC for co-clustering between genes and DCCs which is also supported by Joung et al. (2006). However, the PHVM algorithm is sensitive to outlying observations of gene expression. These outlying observations often occur in the gene expression dataset due to several steps involve in the data generating processes from hybridization to image analysis including scratches or dust on the surface, imperfections in the glass or imperfections in the array production (Gottardo et al., 2006; Upton et al., 2009). The outliers in the dataset may arise following Tukey–Huber contamination model (THCM; Agostinelli et al., 2015) or independent contamination model (ICM; Alqallaf et al., 2009). To overcome the robustness problems of conventional PHVM approach an attempt is made to propose logistic PHVM approach called as LPHVM for robust co-clustering between genes and DCCs to discover toxicogenomic biomarkers and their regulatory DCCs.

METHODS AND MATERIALS

Let us consider a toxicogenomic experimental design as described in **Figure 1** that reflects Japanese Toxicogenomics Project (TGP) (Uehara, 2010) experiment for a single time point from which the toxygates (Nyström-Persson et al., 2013) data were collected. According to this design, gene expression data of both treatment and control group of animal samples are assumed to be generated. Then the fold change gene expression data for a single time point are computed from the treatment and control group of animals. It can measure the actual treatment (DCCs) effects on the genes. The fold change gene-expression value of a gene is defined as follows:

$$Y_{tlq} = log_2\left(\frac{x_{tlq}}{x'_{tlq}}\right) = log_2\left(x_{tlq}\right) - log_2\left(x'_{tlq}\right), \qquad (1)$$

where Y_{tlq} is the fold change expression value of a gene for the qth ($q = 1, 2, 3$) sample under lth (l = Low, Middle, High) dose level of the tth ($t = 1, 2, \cdots, T$) chemical compound, x_{tlq} is the expression value of that gene of mentioned sample under the treatment group and $x_{tlq}{}'$ is the expression value of the same gene of the respective control sample. The effect of compound-dose combination or treatment/DCCs on the animal can be measured by $\overline{Y}_{tl.}$ which is the average fold change value over the samples. In this paper, our objective is to robust co-clustering between genes and DCCs to discover toxicogenomic biomarkers and their regulatory DCCs from the fold change gene expression data using the proposed LPHVM.

Logistic Transformation of Fold Change Gene Expression Data

There are two ways to obtain robust estimates in presence of outlying observations (1) applying the robust methods (2) applying conventional methods on the modified dataset. The modification of the outlier contaminated dataset can be done deleting the outlying observations from the dataset or applying transformation on the dataset. Nonetheless, application of robust methods is complicated than using the conventional methods and deletion of outlying observations loses the information of the dataset. Hence, transformation is the better option for reducing outlier effects. Several authors (Box and Cox, 1964; Atkinson, 1982; Carroll, 1982) have been proved that transformation based robust methods outperform the conventional methods in reducing outlier effects. Thus, in this paper we consider logistic transformation for reducing outlier effects from the dataset. Before application of logistic transformation in the dataset we have taken average value ($\overline{Y}_{tl.}$) of the fold change gene expression (Y_{tlq}) over the samples. We denote this average value by $F\left(G_i, C_j\right)$ for the convenience of further use. In toxicogenomic data the expression profile of a subset of genes is highly correlated across a subset of conditions/treatments (Madeira and Oliveira, 2004; Bicego et al., 2010; Afshari et al., 2011). Interestingly, in the

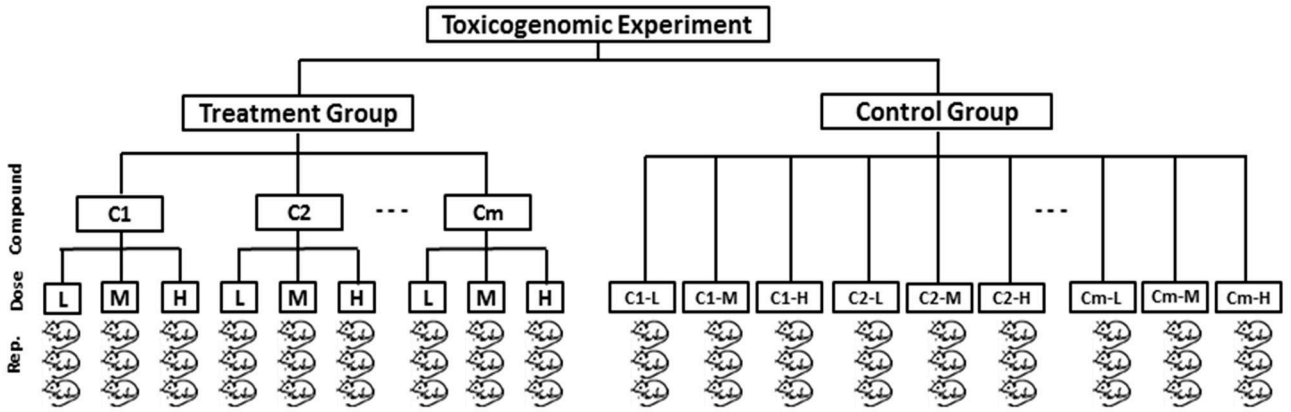


FIGURE 1 | A typical toxicogenomic experimental model for a single time point according to which gene expression data of the animal samples can be collected. In the figure there is a treatment group of animals and a control group of animals from which the fold change gene expression data can be obtained.

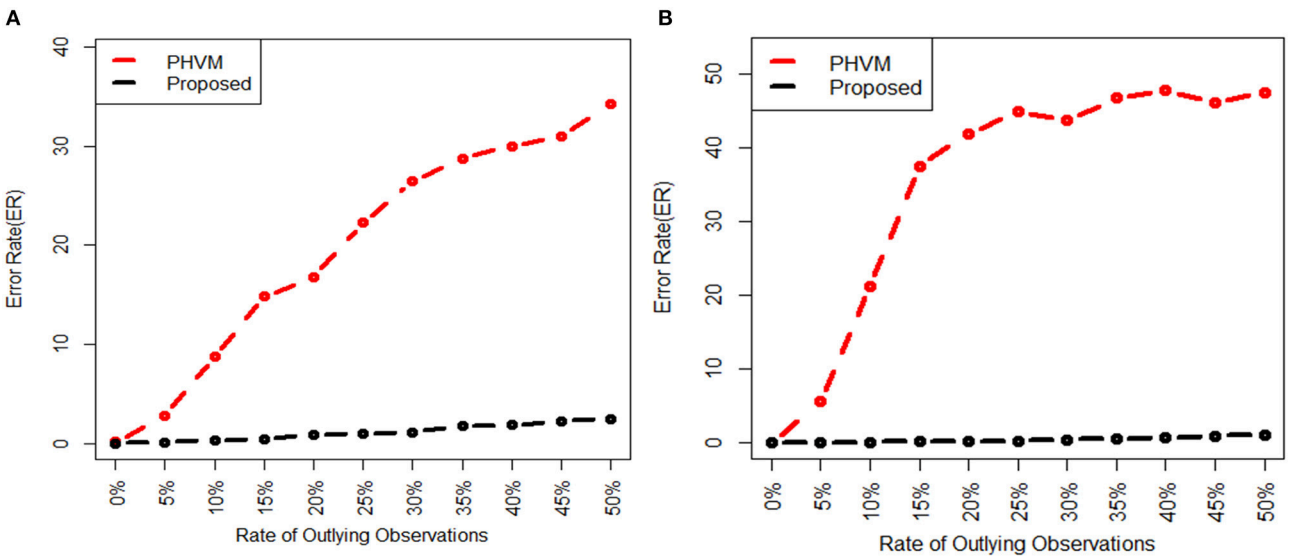


FIGURE 2 | Average gene and doses of chemical compounds co-clustering ER are plotted against the rate of outliers, when each of the data sets are simulated 100 times and outliers in the datasets are introduced using THCM. In the figure **(A)** for D_1 dataset and **(B)** for D_2 dataset.

gene expression or average fold change gene expression data there is a subset of genes which consists an upregulated and a downregulated clusters of genes which is highly correlated over a subset of DCCs. Therefore, we take absolute of the average fold change expression data to merge upregulated and downregulated clusters of genes into a single cluster/subset which are regulated by a subset of DCCs. Thereafter, the subset of genes forms a co-cluster with its regulatory subset of DCCs. Since in this study, we consider all the biomarker and non-biomarker genes (genes are not affected by DCCs) in a pathway, the non-biomarker genes make another co-cluster together with non-regulatory DCCs (which do not affect the expression patterns of the genes in a specific pathway). The term co-cluster refers to the clustering of correlated row (genes) and column (DCCs) simultaneously. Now we apply logistic transformation on the $(|F(G_i, C_j)|)$. If there are extreme values of $|F(G_i, C_j)|$ the logistic transformation bring them within the range of 0–1. The other transformation methods like Box-Cox family of power transformation returns unbounded value for the extreme one. The observed $n \times m$ (gene-DCCs) fold change gene expression data matrix consisting of $G = (G_1, G_2, \ldots, G_n)$ genes and $C = (C_1, C_2, \ldots, C_m)$ DCCs is transformed using logistic function

$$\#(G_i, C_j) = \left(\frac{1}{1 + \exp(-|F(G_i, C_j)|} \right) \times 100$$

Similar to other works (Joung et al., 2006; Bicego et al., 2010) we assume the transformed value $\#(G_i, C_j)$ as the count value for applying PHVM.

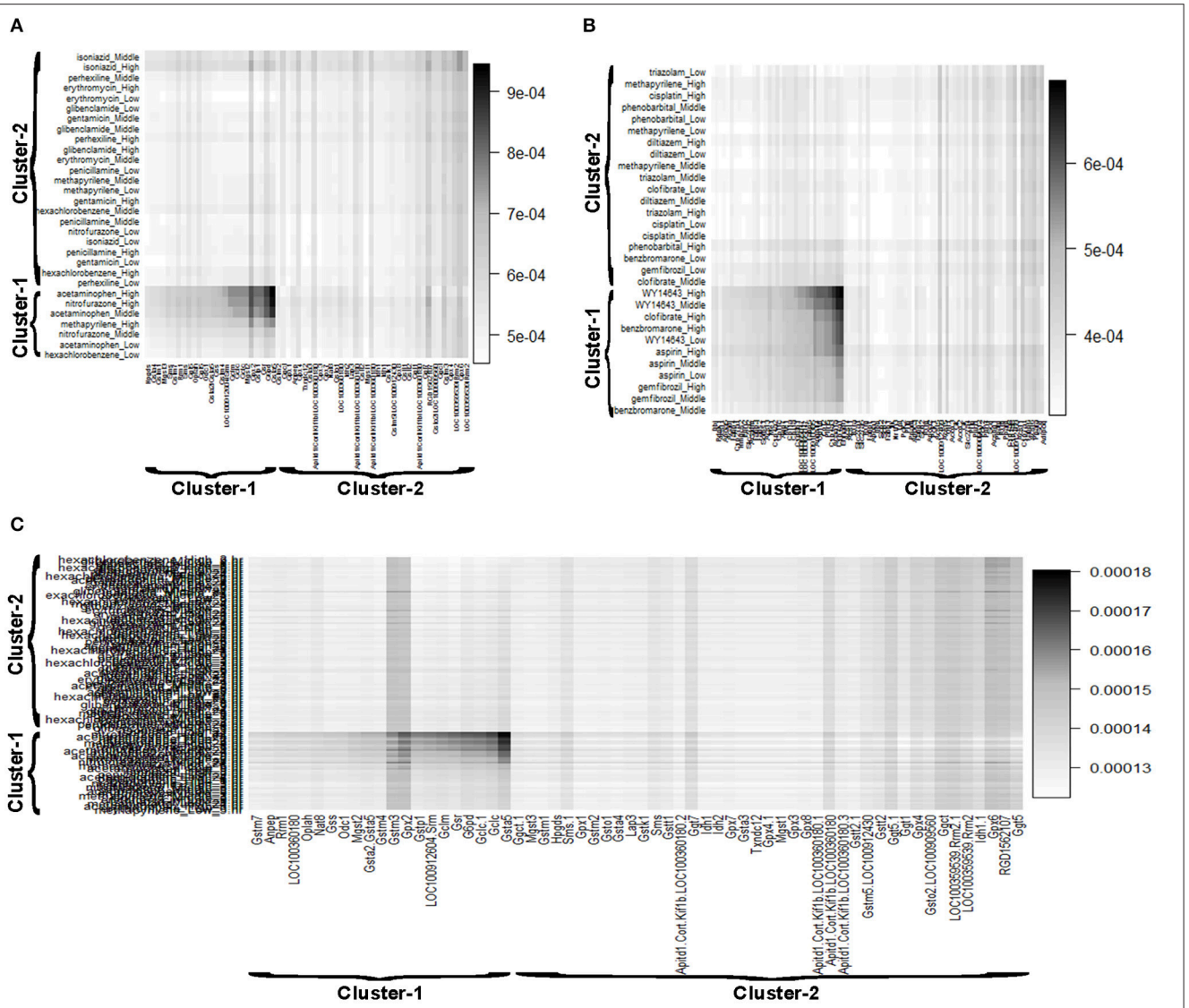


FIGURE 3 | Gene and doses of chemical compounds co-clustered view retrieved from the LPHVM generated gene and DCCs joint probability. **(A)** Represents glutathione metabolism pathway dataset at 24 h time point. **(B)** Represents PPAR signaling pathway dataset at 24 h time point. **(C)** Represents glutathione metabolism pathway dataset for all time points of Toxygates data.

Number of Co-clusters (k) Prediction

As we see from the previous section "logistic transformation of fold change gene expression data" in toxicogenomic dataset there are hidden patters or co-clusters between genes and DCCs. Thus the number of clusters in the DCCs is equal to the number of clusters in the genes. Before applying PHVM it is required to know the number of co-clusters in the dataset. Therefore, in this study, we consider gap statistic (Tibshirani et al., 2001) the most popular and reliable algorithm for predicting the number of co-clusters in the dataset. We use R function "fviz_nbclust" which required packages "factoextra" and "NbClust" (Malika et al., 2014) in order to predict number of co-clusters in the dataset via gap statistic. The detail algorithm of gap statistic is given in the **Supplementary Material**.

Robust Co-clustering Using Logistic Probabilistic Hidden Variable Model

In order to perform robust co-clustering between genes and DCCs we propose LPHVM approach. We define LPHVM as the application of PHVM on the count valued dataset which is obtained transforming absolute value of the fold change gene expression data by logistic transformation. For this standpoint, let us consider $n \times m$ gene-DCC count valued fold change gene expression data matrix consisting of $G = (G_1, G_2, \ldots, G_n)$ genes and $C = (C_1, C_2, \ldots, C_m)$ DCCs. LPHVM assumes that there prevail a certain number of unobserved hidden co-clusters or clusters underlying the gene-DCC count valued data matrix. We have estimated the number of co-clusters (k) in the dataset using gap statistic algorithm proposed by Tibshirani et al. (2001).

Introducing the hidden variable H = (H_1, H_2, …, H_k; r = 1, 2, …, k) the model quantifies the relationships $Pr(G_i|H_r)$, $Pr(C_j|H_r)$, and $Pr(G_i, C_j)$. The following are the probability definition and underlying assumptions of LPHVM accordingly: (1) $Pr(H_r)$ is the probability of the rth co-cluster/cluster and $\sum_{r=1}^{k} Pr(H_r) = 1$. (2) $Pr(G_i|H_r)$ is the probability of the ith gene over the rth co-cluster and $\forall H_r$; $\sum_{i=1}^{n} Pr(G_i|H_r) = 1$. (3) $Pr(C_j|H_r)$ is the probability of the jth DCC over the rth co-cluster and $\forall H_r$; $\sum_{j=1}^{m} Pr(C_j|H_r) = 1$. (4) $Pr(G_i, C_j)$ is the joint probability of the ith gene and the jth DCC and $\sum_{i=1}^{n}\sum_{j=1}^{m} Pr(G_i, C_j) = 1$. Based on these definition and assumptions we obtain the joint probability of the gene-DCC observed pair (G_i, C_j) considering hidden co-cluster H_r as follows:

$$Pr(G_i, C_j) = Pr(C_j)\, Pr(G_i|C_j)$$

Where,

$$Pr(G_i|C_j) = \sum_{r=1}^{k} Pr(G_i|H_r) Pr(H_r|C_j)$$

Applying Bayes' rule, the gene-DCC joint probability $Pr(G_i, C_j)$ can be written as

$$Pr(G_i, C_j) = \sum_{r=1}^{k} Pr(G_i|H_r)\, Pr(C_j|H_r)\, Pr(H_r)$$

So as to estimate the parameters of the model, we need to maximize the total likelihood of the observations:

$$L(G, C) = \sum_{i=1}^{n}\sum_{j=1}^{m} \#(G_i, C_j)\ log\ Pr(G_i, C_j)$$

We have applied the widely used Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for estimating the maximum likelihood parameters of the proposed model. The EM algorithm starts with a random set of initial parameter values and iterates both the expectation (E-step) and maximization (M-step) step alternatively until a certain convergence criteria is satisfied. For this study, we have taken the values of initial parameters from dirichlet distribution and the stopping condition for EM estimation was set to <0.00001 (difference between two log likelihood of successive EM iteration). The E and M-step for the total likelihood can be given as follows:

E-step:

$$Pr(H_r|G_i, C_j) = \frac{Pr(G_i|H_r)\, Pr(C_j|H_r)\, Pr(H_r)}{\sum_{r'=1}^{k} Pr(G_i|H_{r'})\, Pr(C_j|H_{r'})\, Pr(H_{r'})}$$

M-step:

$$Pr(H_r) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \#(G_i, C_j) Pr(H_r|G_i, C_j)}{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{r'=1}^{k} \#(G_i, C_j) Pr(H_{r'}|G_i, C_j)}$$

$$Pr(G_i|H_r) = \frac{\sum_{j=1}^{m} \#(G_i, C_j) Pr(H_r|G_i, C_j)}{\sum_{i'=1}^{n}\sum_{j=1}^{m} \#(G_{i'}, C_j) Pr(H_r|G_{i'}, C_j)}$$

$$Pr(C_j|H_r) = \frac{\sum_{i=1}^{n} \#(G_i, C_j) Pr(H_r|G_i, C_j)}{\sum_{i=1}^{n}\sum_{j'=1}^{m} \#(G_i, C_{j'}) Pr(H_r|G_i, C_{j'})}$$

Once the parameters $Pr(G_i|H_r)$ and $Pr(C_j|H_r)$ have been estimated the genes and DCCs are clustered independently and co-clustered simultaneously. The gene (G_i) and DCC (C_j) will belong to co-cluster r if

$$Pr(G_i|H_r) = argmax_{r'} Pr(G_i|H_{r'})\,;\, i = 1, 2, \cdots, n; r = 1, 2, \ldots, k \text{ and}$$
$$Pr(C_j|H_r) = argmax_{r'} Pr(C_j|H_{r'})\,;\, j = 1, 2, \ldots, m; r = 1, 2, \ldots, k$$

At the same time, if the gene (G_i) and the DCC (C_j) is grouped into a co-cluster (r) and this pair has the highest joint probability $Pr(G_i, C_j)$ in that co-cluster (**Figure 3**).

Extraction of Toxicogenomic Biomarker Genes and Their Regulatory Doses of Chemical Compounds

As described in section "logistic transformation of gene expression data" the biomarker genes form co-clusters with their respective regulatory DCCs. Additionally, the non-biomarker genes in a pathway form another co-cluster with non-regulatory DCCs. The LPHVM grouped the genes and DCCs simultaneously to their respective co-clusters. Zhu et al. (2005) has shown that the PHVM generated co-occurrence probabilities between correlated genes and chemical compounds which co-occur more frequently are higher than others. Biological relationship among these correlated genes and chemical compounds is also stronger. Therefore, we ranked the co-clusters based on the average LPHVM generated joint probability $(Pr(G_i, C_j))$ of gene-DCC within the co-clusters. The co-cluster having largest average joint probability contains most important biomarker genes and their regulatory DCCs and so on. The non-biomarker genes and non-regulatory DCCs in a dataset of a particular pathway are filtered in a co-cluster by LPHVM which have the smallest average joint probability. Except this co-cluster (co-cluster having smallest average joint probability) others are the co-clusters of biomarker genes and their regulatory DCCs and we define these co-clusters as biomarker co-clusters. We extract the toxicogenomic biomarker genes and their regulatory DCCs from these biomarker co-clusters.

Up/Down-Regulated Biomarker Genes and Ranking of Doses of Chemical Compounds

The biomarker co-clusters consisting of biomarker genes and their regulatory DCCs are separated from the whole gene-DCC fold change data matrix which is discussed in the previous section. Within this co-clustering matrix a subset of biomarker genes may be upregulated corresponding to a subset of DCCs or downregulated corresponding to another subset of DCCs. These can be observed from the average fold change value ($\overline{Y}_{tl.}$) of the co-clustering matrix. For example, a biomarker is define as up or down-regulated gene corresponding to the l^{th} dose level of the t^{th} chemical compounds if $\overline{Y}_{tl.} > 0$ or $\overline{Y}_{tl.} < 0$. Then this dose of chemical compound is said to be a regulatory DCC. Furthermore, for ranking the biomarker

gene regulatory DCCs and their relationships with biomarker genes we have separated a sub matrix of biomarker genes and their regulatory DCCs (biomarker co-clusters) from the LPHVM generated gene-DCC joint probability $Pr\left(G_i, C_j\right)$ matrix. The biomarker gene regulatory DCCs are ranked according to their average joint probability value over all biomarkers. We also rank the relationships among biomarker genes and their regulatory DCCs based on their joint probability. The ranking is made considering the formula:

$$\left(\frac{Z_{j/i,j}}{\max\left(Z_{j/i,j}\right)}\right) \times 100$$

where Z_j is the average joint probability of a DCC over the biomarkers or $Z_{i,j}$ is the joint probability of gene G_i and DCC C_j within the biomarker co-clusters.

Robustness of the Proposed Algorithm

We investigate the robustness of the proposed (LPHVM) algorithm and conventional PHVM using simulated datasets in absence and presence of outliers in the dataset based on the co-clustering /clustering error rate (ER). The genes and DCCs which are considered in one co-cluster/cluster in the simulated data are incorrectly assigned in another co-cluster/cluster by the PHVM or LPHVM is considered as the miss co-clustered/clustered observations. The ER is the percentage of miss co-clustered/clustered observations which is calculated as:

$$\left(\frac{tolal\ miss\ co-clustered/clustered\ observations}{Total\ observations}\right) \times 100$$

Computational Steps of LPHVM at a Glance

For detecting the toxicogenomic biomarker genes and their regulatory DCCs from the pathway level toxicogenomic dataset using LPHVM the following steps are to be considered for desired outputs:

Step 1: Obtain gene expression data of treatment and control group of animals from the toxicogenomic experiment (**Figure 1**). Thereafter, compute fold change gene expression data using Equation (1) and then make it absolute.

Step 2: Apply logistic transformation on the dataset obtain from step 1 and assume the transformed value as count value.

Step 3: Estimate the number of co-clusters in the dataset which is obtained from step 2.

Step 4: Obtain robust co-clusters applying PHVM on the dataset obtained from step 2 using the number of co-clusters which we get from step 3.

Step 5: Calculate average joint probability of gene-DCC within the co-clusters and ranked them.

Step 6: Separate the co-clusters of biomarker genes and their regulatory DCCs from the co-cluster which have smallest average joint probability of gene-DCC.

Step 7: The genes and DCCs in the separated co-clusters which we get from step 6 are the toxicogenomic biomarkers and their regulatory DCCs.

Step 8: A biomarker gene obtains from step 7 may be upregulated corresponding to a DCC or downregulated corresponding to another DCC. A biomarker gene is said to be a up or down-regulated if its average fold change value corresponding to the lth dose level of the tth chemical compound is $\overline{Y}_{tl.} > 0$ or $\overline{Y}_{tl.} < 0$.

Simulated Datasets

To investigate the performance of the proposed LPHVM algorithm over the conventional PHVM we have simulated two sets of pathway level fold change gene expression data $D_1(n = 50 \times m = 30)$ and $D_2(n = 50 \times m = 60)$ imitating the toxicogenomic experiment given in **Figure 1**. Alongside these a pathway level dataset considering all time points of toxygates data are analyzed in the real data section. According to this experiment the fold change gene expression data (Y_{tlq}) have been generated using the following model:

$$Y_{tlq} = \begin{array}{lccc} & \textbf{DCCs group-1} & \textbf{DCCs group-2} & \textbf{DCCs group-3} & \\ \text{Gene group-11} & +F11 & 0 & 0 & \\ \text{Gene group-12} & -F12 & 0 & 0 & \\ \text{Gene group-21} & 0 & +F21 & 0 & +N(0,\sigma^2) \\ \text{Gene group-22} & 0 & -F22 & 0 & \\ \text{Gene group-3} & 0 & 0 & 0 & \end{array} \quad (2)$$

In the above model, $+F11$and $+F21$ represent the fold change expression values for upregulated genes under the DCCs group 1 and 2, respectively. Similarly, $-F12$, and $-F22$ represent the fold change expression values for the downregulated genes under the DCCs group 1 and 2, respectively. The 0s represent there is no compound effects on the respective gene group and $N(0, \sigma^2)$ represents the random error term generated from normal distribution with mean 0 and variance σ^2. Now if we take absolute value of the fold change gene expression data generated from the above data generating model (2), the fold change gene expression data $+F11$ and $-F12$ will merge into a single gene group-1 and make a co-cluster with their correlated DCCs group-1. Accordingly, $+F21$ and $-F22$ will merge into a single gene group-2 and make a co-cluster with their correlated DCCs group-2. The rest of the genes which are not regulated by any DCCs make a gene group-3 and the DCCs that do not regulate the expression pattern of genes make a DCCs group-3. The gene group-3 and DCCs group-3 together will make another co-cluster. These co-clusters can be retrieved by the LPHVM. In the simulated datasets n represents the number of genes (G_i; $i = 1, 2, \ldots, n$) and m represents the number of DCCs (C_j; $j = 1, 2, \ldots, m$). The data generation procedures for D_1 and D_2 datasets are given in the **Supplementary Material.**

Real Datasets

Several studies proved that molecular network or pathway based analysis improved the predictive power of gene expression data (Yildirimman et al., 2011; Hofree et al., 2013). Hardt et al. (2016) also analyzed the pathway level data from *in vitro* and *in vivo* experiment of human and rat model. Presently, pathway based analysis in cancer research has also advanced promptly since pathway level analysis able to produce more stable biomarkers (Kim, 2017). Since performance of any method cannot be measured without known dataset. Besides the simulation study, to investigate the performance of the proposed method compare to other existing methods we use two known datasets of glutathione metabolism and PPAR signaling pathways. The fold change expression data of the TGP experiment for glutathione metabolism and PPAR signaling pathway for some selected DCCs of the respective pathway at 24 h time point have been downloaded from toxygates (https://toxygates.nibiohn.go.jp/toxygates/#columns). Because the compounds' toxicity at 24 h time point is more visible compare to other time points (Nyström-Persson et al., 2013). Alongside these a dataset consisting of glutathione metabolism pathway genes and glutathione depleting and non-glutathione depleting compounds (Nyström-Persson et al., 2013) for all time points is also considered for analysis to know about the toxicity of DCCs in other time points.

RESULTS

Simulation Study

We investigate the performance of our proposed method (LPHVM) by comparing it with the conventional PHVM using simulated datasets D_1 and D_2 in absence and presence of outlying observations for robust co-clustering between genes and DCCs to discover biomarker genes and their regulatory DCCs. The number of co-clusters/clusters for both of the simulated datasets is estimated as 3 via gap statistic as per the datasets are simulated (**Figure S1**). For calculating average co-clustering and clustering ER we have simulated each of the datasets 100 times. Every time of data simulation outliers are introduced in the dataset using the data contamination methods THCM and ICM at the same time ER are calculated for PHVM and LPHVM applying these methods on the datasets. The description of the data contamination by outliers, THCM and ICM are given in the **Supplementary Material**. Here it should be mentioned that in the case of THCM we have contaminated the simulated datasets by 5–50% rate of outliers. Similarly, in the case of ICM we have considered the range of probability of at least one component of the dataset is to be contaminated is 0.14–0.60 for D_1 dataset and 0.165–0.5962 for D_2 dataset. **Figure 2** visualizes the average co-clustering ER between genes and DCCs for datasets D_1 and D_2 in absence and presence of outliers when the datasets are contaminated by outliers using the THCM. The **Table 1** shows the average co-clustering ER between genes and DCCs in absence and presence of outliers for the simulated datasets D_1 and D_2 when the datasets are contaminated by outliers using ICM. **Figure S2** and **Table S1** in the Supplementary Material show the average clustering ER for gene and DCCs. It is observed from the mentioned figures and tables that in absence of outlier both of the proposed LPHVM and conventional PHVM approaches produce 0 ER. However, in presence of outlaying observations in the datasets the proposed approach produce far smaller ER than the conventional approach for both of the data contamination methods (THCM and ICM). The simulated data structure, structure of the data when row (gene) and column (DCCs) entities are randomly allocated and proposed method recovered structure of the data are given in the Supplementary Material (**Figures S3**, **S4)** for the datasets D_1 and D_2. From these figures it is observed that the proposed algorithm is efficient for co-clustering between genes and DCCs of the pathway level fold change gene expression data. **Figure S3C** represents the dataset D_1 where all the genes and DCCs are grouped into three co-clusters (co-clusters 1, 2, and 3) and within co-cluster average joint probability of gene-DCC are given in **Table 3**. From where it is found that co-cluster-1 produces the smallest average joint probability of gene-DCC. Therefore, co-cluster 2 and 3 are the co-cluster of biomarker genes and their regulatory DCCs for the dataset D_1. Similarly, for D_2 dataset co-cluster-3 produces the smallest average joint probability of gene-DCC (**Table 3**). Thus, co-cluster 1 and 2 are the biomarker co-clusters consisting of biomarker genes and their regulatory DCCs. The biomarker genes and their regulatory DCCs that we get from the biomarker co-clusters of the simulated datasets are given in the **Table S9**. Ranking of the biomarker regulatory DCCs are performed based on the biomarker gene-DCC joint probability matrix of biomarker co-clusters following the raking method described in sub section (Up/Down-regulated Biomarker Genes and Ranking of Doses of Chemical Compounds). The results are given in

TABLE 1 | Average values of the gene and doses of chemical compounds co-clustering ER for the simulated datasets D_1 and D_2 when each of the datasets are simulated 100 times and contaminated by outlier using ICM.

Dataset	Method	Probability of at least one component in the dataset to be contaminated (ε)						
		0.00	**0.14**	**0.26**	**0.36**	**0.45**	**0.53**	**0.60**
D_1	PHVM	0.175	24.675	28.950	32.912	33.500	35.125	38.487
	Proposed	0.025	0.387	0.612	0.725	1.0	1.862	2.500
		0.00	**0.165**	**0.3031**	**0.4187**	**0.5154**	**0.5962**	
D_2	PHVM	0.00	25.390	26.563	29.554	32.172	39.754	
	Proposed	0.00	0.163	0.945	1.481	1.600	2.072	

TABLE 2 | Upregulated and downregulated biomarker genes and their regulatory doses of chemical compounds for real life datasets.

Dataset	Biomarker genes	Biomarker gene regulatory DCCs
Glutathione metabolism pathway	Gsta4, Gstm1, Sms, Rrm1, Odc1, Gsta2/Gsta5, Gss, Gstm4, LOC100912604/Srm, Gclm, Gclc, Mgst2, Gstp1, Gsr, Gpx2, G6pd, Gsta5, Hpgds, Mgst3, Gstm7, Oplah, Ggt5	hexachlorobenzene_Low acetaminophen_Low nitrofurazone_Middle methapyrilene_High acetaminophen_Middle nitrofurazone_High acetaminophen_High
PPAR signaling pathway	Dbi, Acsl1, Acadl, Hmgcs2, Plin2, Slc27a2, Acadm, Fads2, Fabp3, Me1, Sorbs1, Acsl3, Cyp4a2, Aqp7, Cpt1a, Cyp8b1, OC100365047, LOC100910385, Angptl4, Cpt1b, Cpt2, Plin5, Cyp4a3, Acaa1a, Cyp4a1, Ehhadh, Pdpk1, Apoa5, Fabp4, Cyp27a1, Cpt1c, Fabp5	benzbromarone_Middle gemfibrozil_Middle gemfibrozil_High aspirin_Low aspirin_Middle aspirin_High WY14643_Low benzbromarone_High clofibrate_High WY14643_Middle WY14643_High

TABLE 3 | Average values of the Gene and DCCs joint probabilities within the co-clusters generated by the proposed LPHVM algorithm for the simulated and real life datasets.

Dataset	Co-cluster-1	Co-cluster-2	Cocluster-3
D_1	0.0006095721	0.0010120670	0.0010117088
D_2	0.0005162618	0.0005163485	0.0003147069
Glutathione metabolism pathway	0.0006196723	0.0005331547	
PPAR signaling pathway	0.0004471087	0.0003704091	

TABLE 4 | Biomarker genes regulatory doses of chemical compounds ranking for real datasets (glutathione metabolism and PPAR signaling pathway).

Dataset	Doses of chemical compounds	Percent score
Glutathione metabolism pathway	acetaminophen_High	100.00
	nitrofurazone_High	99.59
	acetaminophen_Middle	95.98
	methapyrilene_High	88.66
	nitrofurazone_Middle	82.24
	acetaminophen_Low	77.84
	hexachlorobenzene_Low	74.57
PPAR signaling pathway	WY14643_High	100.00
	WY14643_Middle	97.59
	clofibrate_High	93.25
	aspirin_High	92.91
	benzbromarone_High	92.25
	WY14643_Low	91.19
	aspirin_Middle	87.93
	aspirin_Low	86.41
	gemfibrozil_High	85.51
	gemfibrozil_Middle	84.52
	benzbromarone_Middle	79.07

the Supplementary Material (**Table S10**) for both D_1 and D_2 datasets.

Analysis of Glutathione Metabolism Pathway Data

Reactive oxygen species (ROS) are produced by living organisms as a normal product as a result of normal cellular metabolism. However, in presence of environmental pollutants or toxic chemical the production of ROS increased dramatically. It is highly reactive molecules and can damage cell structures such as carbohydrates, nucleic acids, lipids, and proteins and alter their functions. In the liver, glutathione is an important antioxidant; a major detoxification player which scavenges ROS. Thus imbalance in the abundance of ROS and glutathione/antioxidant in favor of ROS in the liver in presence of toxic chemicals/drugs causes' drug induced liver injury. Subsequently, gene expression changes occur simultaneously in response to the glutathione depletion or after the glutathione depletion (Gao et al., 2010; Birben et al., 2012; Nyström-Persson et al., 2013). In order to identify glutathione depletion related biomarker genes and their regulatory DCCs as well as to investigate the performance of the proposed LPHVM approach we use known fold change gene expression dataset of glutathione metabolism pathway. The fold change gene expression dataset consists 62 glutathione metabolism pathway genes, three glutathione depleting compounds (acetaminophen, methapyrilene, and nitrofurazone) and seven non-glutathione depleting compounds (erythromycin, hexachlorobenzene, isoniazid, gentamicin, glibenclamide, penicillamine, and perhexilline) (Nyström-Persson et al., 2013) along with the dose levels (low, middle, and high) for 24 h time point. The number of co-clusters which is required in applying LPHVM for this dataset is estimated as 2 (**Figure S1**) via gap statistic. **Figure 3A** shows actual co-clusters in the glutathione metabolism pathway dataset. The genes and DCCs in the co-clusters are given in the **Table S2**. The average joint probabilities of gene-DCC within the co-clusters are 0.0006196723 and 0.0005331547 (**Table 3**), respectively for co-cluster-1 and co-cluster-2. Thus, Co-cluster-1 is the co-cluster of biomarker genes and glutathione depleting DCCs as it produces highest average joint probability. The biomarker genes and their regulatory DCCs in co-cluster-1 are given in **Table 2**. Additionally, the upregulated and downregulated biomarker genes corresponding to their regulatory DCCs are presented in the **Figure S7A**. For the same dataset the clustering results (heatmap) produced by toxygates are given in **Figure S5** where glibenclamide-low, perhexilline-low, and hexachlorobenzene-medium dose level are incorrectly co-clustered whereas only hexachlorobenzene-low dose is incorrectly co-clustered by the proposed LPHVM approach according to Nyström-Persson et al. (2013). The biomarker genes in co-cluster-1 are functionally annotated by the online database DAVID (Huang da et al., 2009) and the results are given in the **Tables S5, S6**. The results show that the biomarker genes are significant in different biological

TABLE 5 | Top 20 (ranked) biomarker gene and their regulatory doses of chemical compound relationships for glutathione metabolism pathway and PPAR signaling pathway datasets.

Glutathione metabolism pathway			PPAR signaling pathway		
Chemical compound and dose combination	**Biomarker gene**	**Ranking score**	**Chemical compound and dose combination**	**Biomarker gene**	**Ranking score**
acetaminophen_High	Gsta5	100.00	WY14643_High	Ehhadh	100.00
nitrofurazone_High	Gsta5	96.26	WY14643_High	Cyp4a1	97.29
acetaminophen_Middle	Gsta5	91.69	WY14643_Middle	Ehhadh	95.32
acetaminophen_High	G6pd	90.85	WY14643_Middle	Cyp4a1	93.17
acetaminophen_High	Gpx2	89.67	WY14643_High	Acaa1a	92.41
nitrofurazone_High	G6pd	89.48	clofibrate_High	Ehhadh	88.93
nitrofurazone_High	Gpx2	89.29	WY14643_Middle	Acaa1a	88.47
acetaminophen_Middle	Gpx2	86.05	clofibrate_High	Cyp4a1	87.34
acetaminophen_Middle	G6pd	85.91	benzbromarone_High	Ehhadh	87.04
acetaminophen_High	Gsr	85.19	WY14643_High	Cyp4a3	86.68
acetaminophen_High	Gstp1	83.54	WY14643_Low	Ehhadh	86.65
nitrofurazone_High	Gsr	83.25	WY14643_High	Plin5	85.99
nitrofurazone_High	Gstp1	81.53	benzbromarone_High	Cyp4a1	85.67
acetaminophen_High	Mgst2	80.46	WY14643_Low	Cyp4a1	85.17
acetaminophen_High	Gclc	80.38	WY14643_High	Cpt2	84.46
methapyrilene_High	Gsta5	80.23	WY14643_High	Cpt1b	84.45
acetaminophen_Middle	Gsr	79.71	WY14643_High	Angptl4	83.99
acetaminophen_High	Gclm	79.56	aspirin_High	Ehhadh	83.60
methapyrilene_High	Gpx2	79.47	WY14643_Middle	Cyp4a3	83.54
nitrofurazone_High	Gclc	78.93	aspirin_High	Cyp4a1	83.10

functions or processes including glutathione metabolism pathway. Ranking of biomarker gene regulatory DCCs and top 20 gene-DCCs relationship along with their ranking score for glutathione metabolism pathway dataset are given in **Tables 4**, **5**. From the tables it is observed that acetaminophen_High, nitrofurazone_High, and acetaminophen_Middle dose etc. are the most important glutathione depleting compounds and Gsta5, G6pd, Gpx2, Gsr, Mgst2, Gstp1, Gclc etc. are the most important biomarker genes. The detail ranked relationships results are given in **Table S12**. Besides this we have analyzed the same dataset considering all time points (3, 6, 9, and 24 h) by LPHVM to know about toxicity mechanism of the glutathione depleting compounds in other time points. The co-clusters produced by LPHVM are given in **Figure 3C**. The detail analyzed results of this dataset are given in **Tables S4**, **S11**. The proposed LPHVM identified 25 genes for the dataset at 24 h time points and 21 genes for the dataset where all time points are considered as biomarker in the glutathione metabolism pathway among which 18 are common.

Analysis of PPAR Signaling Pathway Data

Peroxisome proliferator-activated receptors (PPARs) $PPAR_{\alpha}$, $PPAR_{\beta/\delta}$, and $PPAR_{\gamma}$ are transcription factors which are activated by ligand/drug. They regulate the expression of target genes in response to endogenous and exogenous ligands/chemicals. The PPAR ligands may produce toxicity via receptor-dependent and/or off-target-mediated mechanism(s) (Peraza et al., 2006). To discover PPARs regulated biomarker genes and their regulatory DCCs as well as to investigate the performance of the proposed LPHVM approach we consider known dataset consisting 88 PPAR signaling pathway genes and PPARs related gene regulatory compounds (WY-14643, clofibrate, gemfibrozil, benzbromarone, and aspirin) (Kiyosawa et al., 2006) and some other randomly selected compounds (cisplatin, diltiazem, methapyrilene, phenobarbital, and triazolam) along with their dose levels low, middle and high. The number of hidden co-clusters for this dataset is 2 estimated via gap statistic (**Figure S1**). The LPHVM generates co-clusters of the PPAR signaling pathway dataset which is shown in **Figure 3B**. The average joint probabilities of gene-DCC within co-clusters are 0.0004471087 and 0.0003704091 where co-cluster-1 has the larger value than the co-cluster-2. Therefore, co-cluster-1 is the biomarker co-cluster of biomarker genes and their regulatory DCCs. The non-regulated genes and non-regulatory DCCs consist in co-cluster-2. The detail co-clustering results are given in the **Table S3**. The biomarker genes and their regulatory DCCs in co-cluster-1 are given in **Table 2**. Additionally, up/down-regulated biomarker genes corresponding to their regulatory DCCs are depicted in the **Figure S7B** For the same dataset the toxygates co-clustering result using HC given in **Figure S6** which shows that acetaminophen-middle and methapyrilene-low are incorrectly co-clustered whereas our proposed method properly co-cluster the DCCs (**Table 2**) according to the statement of Kiyosawa et al. (2006). Biomarker genes in co-cluster-1 are functionally

annotated via DAVID the results are given in the **Tables S7**, **S8**. WY14643-High, WY14643-Middle and clofibrate-High are the top most DCCs for regulating PPARs related biomarker genes for detail see **Table 4**. Top 20 (ranked) relationships between biomarker genes and their regulatory DCCs are given in **Table 5** from where it is observed that Ehhadh, Cyp4a1, Acaa1a, Plin5 etc. are the most important biomarker genes and WY14643_High, clofibrate_High, benzbromarone_High, aspirin_High etc. are their important regulatory DCCs in PPAR signaling pathway. The detail results of these relationships are given in the **Table S13**.

DISCUSSION AND CONCLUSIONS

Identification of biomarker genes and their regulatory DCCs is one of the most important tasks in the toxicogenomics studies as well as in drug design and development as mentioned before. In this article, we have proposed a robust co-clustering approach based on logistic probabilistic hidden variable model (LPHVM) to detect important biomarker genes and their regulatory DCCs. The proposed LPHVM approach is robust against outlying gene expressions and more flexible and effective than the application of one-way classical clustering approaches (e.g., k-means, fuzzy, HC, etc.) for co-clustering. The proposed method produces robust results by using the logistic transformation of fold-change gene expression data into the conventional PHVM approach. The logistic transformation reduces unusual/outlying observations into the reasonable space without changing the original hidden patterns of genes and DCCs in the dataset. Thus the proposed LPHVM approach produces robust results.

We investigated the performance of the proposed LPHVM method in a comparison with the traditional PHVM and Toxygates online computational platform using simulated and real life TGP gene expression data, respectively. The simulation results showed that the proposed method improves the performance over the conventional PHVM in presence of outlying observations; otherwise, they perform equally. We also demonstrated the performance of the proposed method in a comparison with the online computational platform "Toxygates" using the real life pathway based fold change gene expression datasets collected from the "Toxygates" database. We observed that three DCCs (glibenclamide-low, perhexilline-low, and hexachlorobenzene-medium) for glutathione metabolism pathway dataset as well as two DCCs (acetaminophen-medium and methapyrilene-low) for PPAR signaling pathway dataset were incorrectly co-clustered by the Toxygates online platform, while only one DCC (hexachlorobenzene-low) for glutathione metabolism pathway was incorrectly co-clustered by the proposed LPHVM approach. Our findings from the real life data analysis are also supported by the other findings in the literature (Kiyosawa et al., 2006; Nyström-Persson et al., 2013). Thus the proposed LPHVM outperform over the classical PHVM and "Toxygates" online coputational platform to detect toxicogenomic biomarkers and their regulatory DCCs.

DATA AVAILABILITY

The demo data and R-code are provided at http://www.bbcba.org/softwares/rCoClust.zip.

AUTHOR CONTRIBUTIONS

MH and MM worked together to develop the algorithm. MH analyzed the data and drafted the manuscript. MM coordinated and supervised the project. MMR, AB, and MR attended at the meeting regarding this article as well as read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

We are grateful to the authority of the bioinformatics laboratory, department of statistics, University of Rajshahi, Rajshahi, Bangladesh for their co-operation and giving chance to do this research work in their laboratory.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00516/full#supplementary-material

REFERENCES

Afshari, C. A., Hamadeh, H. K., and Bushel, P. R. (2011). The evolution of bioinformatics in toxicology: advancing toxicogenomics. *Toxicol. Sci.* 120, S225–S237. doi: 10.1093/toxsci/kfq373

Agostinelli, C. C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test* 24, 441–461. doi: 10.1007/s11749-015-0450-6

Alqallaf, F., Van, A. S., Yohai, V., and Zamar, R. (2009). Propagation of outliers in multivariate data. *Ann. Stat.* 37, 311–331. doi: 10.1214/07-AOS588

Atkinson, A. C. (1982). Regression diagnostics, transformation and constructed variables. *J. R. Stat. Soc. Ser. B* 44, 1–36.

Bicego, M., Lovato, P., Ferrarini, A., and Delledonne, M. (2010). "Biclustering of expression microarray data with topic models," in *International Conference on Pattern Recognition* (Washington, DC: IEEE Computer Society), 2728–2731. doi: 10.1109/ICPR.2010.668

Birben, E., Sahiner, U. M., Cansin Sackesen, C., Serpil Erzurum, S., and Omer Kalayci, O. (2012). Oxidative stress and antioxidant defense. *World Allergy Organ. J.* 5, 9–19. doi: 10.1097/WOX.0b013e3182439613

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. Ser. B* 26, 211–252.

Carroll, R. J. (1982). Two examples of transformations when there are possible outliers. *Appl. Stat.* 31, 149–152. doi: 10.2307/2347978

Chen, M., Zhang, M., Borlak, J., and Tong, W. (2012). A decade of toxicogenomic research and its contribution to toxicological science. *Toxicol. Sci.* 130, 217–228. doi: 10.1093/toxsci/kfs223

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–38.

Fielden, M. R., Brennan, R., and Gollub, J. (2007). A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by non genotoxic chemicals. *Toxicol. Sci.* 99, 90–100. doi: 10.1093/toxsci/kfm156

Gao, W., Mizukawa, Y., Nakatsu, N., Minowa, Y., Yamada, H., Ohno, Y., et al. (2010). Mechanism-based biomarker gene sets for glutathione depletion-r elated hepatotoxicity in rats. *Toxicol. Appl. Pharmacol.* 247, 211–221. doi: 10.1016/j.taap.2010.06.015

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. A. (2010). A review of robust clustering methods. *Adv. Data Anal. Classif.* 4, 89–109. doi: 10.1007/s11634-010-0064-5

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. E. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biomatrics* 62, 10–18. doi: 10.1111/j.1541-0420.2005.00397.x

Hardt, C., Beber, M. E., Rasche, A., Kamburov, A., Hebels, D. G., Kleinjans, J. C., et al. (2016). ToxDB: pathway-level interpretation of drug-treatment data. *Database* 2016:baw052. doi: 10.1093/database/baw052

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.* 42, 177–196. doi: 10.1023/A:1007617005950

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651

Huang da, W., Sherman, B. T. and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., et al. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43(Database issue), D921–D927. doi: 10.1093/nar/gku955

Joung, J. G., Shin, D., Seong, R. H., and Zhang, B. T. (2006). Identification of regulatory modules by co-clustering latent variable models: stem cell differentiation. *Bioinformatics* 22, 2005–2011. doi: 10.1093/bioinformatics/btl343

Kim, S. (2017). Identifying dynamic pathway interactions based on clinical information. *Comput. Biol. Chem.* 68, 260–265. doi: 10.1016/j.compbiolchem.2017.04.009

Kiyosawa, N., Shiwaku, K., Hirode, M., Omura, K., Uehara, T., Shimizu, T., et al. (2006). Utilization of a one-dimensional score for surveying chemical-induced changes in expression levels of multiple biomarker gene sets using a large-scale toxicogenomics database. *J. Toxicol. Sci.* 31, 433–448. doi: 10.2131/jts.31.433

Madeira, S. C., and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEE Trans. Comput. Biol. Bioinform.* 1, 24–45. doi: 10.1109/TCBB.2004.2

Malika, C., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data Set. *J. Stat. Softw.* 61, 1–36. doi: 10.18637/jss.v061.i06

NRC (2007). *National Research Council of the National Academies: Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment*. Washington, DC: National Academies Press.

Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C., and Afshari, C. A. (1999). Microarrays and toxicology: the advent of toxicogenomics. *Mol. Carcinog.* 24, 153–159. doi: 10.1002/(SICI)1098-2744(199903)24:33.0.CO;2-P

Nyström-Persson, J., Igarashi, Y., Ito, M., Morita, M., Nakatsu, N., Yamada, H., et al. (2013). Toxygates: interactive toxicity analysis on a hybrid microarray and linked data platform. *Bioinformatics* 23, 3080–3086. doi: 10.1093/bioinformatics/btt531

Nyström-Persson, J., Natsume-Kitatani, Y., Igarashi, Y., Satoh, D., and Mizuguchi, K. (2017). Interactive toxicogenomics: gene set discovery, clustering and analysis in Toxygates. *Sci Rep*. 7:1390. doi: 10.1038/s41598-017-01500-1

Peraza, M. A., Burdick, A. D., Marin, H. E., Gonzalez, F. J., and Peters, J. M. (2006). The Toxicology of ligands for peroxisome proliferator-activated receptors (PPAR). *Toxicol. Sci.* 90, 269–295. doi: 10.1093/toxsci/kfj062

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc.* 63, 411–423. doi: 10.1111/1467-9868.00293

Uehara, T. (2010). The Japanese toxicogenomics project: application of toxicogenomics. *Mol. Nutr. Food Res.* 54, 218–227. doi: 10.1002/mnfr.200900169

Uehara, T., Hirode, M., Ono, A., Kiyosawa, N., Omura, K., Shimizu, T., et al. (2008). A toxicogenomics approach for early assessment of potential non-genotoxic hepatocarcinogenicity of chemicals in rats. *Toxicology* 250, 15–26. doi: 10.1016/j.tox.2008.05.013

Upton, G. J. G., Sanchez-Graillet, O., Rowsell, J., Arteaga-Salas, J. M., Graham, N. S., Stalteri, M. A., et al. (2009). On the causes of outliers in Affy metrix GeneChip data. *Brief. Funct. Genomic Proteomic.* 8, 119–212. doi: 10.1093/bfgp/elp027

Yildirimman, R., Brolén, G., Vilardell, M., Eriksson, G., Synnergren, J., Gmuender, H., et al. (2011). Human embryonic stem cell derived hepatocyte-like cells as a tool for *in vitro* hazard assessment of chemical carcinogenicity. *Toxicol. Sci.* 124, 278–290. doi: 10.1093/toxsci/kfr225

Zhu, S., Okuno, Y., Tsujimoto, G., and Mamitsuka, H. (2005). A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics* 21(Suppl. 2), 245–251. doi: 10.1093/bioinformatics/bti1141

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 November 2018
doi: 10.3389/fgene.2018.00508

Weighted Gene Correlation Network Analysis (WGCNA) Reveals Novel Transcription Factors Associated With Bisphenol A Dose-Response

*Alexandra Maertens[1], Vy Tran[1], Andre Kleensang[1] and Thomas Hartung[1,2,3]**

[1] Center for Alternatives to Animal Testing, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States, [2] Center for Alternatives to Animal Testing - Europe, University of Konstanz, Konstanz, Germany, [3] Doerenkamp-Zbinden Professor and Chair for Evidence-Based Toxicology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

Despite Bisphenol-A (BPA) being subject to extensive study, a thorough understanding of molecular mechanism remains elusive. Here we show that using weighted gene correlation network analysis (WGCNA), which takes advantage of a graph theoretical approach to understanding correlations amongst genes and grouping genes into modules that typically have co-ordinated biological functions and regulatory mechanisms, that despite some commonality in altered genes, there is minimal overlap between BPA and estrogen in terms of network topology. We confirmed previous findings that ZNF217 and TFAP2C are involved in the estrogen pathway, and are implicated in BPA as well, although for BPA they appear to be active in the absence of canonical estrogen-receptor driven gene expression. Furthermore, our study suggested that PADI4 and RACK7/ZMYNDB8 may be involved in the overlap in gene expression between estradiol and BPA. Lastly, we demonstrated that even at low doses there are unique transcription factors that appear to be driving the biology of BPA, such as SREBF1. Overall, our data is consistent with other reports that BPA leads to subtle gene changes rather than profound aberrations of a conserved estrogen signaling (or other) pathways.

Keywords: bisphenol A, estrogen, WGCNA, ZNF217, TFAP2C, ZMYND8, PADI4, SREBF1

OPEN ACCESS

Edited by:
Paul Jennings,
VU University Amsterdam,
Netherlands

Reviewed by:
Francesco Russo,
University of Copenhagen, Denmark
Matteo Brilli,
Università degli Studi di Milano, Italy

***Correspondence:**
Thomas Hartung
thartun1@jhu.edu;
thartung@jhsph.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *17 July 2018*
Accepted: *10 October 2018*
Published: *12 November 2018*

Citation:
Maertens A, Tran V, Kleensang A and Hartung T (2018) Weighted Gene Correlation Network Analysis (WGCNA) Reveals Novel Transcription Factors Associated With Bisphenol A Dose-Response. Front. Genet. 9:508. doi: 10.3389/fgene.2018.00508

INTRODUCTION

Bisphenol A (BPA) is an industrial chemical used in the manufacture of polycarbonate plastic found in a number of consumer products such as thermal paper, canned foods and epoxy resins (Rubin, 2011) – although many of these uses are being phased out (Zimmerman and Anastas, 2015). Amongst the general population, exposure to BPA is widespread, with very low levels of BPA present in the majority of urinary samples taken in the general population (Calafat et al., 2008), although serum levels are estimated to be lower (Teeguarden et al., 2013). Release of BPA to the environment exceeds one million pounds per year (Rubin, 2011).

Bisphenol-A has been subjected to a high-level of scrutiny – "bisphenol A" returns over 11,500 abstracts in PubMed, with over 700 articles per year being published every year since 2013 ("Pubmed Bisphenol A, n.d.). Within HSDB (the Hazardous Substance Database), there are over 79

peer-reviewed animal studies (TOXNET, n.d.). The CLARITY study, a three generation chronic study with low-levels of BPA used to mimic population exposures, involved 3,500 rats: while it resulted in no revision of safety standards by the FDA, it still failed to bring about a consensus as to a safe level (Academics Urge Caution in Interpreting Clarity-Bpa Results, n.d.). Despite the overwhelming amount of data, the mechanism(s) by which BPA may exert adverse effects remains unclear, nor is there a widely agreed upon endpoint on which to base a safe dose.

Bisphenol-A was presumed to have potentially estrogenic activity, as well as potential carcinogenicity, based on its structural similarity to DES (Diethylstilbestrol) and other synthetic estrogens, as well as appearing to trigger gene expression similar to estrogen receptor agonists, despite its relatively low binding affinity for estrogen receptors (LaPensee et al., 2009). Two different hypotheses have been put forward to explain this discrepancy: one, BPA may bind to different domains of ESR1 or ESR2 and recruit different co-regulators (Safe et al., 2002), or alternatively, BPA may exert its effects through non-classical estrogen receptors, such as membrane-bound ER (GPR30) (LaPensee et al., 2009) or ERRγ (ERRG) (Okada et al., 2008), which is one of several "orphan" receptors that are classified as estrogen-related receptors (Horard and Vanacker, 2003). The question of BPA's ultimate molecular initiating event is not academic – on the presumption that BPA's effects are mediated via estrogen receptors, several alternatives were proposed, such as Bisphenol F and Bisphenol S, but both compounds have proven equally problematic (Rochester and Bolden, 2015).

In our previous work for the Mapping the Human Toxome project (Kleensang et al., 2014; Bouhifd et al., 2015), we demonstrated that using non-inferential statistical methods that did not depend on existing annotations such as IDEA (Pendse et al., 2017) and WGCNA (Maertens et al., 2015) offered a powerful method to untangle possible regulatory mechanisms and providing insight into possible Pathways of Toxicity compared to either inferential-based methods or approaches such as pathway enrichment analysis that depend exclusively on annotations. Building upon our previous work using WGCNA applied to *in vitro* transcriptomic data to more fully understand the transcription factors that are driving the biology of estradiol (Pendse et al., 2017), we used WGCNA to examine a previously published transcriptomic dataset (Shioda et al., 2013). Briefly, Shioda et al. (2013) aimed to study the sensitivities of estrogen responsive genes to various endocrine disrupting chemicals (EDCs) based on the transcriptomic profile of MCF-7 cells exposed to either estrogen or several xenoestrogens (including BPA) over a dose-response curve ranging from picomolar to micromolar concentrations for a 48 h time period. Based on their analysis, they found that a gene signature of "estrogen-responsive genes" allowed the estrogenic substances to be ranked in terms of potency. Additionally, the heat map of BPA-inducible genes demonstrated a weak transcriptional activation at very low BPA concentration as well as a strong peak at high concentration. However, BPA has differences as well as similarities to estrogen in terms of gene signatures: therefore, we sought to explore specifically the differences between estrogen and BPA as well as the differences between low-dose BPA and high-dose BPA for possible regulatory mechanisms.

Our analysis shows that while there is substantial overlap between genes altered by BPA and estrogen, which might imply that BPA is indeed "estrogenic," there are important differences in network topology as well as biological function, and that the overlap appears to be driven by transcription factors such as ZNF217, TFAP2C, PADI4, and RACK7/ZMYND8 rather than the estrogen receptor *per se*. Furthermore, BPA (even at the lower end of the dose response curve - defined here as less than 12.5 μM) has pathways that are likely not mediated by estrogen receptors, but instead by other transcription factors, such as SREBF1. Moreover, our data is consistent with other reports that BPA leads to subtle, diffuse gene changes that are comparatively difficult to capture with inferential methods, and that low-dose BPA has distinct effects compared to higher doses.

MATERIALS AND METHODS

Data

Dataset GSE50705, a comprehensive analysis of estrogen and xenoestrogen dose-response curves on MCF-7 cells after 48 h of exposure, was downloaded from GEO via GEOQuery (Davis and Meltzer, 2007) as normalized data and all analyses were performed with R/Bioconductor (Gentleman et al., 2004).

Weighted Gene Correlation Network Analysis

A WGCNA network (Langfelder and Horvath, 2007) was generated for several subsets of the data: Estrogen (n = 36), BPA (n = 44), and low dose BPA (n = 32), as well as a consensus network for estrogen and BPA together (n = 80) using the 10,000 most highly expressed genes for each subset of the data as determined by rank means expression, the approach in Maertens et al. (2015); consensus networks and module statistics followed overall the approach in Langfelder et al. (2008). Briefly, the network was derived based on a signed Spearman correlation using a β of 10 as a weight function. The topological overlap metric (TOM) (Yip and Horvath, 2007) was derived from the resulting adjacency matrix, and was used to cluster the modules using the blockwiseModules function (blockwiseConsensusModules, for the consensus modules) and the dynamic tree cut algorithm (Langfelder et al., 2008) with a height of 0.25 and a deep split level of 2, a reassign threshold of 0.2 and a minimum module size of 30 (100 for the consensus network). The eigenmodules— essentially the first principal component of the modules, which can be used as a "signature" of the modules gene expression —were then correlated with dose, and each module that was correlated with the dose-response curve with a p-value < 0.01 (p-value < 0.05 for the consensus network) was considered statistically significant.

Transcription Factor Analysis

All statistically significant modules were analyzed in EnrichR (Chen et al., 2013) using the CHEA dataset (Lachmann et al.,

2010) restricted to MCF-7/10 cells –as well as the ARCHS4 TF-Coexpression dataset with an adjusted *p*-value less than 0.01 based on Fisher's exact test.

Functional Annotation Analysis

Module "hubs" were defined as having high-ranking kME (which ranks the connectivity of genes) within the module and predicted as high-degree within the STRING database (Szklarczyk et al., 2017) of protein–protein interactions. The three highest ranking modules were analyzed in STRING for the enrichment of predicted protein interactions as well as functional annotation via GO Biological Process and Molecular Function. All analysis with STRING was done with medium stringency settings, and included all possible interactions (text-mining, database, experiments, co-expression, neighborhood, gene fusion, and co-occurrence).

TCGA Data

Expression and methylation data for FIZ1 was correlated with clinical attributes (estrogen receptor, progesterone receptor, and solid tumor vs. normal vs. metastatic tumor) using MEExpress (Koch et al., 2015) based on the TCGA BRCA dataset.

RESULTS

Consensus Network Analysis Indicates Minimal Overlap Between Estrogen and BPA

We began by analyzing the dose-response curve of the BPA and estrogen dataset combined using WGCNA (which takes advantage of correlations amongst genes and groups genes into modules using network topology) to look for a "consensus network"-a common pattern of genes that are correlated in all conditions. The consensus network identified (**Figure 1**) had clearly delineated modules, and the modules identified were significantly correlated with both estrogen (**Figure 2**) and BPA (**Figure 3**). However, estrogen clearly had a stronger signal in comparison to BPA and quite possibly overwhelmed the signal from BPA. More strikingly, however, the majority of modules in the consensus analysis when analyzed for correlation with both BPA and estrogen showed virtually no similarity – most modules had opposite directions of correlation, and of the few modules with similar correlations, the coefficient of correlation was very weak – only one module ("yellow") was significant with a p-value < 0.05 (**Table 1**), indicating that while there may be some overlap in genetic signatures, from a network topology perspective there is minimal conservation. When analyzed for transcription factors against the CHEA dataset – a collection of ChIP-chip, ChIP-seq, ChIP-PET, and DamID studies collected into a database to infer transcriptional regulation (Lachmann et al., 2010) – the common module was enriched for E2F1, ZNF217, and RACK7, but not ESR1 or ESR2 (**Table 2**).

Estrogen and BPA Network Overlap With Transcription Factors, Including ESR1 and ESR2, but With Different Network Topologies and Different Biological Processes

Next, we derived the *de novo* networks individually for the entire dose-response curve of estrogen and BPA to examine the network topology, common hubs, and biological role of the modules in each network separately. Within the estrogen network (**Supplementary Figure 1A**), there was one large module highly correlated with estrogen dose ("turquoise") (**Table 3**), which was also enriched for ESR1 and ESR2 in addition to E2F1, ZNF217, TFAP2C amongst others (**Table 4**); furthermore, ESR1 was a hub within the module, and the module was predominately enriched with terms related to cell-cycle as well as poly(A) RNA binding (**Supplementary Table 1**).

In comparison, the top module correlated with BPA dose ("lightcyan1") (**Supplementary Figure 1B**) (**Table 5**), was a relatively small module not enriched for any transcription factors, however, SSR1 (Signal Sequence Receptor Subunit 1) was the top hub; annotation analysis revealed that the module was enriched for genes in the GO category of "response to endoplasmic reticulum stress" and "endoplasmic reticulum unfolded protein response" (p-value of 2.93E-17 and 2.02E-12, respectively) (**Supplementary Table 1**). The second module correlated with dose ("royal blue") was enriched for ESR1 and ESR2 genes (**Table 6**), however, neither ESR1 nor ESR2 was present in the module (or any module correlated positively or negatively with dose) and instead the main hub was TOP2A (Topoisomerase IIA). The module had a weak over-representation of genes involved in development (p-value 0.00416) and cytoskeleton organization (p-value 0.0155) (**Supplementary Table 1**). The other module enriched for ESR1 and ESR2 genes ("dark gray") was also annotated to "response to unfolded protein" (p-value of 9.52E-05) (**Supplementary Table 1**).

Low-Dose BPA Network Shows No Enrichment of ESR1 or ESR2 Genes

It has been speculated that BPA at low doses has fundamentally different effects than at high doses; in the original study of the dataset, the authors detected a weak, but distinct, transcriptional activity peak at low doses. Therefore, we restricted the BPA network to doses below 12.5 μM (leaving a highest dose of 6.25 μM, and most of the dose-response curve in the nanomolar/picomolar range) and calculated a network specific for this lower dose range. Despite the smaller sample size, the network still produced several modules that were significantly correlated with dose (**Supplementary Figure 1C** and **Table 7**). This low-dose BPA network shows consistent transcription factors (ZNF217, TFAP2C, RACK7/ZMYND8, and PADI4) with the larger BPA network as well as the estrogen network, but no modules were enriched for genes with ESR1 or ESR2 with a p-value cut-off of < 0.01 (**Table 8**).

The module with the highest correlation with dose ("turquoise") was comparatively dense for predicted protein-protein interactions (p-value of $<$ 1.0e-16, average node

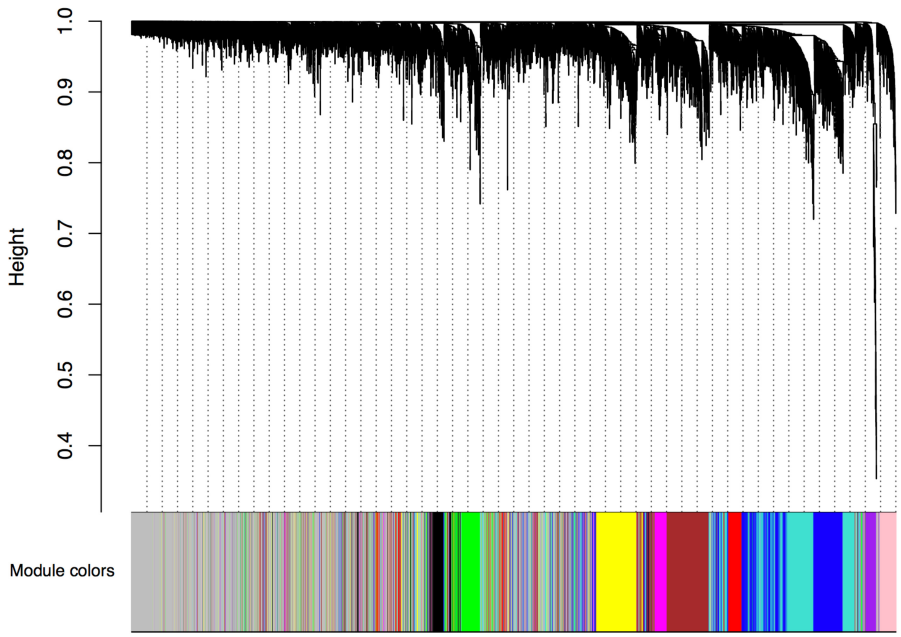


FIGURE 1 | Consensus network from BPA and Estrogen dose-response curve. Gene expression similarity is determined using a pair-wise weighted correlation metric, and clustered according to a topological overlap metric into modules; assigned modules are colored on bottom, gray genes are unassigned to a module.

degree 10.3) as well as genes related to cellular macromolecule metabolic process and poly(A) RNA binding (*p*-value of 1.58E-16 and 2.17eE-18, respectively) (**Supplementary Table 1**) in contradistinction to the estrogen network, where the dominant module was enriched overwhelmingly with cell-cycle genes. Moreover, the module included both ZNF217 and TFAP2C, but neither ESR1 nor ESR2 were in the module, much less hubs. The second module correlated with dose ("Dark Green") showed no enrichment for transcription factors, although it was enriched for protein–protein interactions and the molecular function "enzyme binding"; the third module ("dark red"), also strongly correlated with dose, showed no enrichment for transcription factors or protein–protein interactions, though it was weakly enriched for the KEGG pathway Insulin Signaling (*p*-value 0.0028); this module may simply be an artifact, reflect diffuse alterations that are difficult to detect, or an unknown regulatory mechanism. An additional fairly large module correlated with dose ("brown") was enriched for both E2F1 and PADI4, strongly enriched for protein-protein interactions (*p*-value < 1.0E-16, average node degree 6.56) and as well as cellular metabolic process (*p*-value 1.32E-10) and poly(A) RNA binding (*p*-value 2.5E-16) (**Supplementary Table 1**); but, also in contrast to the estrogen network, was not enriched for cell-cycle genes.

ZNF217 has previously been shown by our work (Pendse et al., 2017) and others (Frietze et al., 2014) to be a critical component of estrogen signaling and an important prognostic factor for breast cancer (Vendrell et al., 2012). Similarly, TFAP2C is known to modulate ESR1 and GPR30 expression, and attenuate the expression of several estrogen-targeted genes (Woodfield et al., 2007). Given the presence of both ZNF217 and TFAP2C in the network as well as the strong enrichment genes targeted by these transcription factors, this suggests that these genes are indeed central to mediating BPAs phenotypic effects; however, our study shows little evidence that they are in exerting their effect in tandem with ESR1 or ESR2.

Moreover, both ZNF217 and TFAP2C were shown independently to be altered by bisphenol A in a rat seminiferous tubule culture model (Ali et al., 2014). The same study also showed alterations (albeit subtle) in PADI4 and RACK7 (official gene symbol: ZYMND8) mRNA as well as RACK7/ZMYND8 methylation; neither of these genes were in our dataset so their role in observed changes remains speculative. However, RACK7/ZYMND8 binds a large set of active enhancers, including almost all "super-enhancers," and is therefore expected to have sweeping transcriptional effects (Shen et al., 2016). Although little is known about its role in breast cancer, it is thought to inhibit HIF-dependent breast-cancer progression (Chen et al., 2018). PADI4 is known to be implicated in cancer and is thought to respond to estrogen-simulation in MCF-7 cells through both genomic and non-genomic mechanisms (Dong et al., 2007). In breast cancer specifically, it is implicated in the ELK1/C-Fos pathway (Zhang et al., 2011). Moreover, BPA was shown to increase protein levels of PADI4 via a reactive oxygen species mechanism in neuroblastoma cells (Park et al., 2012).

Low-Dose BPA Network Had Unique Transcription Factors Not Present in the Estrogen Dataset

To further delineate possible transcription factors unique to BPA signaling compared to estrogen, we examined the list of genes in all modules statistically significantly associated with the

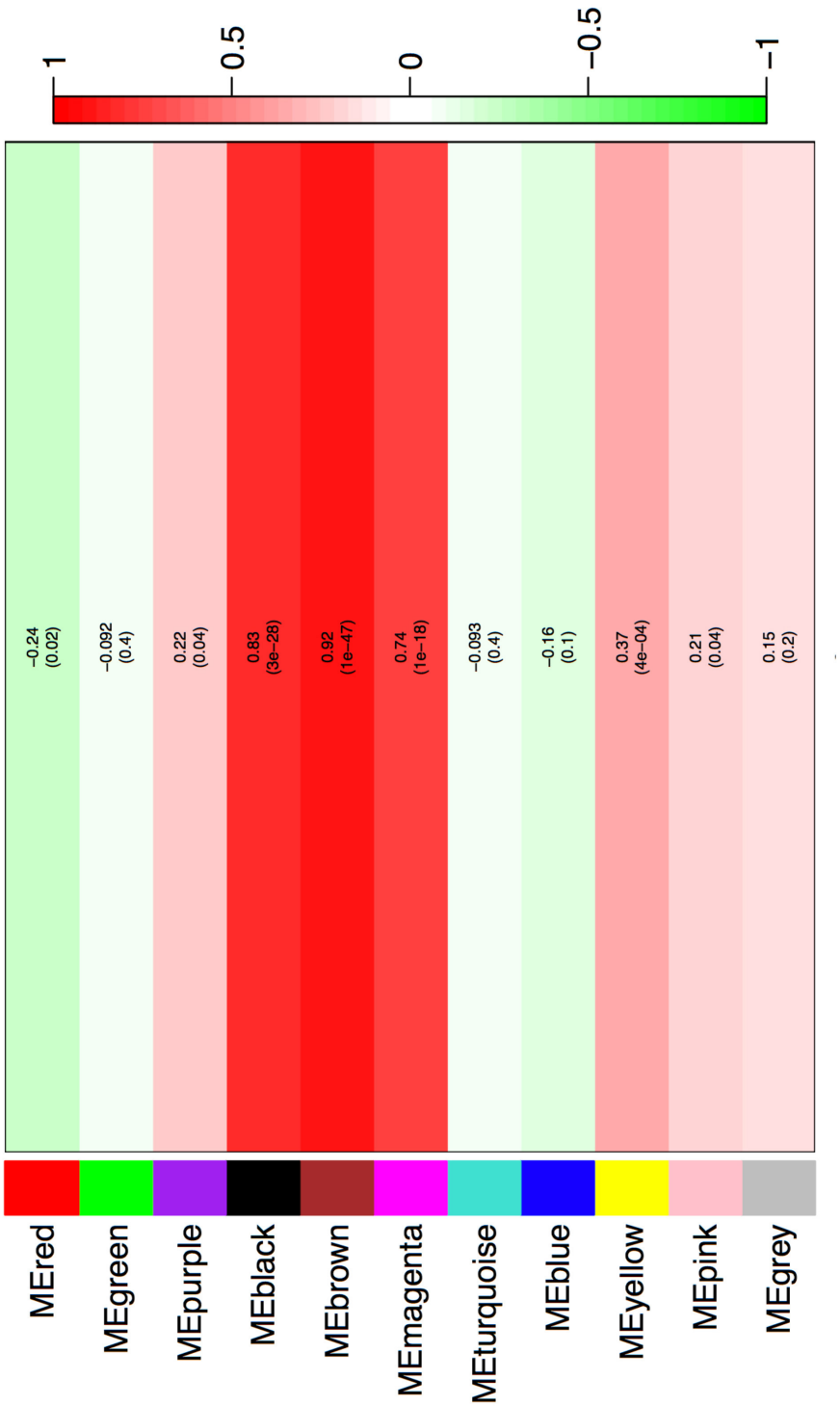


FIGURE 2 | Consensus network modules correlated with estrogen dose using the eigenmodule (the first principal component of the module). Correlation coefficient along with *p*-value in parenthesis underneath; color-coded according to correlation coefficient (legend at right).

low-dose BPA network that were not present in the estrogen network, a total of 1,901 genes. Analyzed against the CHEA dataset, the genes were again enriched for RACK7/ZMYND8, in addition to ELK1 and HIF1A (**Supplementary Table 2**). In order to expand our search for transcription factors that may not have been studied in MCF-7 cells in the CHEA data set, we also analyzed the list of genes for enrichment against the ARCHS4 database (Lachmann et al., 2018), which correlates

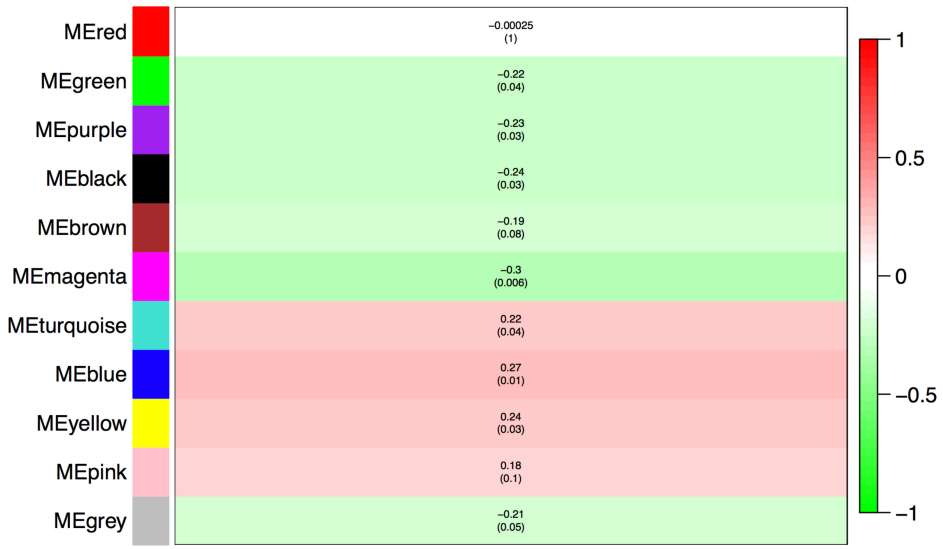


FIGURE 3 | Consensus network modules correlated with BPA dose using the eigenmodule (the first principal component of the module). Correlation coefficient along with *p*-value in parenthesis underneath; color-coded according to correlation coefficient (legend at right).

TABLE 1 | Consensus network modules associated with BPA and estrogen.

Module	Correlation	*p*-value
Red	−0.00025	1
Green	−0.092	0.4
Purple	NA	NA
Black	NA	NA
Brown	NA	NA
Magenta	NA	NA
Blue	NA	NA
Yellow	0.24	0.03
Pink	0.18	0.1
Gray	NA	NA

Consensus network modules were correlated against both estrogen and BPA; NA indicates that the modules had different directions of correlation in estrogen compared to BPA.

TABLE 2 | Enriched transcription factors in conserved module in consensus network.

Transcription factor	Adjusted *p*-value
E2F1	0.000003501
ZNF217	0.000004066
RACK7/ZMYNDB	0.005142

Transcription factors significantly enriched in the conserved module ("yellow") between BPA and estrogen.

TABLE 3 | Estrogen modules correlated with dose

Module	Correlation	*p*-value
Turquoise	0.710543783	1.73E-06
Dark Green	0.667630497	1.18E-05
Dark Red	0.548734271	6.42E-04
Light Yellow	0.470241047	4.36E-03
Brown	0.45994764	5.44E-03
Salmon	0.443256223	7.66E-03
Gray60	−0.47528565	3.91E-03
Blue	−0.5389571	8.36E-04
Yellow	−0.55352085	5.62E-04
Black	−0.63950993	3.54E-05
Midnight Blue	−0.73602275	4.69E-07

All modules correlated with estrogen dose-response curve with a p-value less than 0.01.

transcription factor expression against gene expression in a combined database of over 20,000 RNASeq samples. Of the top 50 transcription factors identified as significantly correlated with the gene list, 18 were also present in the low-dose BPA network (**Table 9**). The highest-ranking transcription factor, FIZ1, is zinc-finger protein with a largely unknown biological role (Wolf and Rohrschneider, 1999) - it has a relatively poor literature base, with only 8 citations in PubMed. However, FIZ1 expression in breast cancer is statistically associated with progesterone receptor status, estrogen receptor status, and sample subtype, and it undergoes extensive CpG-island methylation (**Supplementary Figure 2**), and it is therefore an intriguing candidate for further study. The second highest-ranking transcription factor, SREBF1 is comparatively better characterized: it is known to be central to lipid homeostasis, regulating the LDL receptor gene as well as related fatty acid and cholesterol synthesis genes. Furthermore, SREBF1 mRNA was identified as upregulated in adipocytes by BPA (Boucher et al., 2014). Neither SREBF1 nor FIZ1 were present in the estrogen dataset, and SREBF1- and FIZ1-correlated genes were not enriched in the subset of estrogen-only genes. It is therefore plausible that these two transcription factors are more central to BPAs effects than estrogen, however, because enrichment for transcription factors motifs/regulated genes in

TABLE 4 | Enriched transcription factors in estrogen modules.

Module	TF	Adjusted *p*-value
Black	RACK7	0.005152
Blue	ZNF217	2.79E-14
	PADI4	1.10E-08
	RACK7	8.23E-07
	TFAP2C	0.00001662
	GATA3	0.00004515
	FOXM1	0.000373
	E2F1	0.000674
Turquoise	E2F1	1.43E-20
	ESR1	7.21E-10
	ESR2	2.12E-09
	PADI4	1.34E-07
	RACK7	7.39E-07
	GATA3	0.0000105
	RUNX1	0.0003015
	ZNF217	0.0006679
	TFAP2C	0.0008849
	ELK1	0.001171
	FOXM1	0.001239
Yellow	PADI4	0.001012
	RUNX1	0.001601

All statistically significant modules correlated with dose, with enriched transcription factors. Modules that did not have TFs with an adjusted p-value ≤ 0.01 were excluded (Dark Red, Dark Green, Salmon, Midnight Blue, Gray60).

TABLE 5 | BPA modules associated with dose.

Module	Correlation	*p*-Value
Lightcyan 1	0.725175961	6.17E-15
Royal Blue	0.423995959	5.84E-05
Dark Gray	0.390795349	2.38E-04
Light Yellow	0.383058412	3.23E-04
Ivory	−0.372076037	4.92E-04
Light Cyan	−0.291083662	7.23E-03
Green	−0.280875808	9.65E-03
Gray60	−0.279474408	1.00E-02

All BPA modules correlated with estrogen dose-response curve with a p-value less than 0.01.

TABLE 6 | Enriched transcription factors in BPA modules.

Module	TF	Adjusted *p*-value
Dark Gray	ESR1	1.32E-16
	ESR2	2.74E-08
	ZNF217	0.000002163
	GATA3	0.00001401
Green	ZNF217	1.59E-11
	RACK7	0.00006931
	ESR2	0.0001032
	GATA3	0.0002586
	TFAP2C	0.00055
	ESR1	0.002135
	PADI4	0.002604
	FOXM1	0.003341
Light Cyan	TFAP2C	0.002078
	GATA3	0.002583
Royal Blue	ZNF217	5.24E-08
	ESR2	7.76E-07
	ESR1	0.000005588
	ARNT	0.00005395
	AHR	0.0002705
	GATA3	0.002178

All statistically significant modules correlated with dose, with enriched transcription factors. Modules that did not have TFs with adjusted p-value ≤ 0.01 were excluded (LightCyan1, Ivory, Light Yellow, Plum, Gray60).

TABLE 7 | Low-dose BPA modules associated with dose.

Module	Correlation	*p*-value
Turquoise	0.71054	5.21E-06
Dark Green	0.66763	2.99E-05
Dark Red	0.54873	1.15E-03
Light Yellow	0.47024	6.61E-03
Brown	0.45995	8.08E-03
Salmon	0.44326	1.11E-02
Grey60	−0.4753	5.98E-03
Blue	−0.539	1.46E-03
Yellow	−0.5535	1.02E-03
Black	−0.6395	8.13E-05
Midnight Blue	−0.736	1.58E-06

All low-dose BPA modules correlated with estrogen dose-response curve with a p-value less than 0.01.

any gene list often produce false-positives, understanding their role would require further study.

DISCUSSION

Estrogen signaling is unique amongst nuclear receptors in that substantial number of the genes altered by estrogen do not have canonical estrogen response elements (Miller et al., 2017) – estrogen signaling takes place within a transcriptomic and epigenomic context that markedly influences receptor activation. Our examination of the estrogen dose response curve network both confirmed several of the transcription factors identified previously, such as E2F1, ZNF217 and TFAP2C, as well as suggested other transcriptional factors such as PADI4 and RACK7/ZYMND8 that may impact estrogen signaling.

Our study is consistent with other findings that the assumption that BPA works exclusively or even predominantly on canonical ESR1 or ESR2 gene regulation may be misleading or an oversimplification (Delfosse et al., 2012; MacKay and Abizaid, 2018). To be sure, one can find gene patterns similar to those found in estrogen-induced cells, but the leap from that observation to the presumption that such changes are estrogen-mediated may not be warranted. While this study cannot determine conclusively the ultimate chain of events that leads from the molecular initiating event to the phenotypic

TABLE 8 | Enriched transcription factors in low-dose BPA modules.

Module	TF	Adjusted *p*-value
Black	RACK7	1.78E-07
	TFAP2C	0.00001339
	RUNX1	0.00004104
Blue	ELK1	0.00000437
	ZNF217	6.14E-07
	PADI4	0.000007639
	FOXM1	9.66E-08
	HIF1A	0.005915
	AHR	0.00208
	E2F1	0.002221
	ARNT	0.005915
	RUNX1	0.008124
	GATA3	0.008124
Brown	E2F1	0.002459
	PADI4	0.006908
Turquoise	ZNF217	0.00000179
	RACK7	0.00001912
	GATA3	0.00002574
	PADI4	0.0001018
	FOXM1	0.0001703
	RUNX1	0.0004958
	E2F1	0.001544
Yellow	E2F1	6.34E-18
	PADI4	0.003362
	RACK7	0.00334
	FOXM1	0.009047

All statistically significant modules correlated with dose, with enriched transcription factors. Modules that did not have TFs with adjusted p-value ≤ 0.01 were excluded (Light Yellow, Salmon, Gray60, Dark Green, Midnight Blue, Dark Red).

consequences, it does suggest some hypotheses that are more probable. The lack of overlap in the consensus network indicates that despite similarity of genes, there is minimal conservation of network topology, and the one conserved module was not enriched for ESR1 or ESR2 genes. In networks drawn separately from dose-response curves for estrogen and BPA, the substantial differences in network topology, the absence of ESR1 as a hub gene in the BPA network, and the differences in biological function of the modules suggest that even at high-doses, BPAs effects are fundamentally different than estradiol. The lack of estrogen receptor target genes in the low dose BPA network in the presence of a clear signature of other transcription factors suggests that at low doses BPA's effects are driven by mechanisms other than direct estrogen receptor activation. Additionally, regardless of molecular initiating event, assessing BPAs dose-response by looking at estrogen gene-signatures may miss interesting and important biology, such as the likely role of SREBF1. Furthermore, our study is consistent with other findings that BPA's effects are subtle and phenotypic changes likely reflect modest effects at multiple different points (Porreca et al., 2016) and that analyzing the effects of low-dose BPA can reveal effects that are obscured at higher doses (Shioda et al., 2013). This does not necessarily lead to a "non-monotonic" dose response curve

TABLE 9 | Transcription factors unique to low-dose BPA network.

Transcription factor	Adjusted *p*-value
HSF1	4.54E-32
MBD3	4.72E-30
ZNF787	4.54E-32
ZNF205	4.76E-29
HMG20B	1.48E-29
REPIN1	4.76E-29
FIZ1	4.76E-29
SLC2A4RG	1.48E-29
SREBF1	2.14E-28
ZNF598	9.61E-28
THAP4	1.75E-26
SNAPC4	4.29E-27
ZNF768	1.75E-26
E4F1	6.78E-26
MRPL28	6.78E-26
TUT1	2.67E-25
ERF	2.67E-25
CENPB	6.78E-26
KLF16	2.67E-25
WIZ	1.98E-23
ANAPC2	1.17E-24
ZFP41	1.98E-23
DVL2	6.94E-23
ZNF512B	1.98E-23
SRF	1.98E-23
ELK1	6.94E-23
ZNF282	6.94E-23
AKAP8L	2.65E-22
ZBTB45	6.94E-23
NCOR2	6.94E-23
CIZ1	2.65E-22
TRMT1	2.65E-22
CIC	6.94E-23
NR2F6	1.01E-21
ZNF687	2.65E-22
MTA1	3.81E-21
RBM10	1.01E-21
GATAD2A	1.01E-21
ZNF653	1.01E-21
ZNF777	3.81E-21
EDF1	3.81E-21
PRR12	3.81E-21
SIX5	5.23E-20
TIGD5	5.23E-20
MTA2	1.43E-20
MAZ	1.43E-20
CCDC71	1.43E-20
MLLT1	1.43E-20
SF3A2	2.04E-19
GMEB2	5.23E-20

Genes unique to low-dose BPA compared to estrogen were analyzed against the ARCHS4 for potential transcription factor enrichment; all the transcription factors also present in the dataset were identified in yellow.

– this could be due to technical reasons, or higher-doses could cause non-specific changes that are the result of cellular stress, as evidenced by the identification of modules associated with unfolded protein response. It does, however, point to a need to consider the doses chosen for an *in vitro* study carefully and to not presume linear effects.

This study is certainly not a definitive study of BPA molecular mechanisms: our conclusions cannot confidently be extrapolated to other tissue types, as BPA may have tissue specific effects; MCF-7 cells are prone to artifacts (Kleensang et al., 2016); and our study did not focus on epigenetic mechanisms which are speculated as significantly underpinning much of the observed adverse events seen with BPA exposure, especially at a low dose (Singh and Li, 2012). While using the CHEA dataset and restricting candidate transcription factors to those observed in MCF-7 cells eliminates many of the false-positives intrinsic to such approaches, it also limits findings to those transcription factors that have been studied, and this may miss some important biology. Extending our analysis with the ARCHS4 database added interesting candidates, but all correlation-based approaches must be treated with caution and viewed as "hypothesis-generating," and all exploratory data analysis techniques such as WGCNA require further targeted studies to confirm suggested molecular networks.

Nonetheless, our study does indicate that transcriptomics, especially given a high-dimensional dataset and the use of non-inferential methods, can likely aid toxicologists in having a better understanding of probable molecular targets as well as the complexity of perturbed networks - clearly, understanding BPAs effects will require a systems level approach (Hartung et al., 2017) as well as better characterization of genes that are not as yet confidently mapped as to biological function. More generally speaking, this points to the pitfall of trying to design "greener" substitutes (Maertens et al., 2014; Maertens and Hartung, 2018) in the absence of a clear, comprehensive understanding of molecular mechanism.

AUTHOR CONTRIBUTIONS

AM: main author of the paper. VT: support programming and data analysis. AK: planned the work and revised the manuscript. TH: mentoring, revision of the manuscript, and PI Human Toxome project laying the conceptual ground.

FUNDING

This work was supported by an NIH Transformational Research Grant, "Mapping the Human Toxome by Systems Toxicology" (RO1 ES 020750). VT was supported by NIEHS training grant (T32 ES007141). This work was also supported by the EU-ToxRisk project (An Integrated European "Flagship" Program Driving Mechanism-Based Toxicity Testing and Risk Assessment for the 21st Century) funded by the European Commission under the Horizon 2020 program (Grant Agreement No. 681002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00508/full#supplementary-material

FIGURE S1 | (A) Estrogen and **(B)** BPA Network dendrogram.

FIGURE S2 | Methylation pattern for FIZ1 in BRCA TCGA data set.

TABLE S1 | Protein-protein interaction and enrichment from STRING database.

TABLE S2 | Transcription factors for genes present only in low-dose BPA network.

REFERENCES

Academics Urge Caution in Interpreting Clarity-Bpa Results (n.d.). *Chemical Watch*. Available at: https://chemicalwatch.com/64449/academics-urge-caution-in-interpreting-clarity-bpa-results [accessed July 13, 2018].

Ali, S., Steinmetz, G., Montillet, G., Perrard, M., Loundou, A., Durand, P., et al. (2014). Exposure to low-dose bisphenol A impairs meiosis in the rat seminiferous tubule culture model: a physiotoxicogenomic approach. *PLoS One* 9:e106245. doi: 10.1371/journal.pone.0106245

Boucher, J. G., Husain, M., Rowan-Carroll, A., Williams, A., Yauk, C. L., and Atlas, E. (2014). Identification of mechanisms of action of bisphenol a-induced human preadipocyte differentiation by transcriptional profiling. *Obesity* 22, 2333–2343. doi: 10.1002/oby.20848

Bouhifd, M., Andersen, M. E., Baghdikian, C., Boekelheide, K., Crofton, K. M., Fornace, A. J., et al. (2015). The human toxome project. *ALTEX* 32:112. doi: 10.14573/altex.1502091

Calafat, A. M., Ye, X., Wong, L., Reidy, J. A., and Needham, L. L. (2008). Exposure of the U.S. population to bisphenol A and 4-tertiary-octylphenol: 2003-2004. *Environ. Health Perspect.* 116, 39–44. doi: 10.1289/ehp.10753

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14:128. doi: 10.1186/1471-2105-14-128

Chen, Y., Zhang, B., Bao, L., Jin, L., Yang, M., Peng, Y., et al. (2018). ZMYND8 acetylation mediates HIF-dependent breast cancer progression and metastasis. *J. Clin. Invest.* 128, 1937–1955. doi: 10.1172/JCI95089

Davis, S., and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi: 10.1093/bioinformatics/btm254

Delfosse, V., Grimaldi, M., Pons, J., Boulahtouf, A., le Maire, A., Cavailles, V., et al. (2012). Structural and mechanistic insights into bisphenols action provide guidelines for risk assessment and discovery of bisphenol A substitutes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14930–14935. doi: 10.1073/pnas.1203574109

Dong, S., Zhang, Z., and Takahara, H. (2007). Estrogen-Enhanced Peptidylarginine Deiminase Type IV Gene (PADI4) Expression in MCF-7 Cells Is Mediated by Estrogen Receptor-α-Promoted Transfactors Activator Protein-1. Nuclear Factor-Y and Sp1. *Mol. Endocrinol.* 21, 1617–1629. doi: 10.1210/me.2006-0550

Frietze, S., O'Geen, H., Littlepage, L. E., Simion, C., Sweeney, C. A., Farnham, P. J., et al. (2014). Global analysis of ZNF217 chromatin occupancy in the breast cancer cell genome reveals an association with ERalpha. *BMC Genomics* 15:520. doi: 10.1186/1471-2164-15-520

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80. doi: 10.1186/gb-2004-5-10-r80

Hartung, T., FitzGerald, R. E., Jennings, P., Mirams, G. R., Peitsch, M. C., Rostami-Hodjegan, A., et al. (2017). Systems toxicology: real world applications and opportunities. *Chem. Res. Toxicol.* 30, 870–882. doi: 10.1021/acs.chemrestox.7b00003

Horard, B., and Vanacker, J. (2003). Estrogen receptor-related receptors: orphan receptors desperately seeking a ligand. *J. Mol. Endocrinol.* 31, 349–357. doi: 10.1677/jme.0.0310349

Kleensang, A., Maertens, A., Rosenberg, M., Fitzpatrick, S., Lamb, J., Auerbach, S., et al. (2014). t4 workshop report: Pathways of Toxicity. *ALTEX* 31, 53–61. doi: 10.14573/altex.1309261

Kleensang, A., Vantangoli, M. M., Odwin-DaCosta, S., Andersen, M. E., Boekelheide, K., Bouhifd, M., et al. (2016). Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function. *Sci. Rep.* 6:28994. doi: 10.1038/srep28994

Koch, A., De Meyer, T., Jeschke, J., and Van Criekinge, W. (2015). MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data. *BMC Genomics* 16:636. doi: 10.1186/s12864-015-1847-z

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., et al. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nat.Commun.* 9, 1366. doi: 10.1038/s41467-018-03751-6

Lachmann, A., Xu, H., Krishnan, J., Berger, S. I., Mazloom, A. R., and Ma'ayan, A. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* 26, 2438–2444. doi: 10.1093/bioinformatics/btq466

Langfelder, P., and Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* 1:54. doi: 10.1186/1752-0509-1-54

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

LaPensee, E. W., Tuttle, T. R., Fox, S. R., and Ben-Jonathan, N. (2009). } % The entry below contains non-ASCII chars that could not be converted % to a LaTeX equivalent.). Bisphenol A at Low Nanomolar Doses Confers Chemoresistance in Estrogen Receptor-α–Positive and –Negative Breast Cancer Cells. *Environ. Health Perspect.* 117, 175–180. doi: 10.1289/ehp.11788

MacKay, H., and Abizaid, A. (2018). A plurality of molecular targets: The receptor ecosystem for bisphenol-A (BPA). *Horm. Behav.* 101, 59–67. doi: 10.1016/j.yhbeh.2017.11.001

Maertens, A., Anastas, N., Spencer, P. J., Stephens, M., Goldberg, A., and Hartung, T. (2014). Green toxicology. *ALTEX* 31, 243–249. doi: 10.14573/altex.1406181

Maertens, A., and Hartung, T. (2018). Green Toxicology-know early about and avoid toxic product liabilities. *Toxicol. Sci.* 161, 285–289. doi: 10.1093/toxsci/kfx243

Maertens, A., Luechtefeld, T., Kleensang, A., and Hartung, T. (2015). MPTP's pathway of toxicity indicates central role of transcription factor SP1. *Arch. Toxicol.* 89, 743–755. doi: 10.1007/s00204-015-1509-6

Miller, M. M., McMullen, P. D., Andersen, M. E., and Clewell, R. A. (2017). Multiple receptors shape the estrogen response pathway and are critical considerations for the future of in vitro-based risk assessment efforts. *Crit. Rev. Toxicol.* 47, 564–580. doi: 10.1080/10408444.2017.1289150

Okada, H., Tokunaga, T., Liu, X., Takayanagi, S., Matsushima, A., and Shimohigashi, Y. (2008). Direct evidence revealing structural elements essential for the high binding ability of bisphenol A to human estrogen-related receptor-gamma. *Environ. Health Perspect.* 116, 32–38. doi: 10.1289/ehp.10587

Park, B., Rhee, D., and Pyo, S. (2012). Apoptotic mechanism of bisphenol a in human neuroblastoma. *FASEB J.* 26.

Pendse, S. N., Maertens, A., Rosenberg, M., Roy, D., Fasani, R. A., Vantangoli, M. M., et al. (2017). Information-dependent enrichment analysis reveals time-dependent transcriptional regulation of the estrogen pathway of toxicity. *Arch. Toxicol.* 91, 1749–1762. doi: 10.1007/s00204-016-1824-6

Porreca, I., Ulloa Severino, L., D'Angelo, F., Cuomo, D., Ceccarelli, M., Altucci, L., et al. (2016). "Stockpile" of Slight Transcriptomic Changes Determines the Indirect Genotoxicity of Low-Dose BPA in Thyroid Cells. *PLoS One* 11:e0151618. doi: 10.1371/journal.pone.0151618

Pubmed Bisphenol A (n.d.). *PubMED*. Available at: https://www.ncbi.nlm.nih.gov/pubmed/?term = Bisphenol + A [accessed January 7, 2018].

Rochester, J. R., and Bolden, A. L. (2015). Bisphenol and F: a systematic review and comparison of the hormonal activity of bisphenol a substitutes. *Environ. Health Perspect.* 123, 643–650. doi: 10.1289/ehp.1408989

Rubin, B. S. (2011). Bisphenol A: an endocrine disruptor with widespread exposure and multiple effects. *J. Steroid Biochem. Mol. Biol.* 127, 27–34. doi: 10.1016/j.jsbmb.2011.05.002

Safe, S. H., Pallaroni, L., Yoon, K., Gaido, K., Ross, S., and McDonnell, D. (2002). Problems for risk assessment of endocrine-active estrogenic compounds. *Environ. Health Perspect.* 110(Suppl. 6), 925–929. doi: 10.1289/ehp.02110s6925

Shen, H., Xu, W., Guo, R., Rong, B., Gu, L., Wang, Z., et al. (2016). Suppression of enhancer overactivation by a RACK7-histone demethylase complex. *Cell* 165, 331–342. doi: 10.1016/j.cell.2016.02.064

Shioda, T., Rosenthal, N. F., Coser, K. R., Suto, M., Phatak, M., Medvedovic, M., et al. (2013). Expressomal approach for comprehensive analysis and visualization of ligand sensitivities of xenoestrogen responsive genes. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16508–16513. doi: 10.1073/pnas.1315929110

Singh, S., and Li, S. S. (2012). Epigenetic effects of environmental chemicals bisphenol A and phthalates. *Int. J. Mol. Sci.* 13, 10143–10153. doi: 10.3390/ijms130810143

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Teeguarden, J., Hanson-Drury, S., Fisher, J. W., and Doerge, D. R. (2013). Are typical human serum BPA concentrations measurable and sufficient to be estrogenic in the general population? *Food Chem. Toxicol.* 62, 949–963. doi: 10.1016/j.fct.2013.08.001

TOXNET (n.d.). *TOXNET Databases*. Available at: https://toxnet.nlm.nih.gov/cgi-bin/sis/search2/r?dbs + hsdb:@term + @DOCNO + 513[accessed July 11, 2018].

Vendrell, J. A., Thollet, A., Nguyen, N. T., Ghayad, S. E., Vinot, S., Bi\'eche, I., et al. (2012). ZNF217 is a marker of poor prognosis in breast cancer that drives epithelial-mesenchyme transition and invasion. *Cancer Res.* 72, 3593–3606. doi: 10.1158/0008-5472.CAN-11-3095

Wolf, I., and Rohrschneider, L. R. (1999). Fiz1, a novel zinc finger protein interacting with the receptor tyrosine kinase Flt3. *J. Biol. Chem.* 274, 21478–21484. doi: 10.1074/jbc.274.30.21478

Woodfield, G. W., Horan, A. D., Chen, Y., and Weigel, R. J. (2007). TFAP2C controls hormone response in breast cancer cells through multiple pathways of estrogen signaling. *Cancer Res.* 67, 8439–8443. doi: 10.1158/0008-5472.CAN-07-2293

Yip, A. M., and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 8:22. doi: 10.1186/1471-2105-8-22

Zhang, X., Gamble, M. J., Stadler, S., Cherrington, B. D., Causey, C. P., Thompson, P. R., et al. (2011). Genome-wide analysis reveals PADI4 cooperates with Elk-1 to activate c-Fos expression in breast cancer cells. *PLoS Genet.* 7:e1002112. doi: 10.1371/journal.pgen.1002112

Zimmerman, J. B., and Anastas, P. T. (2015). Chemistry. Toward substitution with no regrets. *Science* 347, 1198–1199. doi: 10.1126/science.aaa0812

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 November 2018
doi: 10.3389/fgene.2018.00527

Embracing the Dark Side: Computational Approaches to Unveil the Functionality of Genes Lacking Biological Annotation in Drug-Induced Liver Injury

Terezinha Souza[1†], Panuwat Trairatphisan[2†‡], Janet Piñero[3†], Laura I. Furlong[3], Julio Saez-Rodriguez[2,4‡], Jos Kleinjans[1] and Danyel Jennen[1]*

[1] Department of Toxicogenomics, GROW School for Oncology and Developmental Biology, Maastricht University, Maastricht, Netherlands, [2] Joint Research Center for Computational Biomedicine (JRC-COMBINE), Faculty of Medicine, RWTH Aachen University, Aachen, Germany, [3] Integrative Biomedical Informatics Group, Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences (DCEXS), Hospital del Mar Medical Research Institute (IMIM), Universitat Pompeu Fabra, Barcelona, Spain, [4] European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Cambridge, United Kingdom

In toxicogenomics, functional annotation is an important step to gain additional insights into genes with aberrant expression that drive pathophysiological mechanisms. Nevertheless, there exists a gap on annotation of these genes which often hampers the interpretation of results and limits their applicability in translational medicine. In this study, we evaluated the coverage of functional annotations of differentially expressed genes (DEGs) induced by 10 selected compounds from the TG-GATEs database identified as high- or no-risk in causing drug-induced liver injury (most-DILI or no-DILI, respectively) using *in vitro* human data. Functional roles of DEGs not present in the most common biological annotation databases – termed "dark genes" – were unveiled via literature mining and via the identification of shared regulatory transcription factors or signaling pathways. Our results demonstrated that there were approximately 13% of dark genes induced by these compounds *in vitro* and we were able to obtain additional relevant information for up to 76% of those. Using interactome data from several sources, we have uncovered genes such as *LRBA*, and *WDR26* as highly connected in the protein network that play roles in drug response. Genes such as *MALAT1, H19,* and *MIR29C* – whose links to hepatotoxicity have been confirmed – were identified as markers for the most-DILI group and appeared as top hits across all literature-based mining methods. Furthermore, we investigated the potential impact of dark genes on liver toxicity by identifying their rat orthologs in combination with their correlation to drug-induced liver pathologies observed *in vivo* following chemical exposure. We identified a set of important regulatory transcription factors of dark genes for all most-DILI compounds including E2F1 and JUND with supporting evidences in literature and we found *Magee1* correlated with chemically induced bile duct hyperplasia and adverse responses at

OPEN ACCESS

Edited by:
Pierre R. Bushel, National Institute of Environmental Health Sciences (NIEHS), United States

Reviewed by:
Mohamed Diwan M. AbdulHameed, Independent Researcher, Frederick, United States
Terrence Furey, The University of North Carolina at Chapel Hill, United States

***Correspondence:**
Terezinha Souza
terezinhamsouza@gmail.com

† These authors have contributed equally to this work

‡ Present address:
Panuwat Trairatphisan and Julio Saez-Rodriguez, Institute of Computational Biomedicine, Faculty of Medicine, Heidelberg University, Heidelberg, Germany

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Genetics

Received: *12 July 2018*
Accepted: *19 October 2018*
Published: *20 November 2018*

Citation:
Souza T, Trairatphisan P, Piñero J, Furlong LI, Saez-Rodriguez J, Kleinjans J and Jennen D (2018) Embracing the Dark Side: Computational Approaches to Unveil the Functionality of Genes Lacking Biological Annotation in Drug-Induced Liver Injury. Front. Genet. 9:527. doi: 10.3389/fgene.2018.00527

29 days in rats *in vivo*. In conclusion, in this study we show the potential role of these poorly annotated genes in mechanisms underlying hepatotoxicity and offer a number of computational approaches that may help to minimize current gaps in gene annotation and highlight their values as potential biomarkers in toxicological studies.

Keywords: annotation, DILI, gene ontology, text mining, network biology, translational bioinformatics

INTRODUCTION

In the field of toxicogenomics, various computational approaches have been developed and upgraded over the years. Nowadays, the most commonly applied method consists of the use of differential analysis, i.e., the application of statistical approaches to identify and biologically annotate differentially expressed genes (DEGs) upon compounds' perturbation (Khatri et al., 2012; Souza et al., 2016). Genome-wide, unsupervised methods such as gene set enrichment analysis (GSEA), biclustering and weighted co-expression analysis (WGCNA) can be used to identify gene sets associated with specific phenotypes (AbdulHameed et al., 2014; Tawa et al., 2014; Sutherland et al., 2016). Another branch of methods includes network-based analyses such as the clustering of gene sets based on their centrality in molecular networks (Kotlyar et al., 2012), as well as mechanistic modeling in smaller scales such as Boolean logic modeling (Zhang J.D. et al., 2014) and ordinary differential equation (ODE)-based models (Hendrickx et al., 2017)– the latter providing dynamical information of the systems in a more refined granularity.

An important bottleneck across all methodologies, however, is the biological annotation of the gene sets. This biological annotation is provided by collections of pathways or gene sets stored in popular knowledge-driven resources such as Reactome (Fabregat et al., 2018) and the Gene Ontology (The Gene Ontology Consortium, 2017). Despite the ever-increasing amount of information deposited in pathway knowledge databases, gaps on functional protein interaction and other types of biological annotation still exist. In addition, a large number of non-coding genes, i.e., small- and long- non-coding genes and pseudogenes, covering around 37,000 molecular entities whose biological roles elucidation is an ongoing task. The "biological process" branch of the Gene Ontology (GO BP) is one of the most commonly used sources of biological annotations. Nevertheless, GO BP terms only cover 33% (19,691 genes) from the entire human genome (estimated in approximately 60,200 genes according to NCBI's gene annotation) (Brown et al., 2015). On the pathway side, high-confidence databases such as Reactome comprise only around half of all human protein-coding genes (10,762 genes) (Fabregat et al., 2018) while low-confidence high-coverage databases such as Pathway Commons coverage for coding and non-coding portions of the genome is around 38% (22,754 genes). Furthermore, most common pathway resources only cover information regarding protein coding genes, while the role of non-coding RNAs (ncRNAs) in processes such as disease or drug response, remains uncovered. We argue here that these missing entities should not be neglected due to their potential biological functionality with respect to human health.

Community-based efforts can help to fill this gap. An example of this is the creation of GeneRIF (Mitchell et al., 2003), a platform to share short functional descriptions of genes which are generally observed by experimentalists. Such a database allows users to rapidly scan through the additional functional information on genes of interest which are stored in a standardized format. In parallel, user-friendly text mining tools that allow automatic retrieval of information about gene function from the literature have been developed. One such tool is PubTator (Wei et al., 2013), which supports manual literature curation besides offering a collection of annotated abstracts, including relationships among diseases, genes, and drugs. In addition, even if genes are not annotated for their biological processes, they can still be linked to verified disease signatures with, e.g., DisGeNET (Piñero et al., 2017).

Besides text mining, various emerging computational approaches in Systems Biology have been developed with high potential to be applied for unveiling the functional roles of genes. For instance, the inference of transcription factor (TF) activities based on gene expression data may reflect the common regulatory patterns of signaling pathways which are shared among downstream targets with or without functional annotation (Alvarez et al., 2016; Garcia-Alonso et al., 2018). In parallel, the activity of regulatory signaling pathways can be independently predicted by computational approaches based on the expression of genes that reflect the activities of the respective pathway upon perturbation, thus highlighting possible involvement of signaling modulation via unannotated genes (Tarca et al., 2009; Khatri et al., 2012; Schubert et al., 2018). By investigating the list of genes with unknown function which were applied to derive transcription factors' activities and signaling pathways' signatures, one could infer their biological functions associated to the role of the predicted upstream regulatory modules.

Recently, Sutherland et al. (2016, 2017) have shown that gene expression in chemically exposed rats coalesce into groups of co-expressed genes (i.e., modules) – some of which appear to be correlated to phenotypes indicative of toxicity or adverse outcomes. Interestingly, this approach highlighted branches comprising a number of modules of interest with little or no biological annotation, some of which containing ncRNAs. Their roles in cellular functioning and disease are slowly being elucidated (Luo et al., 2016; Xu et al., 2017), but their modulation upon drug exposure remains largely uncovered. In spite of that, toxicologists have pointed that their involvement in apical effects should be investigated and considered in regulatory frameworks, i.e., mode-of-action (MoA) and adverse outcome pathway (AOP) analyses (Aigner et al., 2016). Studies to unveil the functionality

of these poorly annotated genes are therefore necessary to generate potentially novel biomarkers to improve risk assessment during the preclinical phase. In addition, connecting the poorly annotated genes to the pathological outcomes of rodent studies will further aid to identify their function. Therefore, the identification of human orthologs is imperative to allow and improve translation of the rodent data to the human context.

Therefore, in this work we aim to assess the coverage of the current functional annotation of genes represented in public databases using toxicogenomics sets; those not found in these representative biological annotation databases were coined "dark genes" in this study. Our second goal is to (a) estimate the relevance for cellular functions of dark genes involved in drug response, and (b) assign putative functions to them. For the first task, we assess the presence of these genes in human interactomes built from several sources, in literature-based resources and their association to diseases. For the second, we employed computational approaches to identify (i) common regulatory transcription factors and (ii) signaling pathways' signatures which are shared between annotated and unannotated genes. Finally, we examine these chemical-induced changes in the light of toxicity and as potential markers of drug-induced liver injury (DILI) given their regulation in human *in vitro* and associations to pathological responses in rat *in vivo*.

MATERIALS AND METHODS

Compound Selection

In order to obtain robust modulation of genes and minimize noisy expression, we opted for analyzing inducible responses across multiple compounds. To investigate whether gene modulation of entities of interest is associated with distinct toxicities, we created two equally sized groups of chemicals to avoid sample bias, selected according to their current classification as agents involved in human DILI. For this, we used a classification based on weight of evidence of causality (DILIRank) (Chen et al., 2016), which categorize compounds in three main classes: most-DILI (drugs withdrawn or with severe DILI indication), less-DILI (drugs with mild DILI indication or adverse reactions) and no-DILI. Here, we selected compounds available on TG-GATEs either classified as most-DILI (acetaminophen, diclofenac, isoniazid, nimesulide, and valproic acid) and no-DILI (caffeine, chloramphenicol, chlorpheniramine, hydroxyzine, and theophylline) to enable an unambiguous separation of gene modulation responses. Further information on the compounds and classification proposed by Chen et al. (2016) can be found in **Supplementary Table S1**.

Gene Expression Data: Processing and Differential Gene Expression

Gene expression data were obtained from TG-GATEs[1] (Igarashi et al., 2015). Raw data files generated *in vitro* from primary human hepatocytes from each compound selected were processed (quality control, background correction, RMA normalization) using the R package affy (Gautier et al., 2004). Genes were annotated with a customCDF (v. 19) with Entrez gene identifiers for Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays. Here, we opted for a traditional approach (i.e., comparison of treated vs. control mean expression) to obtain DEGs; to obtain maximal transcriptional response, we selected the highest dose and latest time point (24 h) from each compound. Differential expression analysis was then performed on each set using the R package LIMMA and comparing to time-matched controls from each compound treatment. DEGs were selected based on their significance after multiple testing correction (false discovery rate, FDR) and an absolute fold change of 1.5 (equivalent to log2 fold change of 0.585) with FDR < 0.05.

Coverage of Biological Annotation Across Databases

To compute the number of DEGs that were not included in the most commonly used resources in the field of toxicology and network biology, we downloaded the files from Gene Ontology[2] (The Gene Ontology Consortium, 2017), Reactome[3] (Fabregat et al., 2018), MSigDB (Liberzon et al., 2015) curated pathways[4], Pathway Commons[5] (Cerami et al., 2011), and OmniPath[6] (Türei et al., 2016) on May, 2018.

We mapped the gene symbols to Entrez gene identifiers using the file http://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz downloaded on April, 2018. For those genes for which we could not find an Entrez gene identifier, we used the correspondence between UniProt identifiers and Entrez gene identifiers from the file http://ftp.ebi.ac.uk/pub/databases/genenames/new/tsv/hgnc_complete_set.txt downloaded on May, 2018. From the Gene Ontology file, we only took into account the GO BP branch as this branch provides a better insight into the biological mechanisms compared to molecular function (MF) and cellular component (CC). From Pathway Commons, we removed interactions without pathway annotations. From OmniPath, we removed interactions that were supported only by protein-protein interaction databases (BioGRID, HPRD, and IntAct). A DEG was tagged as dark gene if it was absent in the pathway databases and GO BP branch.

Furthermore, to assess the global coverage of the biological annotations, the same steps were performed to categorize all genes measured within the Affymetrix array platform.

Protein Interaction Networks

We built four protein interaction networks (PINs) using data from the most comprehensive, and updated databases: INBIOMAP (Li et al., 2017), HIPPIE (Alanis-Lobato et al., 2017),

[1] http://toxico.nibiohn.go.jp/english/datalist.html

[2] http://geneontology.org/gene-associations/goa_human.gaf.gz

[3] https://reactome.org/download/current/NCBI2Reactome_All_Levels.txt

[4] http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C2

[5] http://www.pathwaycommons.org/archives/PC2/v10/PathwayCommons10.All.hgnc.txt.gz

[6] http://omnipathdb.org/interactions/?fields=sources&fields=references

BIANA (Garcia-Garcia et al., 2010), and IntAct (Orchard et al., 2014).

To build a HIPPIE-based network, we downloaded the file http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/hippie_current.txt on January, 2018. In the case of INBIOMAP, we downloaded the file from https://www.intomics.com/inbio/map/#downloads. We removed predicted interactions. To build an interactome from BIANA, we downloaded the *Homo sapiens* data from http://sbi.imim.es/web/GUILDify2.php/downloads on January, 2018. For IntAct, we downloaded the file http://ftp.ebi.ac.uk/pub/databases/intact/current/all.zip on October, 2017.

Literature-Based Resources

To provide further insight on the relevance of the role of the dark genes, we checked if they were involved in human diseases using DisGeNET data, version 5 (Piñero et al., 2017). Additionally, we assessed the presence of dark genes in the scientific literature. For that goal we used GeneRIF (Mitchell et al., 2003), that describe in a short phrase (less than 25 characters in length) the function or functions of a gene, and PubTator (Wei et al., 2013), a web tool that supports manual literature curation using text-mining techniques.

GeneRIFs were downloaded from http://ftp.ncbi.nih.gov/gene/GeneRIF/generifs_basic.gz and PubTator data was downloaded from http://ftp.ncbi.nlm.nih.gov/pub/lu/PubTator/gene2pubtator.gz and http://ftp.ncbi.nlm.nih.gov/pub/lu/PubTator/bioconcepts2pubtator.gz on January 2018.

Identification of Common Regulatory Transcription Factors and Signaling Pathways

The list of dark genes was mapped to the list of transcription factors and their regulated genes ("regulons") from the tool DoRothEA (Garcia-Alonso et al., 2018) and to the list of gene signatures used for the inference of signaling pathways' activities from the tool PROGENy (Schubert et al., 2018). The mapping was classified and compared according to the group of compounds. The shared common transcription factors and signaling pathways in each group were intersected to derive the most representative proxies which represent the corresponding dark genes. Venn diagrams of these results as the ones from PINs (see section "Protein Interaction Networks") were generated with the following web tool: http://bioinformatics.psb.ugent.be/webtools/Venn.

Comparison to Weighted Gene Co-expression Network Analysis (WGCNA) Modules

Co-expression analyses aim to obtain significant relationships among genes showing similar patterns of expression across samples. The resulting gene sets (also known as modules) are useful for reducing dimensionality and correlating molecular changes to an observed phenotype. Since clusters are generated in an unbiased manner, it is possible to identify modules encompassing genes with multiple levels of biological annotation (e.g., GO terms or pathways).

To investigate the relevance of these dark genes in an animal model and its implications in adverse outcomes, we identified rat orthologs of the dark genes present in co-expression modules detected in Sutherland et al. (2017). The rat orthologs to human genes were then mapped to modules identified using the annotation available in the Rat Genome Database (rgd.mcw.edu). From there, modules associated with pathological outcomes and underlying GO BPs were further investigated.

RESULTS

Compound-Induced Gene Expression

The number of DEGs modulated by each compound can be found in **Table 1**. By merging the DEGs groupwise, a total of 5,446 and 3,845 genes were found to be induced by most-DILI and no-DILI groups, respectively, comprising in total 6,918 unique genes. These genes were classified using the Ensembl gene annotation information, which showed that the majority of all genes identified were protein coding (95%), followed by non-coding RNA (ncRNA, 4.2%), pseudogenes, snoRNA and others (less than 1% each). An overview of the number of DEGs shared by compounds from the same DILI risk group can be found in the **Supplementary Table S1**.

Biological Annotation and Gene Annotation of Dark Genes

Among the 6,918 genes deemed significantly affected by chemical exposure, 916 genes (~13%) were not included in any biological pathway or process. This number is lower than the number of genes in the array lacking this type of annotation, identified as 22% (4,210 out of 19,441 genes). In total, 760 out of 916 entities were categorized into gene types based on Ensembl annotation; the majority of those is considered protein coding (**Table 2**). A detailed description of gene types from the array and modulated by chemicals can be found in the **Supplementary Table S1**. A comparison of database coverage can be found in **Supplementary Date Sheet S1**. In addition, a comprehensive list encompassing gene modulation per compound/DILI risk group, as well as pathway and GO annotation and results from the methodologies applied for annotation of the dark genes

TABLE 1 | Number of differentially expressed genes (DEGs, absolute FC > 1.5 and FDR < 0.05) of compounds from most-DILI and no-DILI groups.

Most-DILI	Number of DEGs	No-DILI	Number of DEGs
Acetaminophen	2,280	Caffeine	2,316
Diclofenac	1,888	Chloramphenicol	108
Isoniazid	1,024	Chlorpheniramine	93
Nimesulide	1,697	Hydroxyzine	815
Valproic acid	2,290	Theophylline	2,918
Total unique DEGs	**5,446**	**Total unique DEGs**	**3,845**

TABLE 2 | Classification of genes without GO BP annotation and absent on Reactome, MSigDB, OmniPath, and Pathway Commons databases (dark genes) modulated by compounds from most-DILI and no-DILI groups.

Gene type	Array dark genes	Most-DILI	No-DILI	Dark DEGes
Protein coding	1,756	444	278	567
Antisense RNA	527	69	33	78
lincRNA	722	53	33	63
Processed transcript	113	11	8	15
Pseudogenes[1]	56	18	14	25
snoRNA	8	4	3	5
Sense intronic	25	3	1	3
Sense overlapping	10	1	1	2
miRNA	3	1	0	1
TEC[2]	11	1	1	1
Total	**3231**	**605**	**372**	**760**

[1]Pseudogenes from the categories "transcribed unprocessed pseudogene," "transcribed unitary pseudogene" and "transcribed processed pseudogene." [2]TEC: to be experimentally confirmed.

is available as **Supplementary Table S1** while an overview of gene modulation shared across compounds from each group is available in **Supplementary Data Sheet S1**.

Characterization of Dark Genes in the Human Interactome

Furthermore, we investigated the coverage of the dark genes in four different sources of human protein–protein interactions. We found 492, 420, 475, and 285 dark genes included in HIPPIE, IntAct, Inbiomap and Biana interactomes, respectively. Among them, 536 dark genes were present in at least one of these resources, while 268 were included in all four resources. The overlaps can be found in **Supplementary Data Sheet S1**.

We further characterized the dark genes present in the interactomes. **Figure 1** shows histograms of the degree distribution of the dark genes in each interactome. A large fraction of the dark genes has low connectivity in all four interactomes, although there are some genes with relatively high degrees. Some examples of these latter genes, more connected than the rest of dark genes in the four interactomes, are shown in **Table 3**.

Literature Mining: Disease Association, GeneRIF, and PubTator

We also evaluated other literature-based resources containing functional information. First, we used DisGeNET v5.0 to determine whether the dark genes are associated to human diseases. We found 60 dark genes with disease annotations reported by curated databases, and 255 dark genes in DisGeNET ALL dataset, which also includes the results from automatic text mining in the scientific literature. The top genes with disease annotations in the curated data in DisGeNET are shown in **Table 4**. The diseases in which these genes were more frequently involved were different types of neoplasms, although they seem to play a role in a wide variety of diseases, and abnormal phenotypes (**Supplementary Table S1**).

We also evaluated the coverage of the dark genes in GeneRIF which contains users-submitted compact information regarding the function of the genes. We found 356 dark genes with GeneRIF annotations. Twenty-three dark genes had 10 or more GeneRIFs, and among those, several ncRNAs (**Table 5**). Some relevant examples of the GeneRIFs for *MALAT1* are "*MALAT1* level is associated with liver damage, and has clinical utility for predicting development of hepatocellular carcinoma" or "observations suggest that *MALAT1* promotes hepatic steatosis and insulin resistance by increasing nuclear SREBP-1c protein stability."

A similar exercise was performed using PubTator to obtain additional information with a unbiased text-mining approach. We found that 550 dark genes matched the entries in PubTator. Interestingly, the two genes with the highest number of hits were, again, two long non-coding RNAs, *MALAT1* and *H19* (**Table 5**), with over 1,000 papers each. In some cases a single entry on PubTator was a match for multiple hits, as for instance "Central role of the p53 pathway in the non-coding-RNA response to oxidative stress," which related *MALAT1*, *NEAT1,* and *PVT1* (3 dark ncRNAs) to oxidative stress produced by H2O2 (Fuschi et al., 2017).

Mapping Functional Information of the Dark Genes With Common Regulatory TF and Signaling Pathways

By mapping the DEGs of the selected compounds, we found that about 16% of dark genes are the targets genes of regulatory TFs in DoRothEA (**Table 6**). The intersections of regulatory TFs between most-DILI and no-DILI compounds are shown in **Figure 2**. Here, the most representative TFs for most-DILI group overlapped across all five compounds (n = 14) were AR, E2F1, E2F4, ETS1, FOXA1, FOXP3, GATA1, GATA2, GATA3, HNF4A, JUND, REST, SPI1, and TFAP2C, while the most representative for non-DILI group shared by all five compounds (n = 1) was GATA2.

In parallel, we found that about 4% of dark genes can be grouped together with the gene signatures used for the inference of signaling pathways' activities in PROGENy (**Table 6**). The most representative signaling pathways overlapped among all five most-DILI compounds (n = 2) were Hypoxia and PI3K, whereas TNF-alpha was the most representative one for non-DILI compounds (excluding chloramphenicol which did not have an enriched pathway), see **Figure 2**.

The scripts for all analyses conducted in this study are available in **Supplementary Data Sheets S2, S3**.

Rat Orthologs to Human Dark Genes in Co-expression Modules

Identification of rat orthologs to human dark genes and comparison to co-expression modules generated from rats exposed to chemicals showed that 544 human dark genes had an ortholog in rat and, from these, 241 were included in at least one WGCNA module. Among these genes, at least 20 comprised

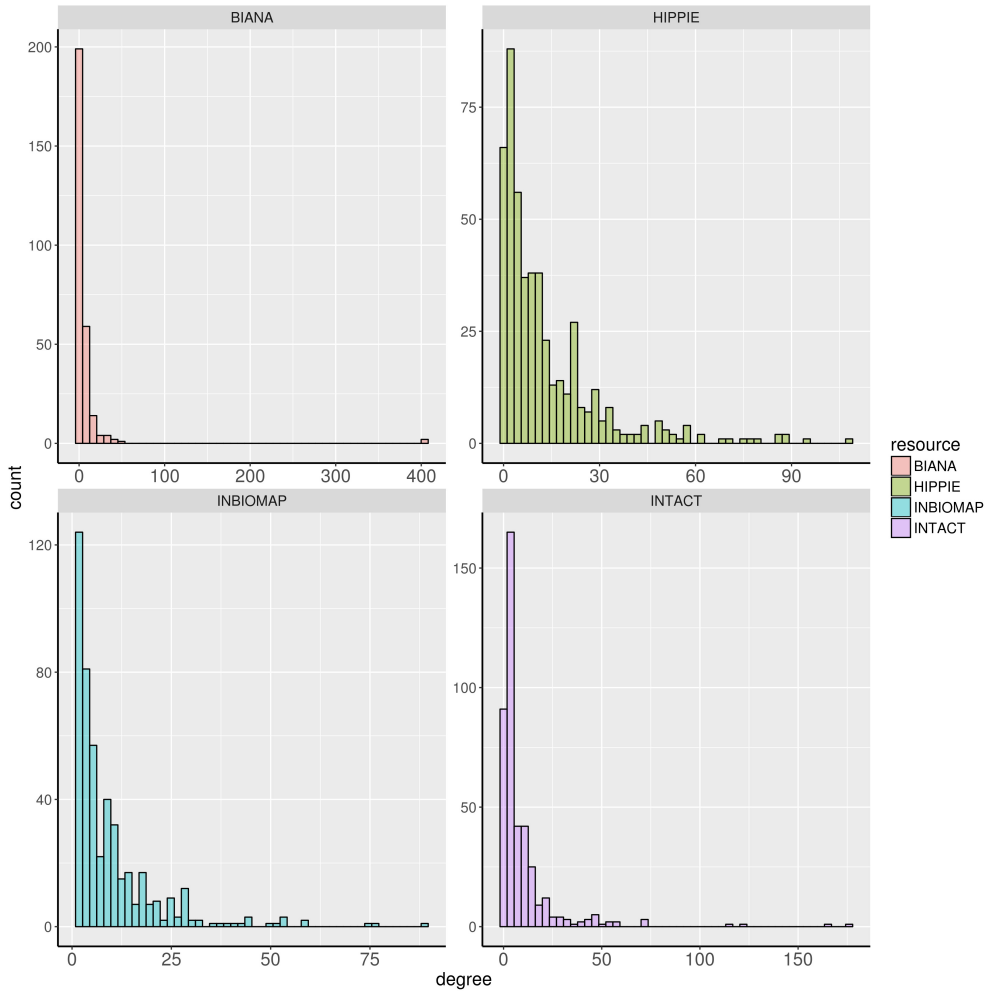


FIGURE 1 | Degree distribution of the dark genes in human interactome databases Biana, HIPPIE, Inbiomap, and IntAct.

those coding for transmembrane proteins (TMEM family). These dark genes were found in (1) modules from branches with global poor GO BP annotation (branches C.I and C.II indicated by Sutherland et al., 2017) and (2) modules associated with pathology. **Table 7** contains a list of dark gene orthologs whose modules were associated with specific pathologies and the underlying GO BP (whenever available). The complete list of dark genes orthologs mapped to modules can be found in **Supplementary Table S1**.

DISCUSSION

Pathway and network analyses are essential steps downstream to the identification of interesting features (e.g., differential analysis) in diverse fields of 'omics research. Despite advances in biological annotation of the human genome, there is still a considerable gap in knowledge, owed mainly to experimental evaluation of already well-studied entities, which hampers biomedical research (Haynes et al., 2018). In this study, we aimed to investigate these poorly annotated entities (coined dark genes) in the light of chemical exposure since many studies in mechanistic toxicology are heavily attached to biological roles and many genes with potential mechanistic and predictive roles may remain uncovered as a result.

From our analysis, we observed that approximately 13% of DEGs and 22% of all genes in the array were not mapped to GO BP, OmniPath, MSigDB, Reactome or Pathway Commons. This finding highlights that the issue with unannotated genes is generalized and the biological functions of a number of DEGs identified in gene expression studies remain to be uncovered. Genes with Ensembl classification were mostly categorized as protein coding (73%), while 8% of dark genes were classified as long-intergenic non-coding RNA (lincRNAs), which have increasing evidences to play a role in drug-induced organ toxicity (Zhou et al., 2015; Dempsey and Cui, 2017).

It was demonstrated that up to 59% of dark genes are present in at least one of the human interactome databases. Of these, a few have higher degree of connectivities to the other genes as shown in **Table 3**. In the context of drug development, PINs

TABLE 3 | Degree of connectivity for top 10 genes in human protein–protein interaction databases.

Gene symbol	Description	BIANA	HIPPIE	INBIOMAP	IntAct
RBM12	RNA binding motif protein 12	405	44	17	8
LRBA	LPS responsive beige-like anchor protein	402	28	16	9
SGTB	Small glutamine rich tetratricopeptide repeat containing beta	8	87	88	177
TMEM25	Transmembrane protein 25	3	85	77	2
FAM189A2	Family with sequence similarity 189 member A2	10	78	75	10
ZCCHC10	Zinc finger CCHC-type containing 10	51	52	53	59
C1orf109	Chromosome 1 open reading frame 109	45	51	53	116
TSSC4	Tumor suppressing subtransferable candidate 4	13	68	59	17
WDR26	WD repeat domain 26	3	79	51	33
FAM90A1	Family with sequence similarity 90 member A1	33	49	50	122

have been employed to understand the perturbations elicited by drug treatment in cellular processes, and to characterize drug targets (Yıldırım et al., 2007) and side effects (Wang et al., 2013). Recently, Piñero et al. (2018) has shown that within

TABLE 4 | Top 10 genes associated to diseases in DisGeNET (curated data).

Symbol	Description	Gene type	DILI risk group(s)	Number of diseases
CLIP2	CAP-Gly domain containing linker protein 2	Protein-coding	Most-DILI	141
IPW	Imprinted in Prader-Willi syndrome (non-protein coding)	ncRNA	Most-DILI, no-DILI	66
TGDS	TDP-glucose 4,6-dehydratase	Protein-coding	Most-DILI	62
LRBA	LPS responsive beige-like anchor protein	Protein-coding	Most-DILI	33
AMMECR1	Alport syndrome, mental retardation, midface hypoplasia and elliptocytosis chromosomal region gene 1	Protein-coding	Most-DILI, no-DILI	27
TMEM98	Transmembrane protein 98	Protein-coding	Most-DILI	9
H19	H19, imprinted maternally expressed transcript (non-protein coding)	ncRNA	Most-DILI	7
MALAT1	Metastasis associated lung adenocarcinoma transcript 1 (non-protein coding)	ncRNA	Most-DILI	7
WDR11	WD repeat domain 11	Protein-coding	Most-DILI	6
CMYA5	Cardiomyopathy associated 5	Protein-coding	no-DILI	3

TABLE 5 | Top 10 dark genes by number of GeneRIFs with their corresponding number of publications indexed on PubTator.

Symbol	Description	Gene Type	DILI risk group(s)	GeneRIFs	Number of publications
H19	H19, imprinted maternally expressed transcript (non-protein coding)	ncRNA	Most-DILI	193	1169
MALAT1	Metastasis associated lung adenocarcinoma transcript 1 (non-protein coding)	ncRNA	Most-DILI	156	1203
MIR29C	microRNA 29c	ncRNA	Most-DILI	77	234
UCA1	Urothelial cancer associated 1 (non-protein coding)	ncRNA	Most-DILI, no-DILI	63	152
NEAT1	Nuclear paraspeckle assembly transcript 1 (non-protein coding)	ncRNA	Most-DILI, no-DILI	56	223
PVT1	Pvt1 oncogene (non-protein coding)	ncRNA	Most-DILI	56	182
TUG1	Taurine up-regulated 1 (non-protein coding)	ncRNA	Most-DILI, no-DILI	41	99
MTUS1	Microtubule associated scaffold protein 1	Protein-coding	Most-DILI, no-DILI	26	71
TM4SF5	Transmembrane 4 L six family member 5	Protein-coding	Most-DILI, no-DILI	20	37
FAM167A	Family with sequence similarity 167 member A	Protein-coding	Most-DILI	19	32

TABLE 6 | Overview of mapped dark genes based on transcriptional regulation (DoRothEA) and on signaling pathway signatures (PROGENy).

Compound	Dark genes	Dark genes in DoRothEA	Number of mapped TFs	Dark genes in PROGENy	Number of mapped signaling pathways
Acetaminophen	294	46	24	11	8
Valproic acid	330	51	28	11	6
Isoniazid	152	22	21	8	5
Diclofenac	221	32	26	10	7
Nimesulide	145	34	26	5	4
Total Most-DILI	732	115	40	29	10
Theophylline	326	48	29	16	6
Caffeine	271	47	30	16	7
Hydroxyzine	81	17	19	3	3
Chloramphenicol	6	2	4	0	0
Chlorpheniramine	7	1	1	1	1
Total No-DILI	451	70	36	19	7

the set of drug targets, those that are related to side effects are more central in the interactome at local, global and meso-scale level. In the current study, we have used interactome data to highlight genes with strong molecular data, such as genes *LRBA*, which showed over 400 interaction partners in BIANA database, being associated to several diseases and involved in the response to DNA damage (Matsuoka et al., 2007). Another example is *WDR26* – with over 70 partners in HIPPIE database and also disease-associated, that has been found to protect cells from oxidative stress-induced apoptosis (Zhao et al., 2009). Furthermore, genes such as *MYO15B, BEX5, C12orf75,* and *SPATA2L,* that appear differentially expressed in at least 4 of the 5 DILI compounds and not perturbed upon no-DILI drugs, are also involved in protein-protein interactions according to most PPI databases, thus making them interesting potential DILI biomarker candidates to further pursue.

On the other hand, the use of text mining tools allowed to obtain information about non-coding RNAs – entities which are not included in PINs. With these methods we identified genes such as microRNA MIR29C, and non-coding RNAs H19 and MALAT1, all found exclusively in the most-DILI risk group. Deregulation of *H19* and *MALAT1* has been associated with liver disease (Takahashi et al., 2014). Downregulation of *H19*, which was consistently observed in all most-DILI compounds except nimesulide, has been associated with formation of Mallory-Denk bodies (MDBs), aggresomes of proteins found in many types of liver diseases (Oliva et al., 2009). Furthermore, downregulation of circulating microRNAs from the mir29 family were shown in liver cirrhosis patients (Loosen et al., 2017) and *MIR29C* in particular has been associated to acute and chronic

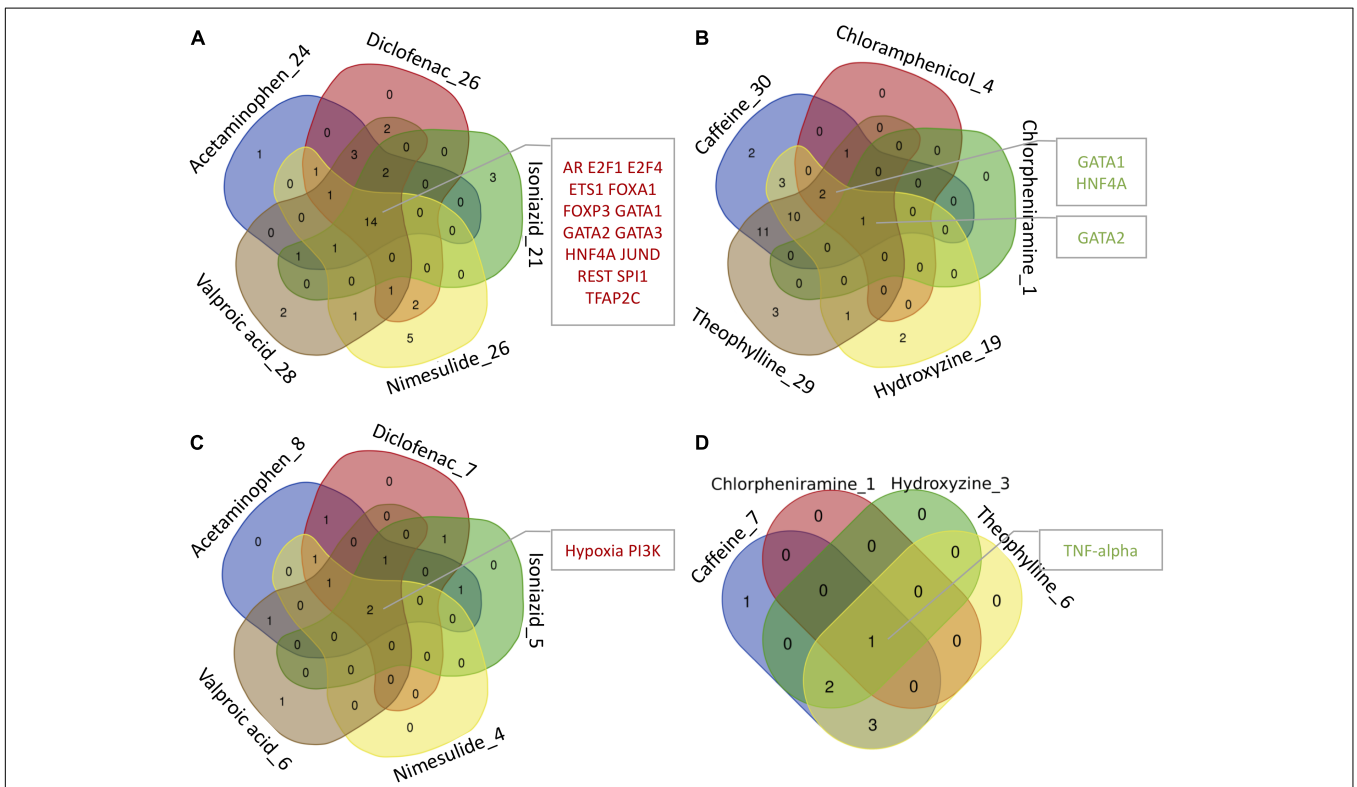


FIGURE 2 | Venn diagrams showing the intersection of transcription factors (TFs) and signaling pathways regulating at least one dark gene. The number accompanying each compound refers to the number of transcription factors and signaling pathways enriched by dark genes and the intersected modules by all or most of the compounds are highlighted in the adjacent boxes. **(A,B)** Regulatory TFs of hepatotoxic and non-hepatotoxic compounds, respectively. **(C,D)** Regulatory signaling pathways of hepatotoxic and non-hepatotoxic compounds, respectively. No enriched signaling pathway was found for Chloramphenicol (absent in **D**).

TABLE 7 | Orthologs to human dark genes present in modules associated to pathologies in rats described by Sutherland et al. (2017).

Module	Gene symbol	Pathology association	GO-BP
13m	*Smim14*	Adverse at 29 days, Hematopoiesis	Complement activation; Inflammatory response, Leukocyte chemotaxis
39	*Lhfpl6*	BDH	Extracellular matrix organization, Collagen fibril organization
205	*Thyn1*	BDH, Adverse at 29 days	Cellular response to DNA damage stimulus, Signal transduction by p53 class mediator
293	*Magee1*	BDH	–
55m	*Abracl*	Fibrosis, BDH, Necrosis	Membrane raft assembly, Regulation of cytoskeleton organization
14m	*Wdr70, Lyrm1, Tmem209*	Hypertrophy	Protein folding, tRNA metabolic process
10	*RGD1560010, Abhd8, Tbc1d31*	Increased mitosis	Cell cycle, Mitotic cell cycle
81	*Jpt1*	Increased mitosis, BDH	Actin polymerization or depolymerization
70	*Spata2l, Ubald1*	Single cell necrosis	Cell cycle arrest
309	*RGD1359127*	Single cell necrosis	–
147	*Oser1*	Single cell necrosis	–
27m	*C2cd2*	Vacuolation	

BDH, bile duct hyperplasia.

models of hepatotoxicity (Schueller et al., 2018). The relevance of these genes in diseases, in particular liver diseases, was demonstrated in the disease association analysis with DisGeNET (**Figure 3**). Clear associations to common compound-induced liver injuries (fatty liver, fibrosis, steatohepatitis, and cirrhosis), in addition to cancer-related processes, were observed. Drug-disease relationships are regarded as important ways to improve toxicity testing and drug safety and discovery; methods such as Connectivity map have been successfully applied to datasets, showing that correlation of 'omics' profiles between certain drugs and disease profiles recapitulate drug disease risks (Lamb et al., 2006; Caiment et al., 2014). Here, we show the potential of poorly annotated genes to strengthen these connections, impacting the discovery of potentially novel toxicity markers.

On another perspective, even though regulatory TF and pathway enrichment analyses have already been widely applied to many fields in biomedicine especially in cancer research (Darnell, 2002; Bhagwat and Vakoc, 2015), only a few case studies were shown in the field of drug safety and toxicity (Souza et al., 2017). Our unbiased enrichment analysis of regulatory TFs and pathways is one of the first studies to combine the analysis of both transcription factors and signaling pathways related to drug toxicity, especially focusing on poorly annotated entities regulated by these systems in an effort to propose additional markers of drug toxicity (Andersen et al., 2013; Jennings et al., 2013).

In our analyses we show that approximately 16% of the dark genes were mapped in TF-regulon database DoRothEA (**Table 6**). Among the enriched TFs of dark genes in the most-DILI group, we detected, for instance, E2F1, which has been demonstrated to be involved in liver fibrosis, a common end-point of compound-induced liver injury (Zhang Y. et al., 2014), as well as JUND in the inflammatory process in liver (Seki et al., 2012). Pathways' signatures, which are largely curated and expected to represent the activity states of signaling pathways, were also found to contain approximately 4% of dark genes modulated in this study. Enriched pathways for these entities included the Hypoxia pathway, known to play a role in inflammation and fibrosis (Nath and Szabo, 2012), PI3K pathway, that mediates liver injury in chronic fluorosis (Fan et al., 2015), as well as that of TNF-alpha pathway as the mediator of hepatotoxicity and regeneration

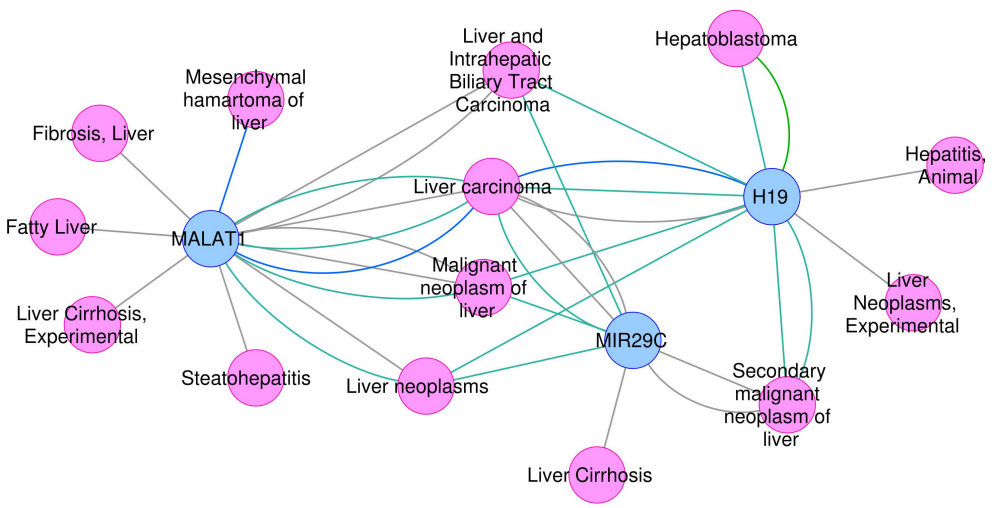


FIGURE 3 | Association between *H19*, *MALAT1*, and *MIR29C* and liver -related disease phenotypes.

(Schwabe and Brenner, 2006), inflammation and homeostasis (Tacke et al., 2009). These liver-injury mediating TFs E2F1 and JUND together with the representatives from hepatotoxic-related pathways such as HIF1A, AKT, and TNF-receptor could be perceived as potential markers to demonstrate the involvement of the dark genes in the context of DILI.

By comparing the dark genes identified in human hepatocytes to corresponding orthologs *in vivo* in a murine model, we found a consistency in the expression of these entities across species. More importantly, we show that these genes are associated with pathological outcomes (**Table 7**), highlighting their potential value in pre-clinical studies. We were not able to assess the relevance of aforementioned genes linked to DILI (*MALAT1*, *H19,* and *MIR29C*) since these genes, although possessing rat orthologs, were not measured in the arrays. However, functional annotation performed *in vitro* pointed similarities to most-DILI risk – demonstrated through genes such as *Magee1*. *Magee1* was modulated *in vitro* only by compounds in the most-DILI group, and associated to "Liver Cirrhosis, Experimental" according to DisGeNET data; *in vivo*, it was found in a module associated with hepatobiliary outcomes (Sutherland et al., 2017). Furthermore, genes such as *Smim14* and *Thyn1* were included in modules with biological processes; these functions may be putatively associated to these genes, as it has been shown that genes acting simultaneously often share the same biological process(es), and therefore gene co-expression networks can be used for the purpose of functional annotation (van Dam et al., 2017).

By combining the results of all approaches employed in this study, we were able to find evidence in at least one approach for 701 out of the initial 916 dark genes, i.e., 76% (**Supplementary Table S1**). Some genes were consistently found across all methodologies in addition to rat-human orthologs mapped to co-expression modules (e.g., *ST7, KLHDC2, CCDC28A, TMEM140, TRIM47*), all of which were included in clusters with GO BP annotation (**Supplementary Table S1**) (Sutherland et al., 2017). Genes exclusively modulated by the most-DILI group with (i) hits across several methods (i.e., sum of evidences equal or higher to 8, see **Supplementary Table S1**) (e.g., *ST7, LRBA, TPD52L2, TSSC4, BOLA1, YIPF1, TMEM168, RSRC2, CCDC92, ITFG1, ZMYND19, TTC14,* and *TMEM9*) and (ii) moderate amount of evidence (sum equal or higher than 4) and associated with pathologies (*MAGEE1*, *TBC1D31*, *SPATA2L*, *ABHD8,* and *LHFPL6*) were also identified. Although there are reports on their involvement in different liver diseases, including non-alcoholic steatohepatitis and hepatocellular carcinoma (Cai et al., 2018; Zhu et al., 2018), their roles in drug-induced organ injury has not yet been investigated. In addition to that, 215 dark genes modulated by the chemicals investigated here remain obscure – the majority (174) being classified as ncRNAs – which have been presented as potential non-invasive disease biomarkers (Teng and Ghoshal, 2015; de Gonzalo-Calvo et al., 2018). Regardless the level of findings, our results indicate concordance *in silico*, *in vitro,* and *in vivo* and potential roles in toxicity that should pave the way for further investigations aiming at the confirmation and uncovering of their biological function.

Overall, our study indicated how limitations arising from the biological annotation of genes can be minimized using a number of computational approaches, especially in the field of toxicogenomics in which uncovering and understanding of drug-gene responses is necessary to obtain novel/robust markers of toxicity. Although comprehensive databases such as Harmonizome (Rouillard et al., 2016) exist, they do not offer advanced mapping into the TF and pathway signatures nor cross-species concordance as performed in this study. It should also be noted that this study was based on a predefined set of approximately 19,000 genes; analyses of data from unconstrained methods (e.g., RNA-seq) using the methods described here will likely be able to provide a more accurate picture of the state of functional annotation of the whole human genome and shed light onto new, potentially relevant features in toxicological analysis.

CONCLUSION

In summary, this study highlighted a gap in functional gene annotation in the field of toxicogenomics and presented potential methods that can generate a pipeline to fill such gap through mapping using several resources. We showed that text mining tools and biocuration offer important insights by revealing potential chemical-disease associations and functional roles. The presented microRNA, ncRNAs and regulatory transcription factors in this study may also be further investigated as potential biomarkers of DILI. Nevertheless, further experimental validation of their biological roles are still necessary not only to extend the biological knowledge beyond the scope of well-annotated entities, but in order to also fully understand their roles in toxicity and disease development which would help to unlock their prognostic and translational value.

AUTHOR CONTRIBUTIONS

TS performed microarray analysis, cross-species comparison, and module enrichment. PT performed the annotations with transcription factor regulation and signaling pathways' signatures. JP characterized the genes and performed functional analysis. TS, PT, JP, LF, JS-R, JK, and DJ designed and revised the analyses. TS, PT, and JP wrote the manuscript. All authors read and revised the manuscript.

FUNDING

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement no. 116030 (TransQST). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00527/full#supplementary-material

TABLE S1 | Comprehensive overview of the dark genes analyzed, including (i) modulation by individual chemicals and per DILI risk, (ii) current mapping status to gene ontology (GO) and pathways (several databases), and (iii) results obtained from different methodologies applied.

DATA SHEET S1 | Supplementary Figures.

DATA SHEETS S2 and S3 | Scripts (.R) used in all analyses conducted in this study as well as mapping to databases and IDs.

REFERENCES

AbdulHameed, M. D. M., Tawa, G. J., Kumar, K., Ippolito, D. L., Lewis, J. A., Stallings, J. D., et al. (2014). Systems level analysis and identification of pathways and networks associated with liver fibrosis. *PLoS One* 9:e112193. doi: 10.1371/journal.pone.0112193

Aigner, A., Buesen, R., Gant, T., Gooderham, N., Greim, H., Hackermüller, J., et al. (2016). Advancing the use of noncoding RNA in regulatory toxicology: report of an ECETOC workshop. *Regul. Toxicol. Pharmacol.* 82, 127–139. doi: 10.1016/J.YRTPH.2016.09.018

Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res.* 45, D408–D414. doi: 10.1093/nar/gkw985

Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H., et al. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* 48, 838–847. doi: 10.1038/ng.3593

Andersen, M. E., McMullen, P. D., and Bhattacharya, S. (2013). Toxicogenomics for transcription factor-governed molecular pathways: moving on to roles beyond classification and prediction. *Arch. Toxicol.* 87, 7–11. doi: 10.1007/s00204-012-0980-6

Bhagwat, A. S., and Vakoc, C. R. (2015). Targeting transcription factors in cancer. *Trends Cancer* 1, 53–65. doi: 10.1016/j.trecan.2015.07.001

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., et al. (2015). Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* 43, D36–D42. doi: 10.1093/nar/gku1055

Cai, J., Zhang, X.-J., and Li, H. (2018). Progress and challenges in the prevention and control of nonalcoholic fatty liver disease. *Med. Res. Rev.* doi: 10.1002/med.21515 [Epub ahead of print].

Caiment, F., Tsamou, M., Jennen, D., and Kleinjans, J. (2014). Assessing compound carcinogenicity *in vitro* using connectivity mapping. *Carcinogenesis* 35, 201–207. doi: 10.1093/carcin/bgt278

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., et al. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* 39, D685–D690. doi: 10.1093/nar/gkq1039

Chen, M., Suzuki, A., Thakkar, S., Yu, K., Hu, C., and Tong, W. (2016). DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov. Today* 21, 648–653. doi: 10.1016/j.drudis.2016.02.015

Darnell, J. E. (2002). Transcription factors as targets for cancer therapy. *Nat. Rev. Cancer* 2, 740–749. doi: 10.1038/nrc906

de Gonzalo-Calvo, D., Vea, A., Bär, C., Fiedler, J., Couch, L. S., Brotons, C., et al. (2018). Circulating non-coding RNAs in biomarker-guided cardiovascular therapy: a novel tool for personalized medicine? *Eur. Heart J.* doi: 10.1093/eurheartj/ehy234 [Epub ahead of print].

Dempsey, J. L., and Cui, J. Y. (2017). Long non-coding RNAs: a novel paradigm for toxicology. *Toxicol. Sci.* 155, 3–21. doi: 10.1093/toxsci/kfw203

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46, D649–D655. doi: 10.1093/nar/gkx1132

Fan, B., Yu, Y., and Zhang, Y. (2015). PI3K-Akt1 expression and its significance in liver tissues with chronic fluorosis. *Int. J. Clin. Exp. Pathol.* 8, 1226–1236.

Fuschi, P., Carrara, M., Voellenkle, C., Garcia-Manteiga, J. M., Righini, P., Maimone, B., et al. (2017). Central role of the p53 pathway in the noncoding-RNA response to oxidative stress. *Aging* 9, 2559–2586. doi: 10.18632/aging.101341

Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Peat, G., et al. (2018). Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Res.* 78, 769–780. doi: 10.1158/0008-5472.CAN-17-1679

Garcia-Garcia, J., Guney, E., Aragues, R., Planas-Iglesias, J., and Oliva, B. (2010). Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics* 11:56. doi: 10.1186/1471-2105-11-56

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Haynes, W. A., Tomczak, A., and Khatri, P. (2018). Gene annotation bias impedes biomedical research. *Sci. Rep.* 8:1362. doi: 10.1038/s41598-018-19333-x

Hendrickx, D. M., Souza, T., Jennen, D. G. J., and Kleinjans, J. C. S. (2017). DTNI: a novel toxicogenomics data analysis tool for identifying the molecular mechanisms underlying the adverse effects of toxic compounds. *Arch. Toxicol.* 91, 2343–2352. doi: 10.1007/s00204-016-1922-5

Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., et al. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43, D921–D927. doi: 10.1093/nar/gku955

Jennings, P., Limonciel, A., Felice, L., and Leonard, M. O. (2013). An overview of transcriptional regulation in response to toxicological insult. *Arch. Toxicol.* 87, 49–72. doi: 10.1007/s00204-012-0919-y

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8:e1002375. doi: 10.1371/journal.pcbi.1002375

Kotlyar, M., Fortney, K., and Jurisica, I. (2012). Network-based characterization of drug-regulated genes, drug targets, and toxicity. *Methods* 57, 499–507. doi: 10.1016/J.YMETH.2012.06.003

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., et al. (2017). A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* 14, 61–64. doi: 10.1038/nmeth.4083

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Loosen, S. H., Schueller, F., Trautwein, C., Roy, S., and Roderburg, C. (2017). Role of circulating microRNAs in liver diseases. *World J. Hepatol.* 9, 586–594. doi: 10.4254/wjh.v9.i12.586

Luo, F., Liu, X., Ling, M., Lu, L., Shi, L., Lu, X., et al. (2016). The lncRNA MALAT1, acting through HIF-1α stabilization, enhances arsenite-induced glycolysis in human hepatic L-02 cells. *Biochim. Biophys. Acta Mol. Basis Dis.* 1862, 1685–1695. doi: 10.1016/J.BBADIS.2016.06.004

Matsuoka, S., Ballif, B. A., Smogorzewska, A., McDonald, E. R., Hurov, K. E., Luo, J., et al. (2007). ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* 316, 1160–1166. doi: 10.1126/science.1140321

Mitchell, J. A., Aronson, A. R., Mork, J. G., Folk, L. C., Humphrey, S. M., and Ward, J. M. (2003). Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu. Symp. Proc.* 2003, 460–464.

Nath, B., and Szabo, G. (2012). Hypoxia and hypoxia inducible factors: diverse roles in liver diseases. *Hepatology* 55, 622–633. doi: 10.1002/hep.25497

Oliva, J., Bardag-Gorce, F., French, B. A., Li, J., and French, S. W. (2009). The regulation of non-coding RNA expression in the liver of mice fed DDC. *Exp. Mol. Pathol.* 87, 12–19. doi: 10.1016/j.yexmp.2009.03.006

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, D358–D363. doi: 10.1093/nar/gkt1115

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Piñero, J., Gonzalez-Perez, A., Guney, E., Aguirre-Plans, J., Sanz, F., Oliva, B., et al. (2018). Network, transcriptomic and genomic features differentiate genes relevant for drug response. *Front. Genet.* 9:412. doi: 10.3389/fgene.2018.00412

Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G., et al. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016:baw100. doi: 10.1093/database/baw100

Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., et al. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* 9:20. doi: 10.1038/s41467-017-02391-6

Schueller, F., Roy, S., Vucur, M., Trautwein, C., Luedde, T., and Roderburg, C. (2018). The role of miRNAs in the pathophysiology of liver diseases and toxicity. *Int. J. Mol. Sci.* 19:E261. doi: 10.3390/ijms19010261

Schwabe, R. F., and Brenner, D. A. (2006). Mechanisms of liver injury. I. TNF-α-induced liver injury: role of IKK, JNK, and ROS pathways. *Am. J. Physiol. Liver Physiol.* 290, G583–G589. doi: 10.1152/ajpgi.00422.2005

Seki, E., Brenner, D. A., and Karin, M. (2012). A liver full of JNK: signaling in regulation of cell function and disease pathogenesis, and clinical approaches. *Gastroenterology* 143, 307–320. doi: 10.1053/j.gastro.2012.06.004

Souza, T., Jennen, D., van Delft, J., van Herwijnen, M., Kyrtoupolos, S., and Kleinjans, J. (2016). New insights into BaP-induced toxicity: role of major metabolites in transcriptomics and contribution to hepatocarcinogenesis. *Arch. Toxicol.* 90, 1449–1458. doi: 10.1007/s00204-015-1572-z

Souza, T. M., van den Beucken, T., Kleinjans, J. C. S., and Jennen, D. G. J. (2017). Inferring transcription factor activity from microarray data reveals novel targets for toxicological investigations. *Toxicology* 389, 101–107. doi: 10.1016/J.TOX.2017.07.008

Sutherland, J. J., Jolly, R. A., Goldstein, K. M., and Stevens, J. L. (2016). Assessing concordance of drug-induced transcriptional response in rodent liver and cultured hepatocytes. *PLoS Comput. Biol.* 12:e1004847. doi: 10.1371/journal.pcbi.1004847

Sutherland, J. J., Webster, Y. W., Willy, J. A., Searfoss, G. H., Goldstein, K. M., Irizarry, A. R., et al. (2017). Toxicogenomic module associations with pathogenesis: a network-based approach to understanding drug toxicity. *Pharmacogenomics J.* 18, 377–390. doi: 10.1038/tpj.2017.17

Tacke, F., Luedde, T., and Trautwein, C. (2009). Inflammatory pathways in liver homeostasis and liver injury. *Clin. Rev. Allergy Immunol.* 36, 4–12. doi: 10.1007/s12016-008-8091-0

Takahashi, K., Yan, I., Haga, H., and Patel, T. (2014). Long noncoding RNA in liver diseases. *Hepatology* 60, 744–753. doi: 10.1002/hep.27043

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J., et al. (2009). A novel signaling pathway impact analysis. *Bioinformatics* 25, 75–82. doi: 10.1093/bioinformatics/btn577

Tawa, G. J., AbdulHameed, M. D. M., Yu, X., Kumar, K., Ippolito, D. L., Lewis, J. A., et al. (2014). Characterization of chemically induced liver injuries using gene co-expression modules. *PLoS One* 9:e107230. doi: 10.1371/journal.pone.0107230

Teng, K.-Y., and Ghoshal, K. (2015). Role of noncoding RNAs as biomarker and therapeutic targets for liver fibrosis. *Gene Expr.* 16, 155–162. doi: 10.3727/105221615X14399878166078

The Gene Ontology Consortium (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi: 10.1093/nar/gkw1108

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13, 966–967. doi: 10.1038/nmeth.4077

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.* 19, 575–592. doi: 10.1093/bib/bbw139

Wang, X., Thijssen, B., and Yu, H. (2013). Target essentiality and centrality characterize drug side effects. *PLoS Comput. Biol.* 9:e1003119. doi: 10.1371/journal.pcbi.1003119

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* 41, W518–W522. doi: 10.1093/nar/gkt441

Xu, Y., Wu, J., Peng, X., Yang, T., Liu, M., Chen, L., et al. (2017). LncRNA LINC00341 mediates PM 2.5 -induced cell cycle arrest in human bronchial epithelial cells. *Toxicol. Lett.* 276, 1–10. doi: 10.1016/j.toxlet.2017.03.026

Yıldırım, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L., and Vidal, M. (2007). Drug—target network. *Nat. Biotechnol.* 25, 1119–1126. doi: 10.1038/nbt1338

Zhang, J. D., Berntenis, N., Roth, A., and Ebeling, M. (2014). Data mining reveals a network of early-response genes as a consensus signature of drug-induced *in vitro* and *in vivo* toxicity. *Pharmacogenomics J.* 14, 208–216. doi: 10.1038/tpj.2013.39

Zhang, Y., Xu, N., Xu, J., Kong, B., Copple, B., Guo, G. L., et al. (2014). E2F1 is a novel fibrogenic gene that regulates cholestatic liver fibrosis through the Egr-1/SHP/EID1 network. *Hepatology* 60, 919–930. doi: 10.1002/hep.27121

Zhao, J., Liu, Y., Wei, X., Yuan, C., Yuan, X., and Xiao, X. (2009). A novel WD-40 repeat protein WDR26 suppresses H_2O_2-induced cell death in neural cells. *Neurosci. Lett.* 460, 66–71. doi: 10.1016/j.neulet.2009.05.024

Zhou, Z., Liu, H., Wang, C., Lu, Q., Huang, Q., Zheng, C., et al. (2015). Long non-coding RNAs as novel expression signatures modulate DNA damage and repair in cadmium toxicology. *Sci. Rep.* 5:15293. doi: 10.1038/srep15293

Zhu, Q., Luo, Z., Lu, G., Gui, F., Wu, J., Li, F., et al. (2018). LncRNA FABP5P3/miR-589-5p/ZMYND19 axis contributes to hepatocellular carcinoma cell proliferation, migration and invasion. *Biochem. Biophys. Res. Commun.* 498, 551–558. doi: 10.1016/j.bbrc.2018.03.017

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 December 2018
doi: 10.3389/fgene.2018.00558

Persistence of Epigenomic Effects After Recovery From Repeated Treatment With Two Nephrocarcinogens

Alice Limonciel[1,2], Simone G. van Breda[3], Xiaoqi Jiang[4], Gregory D. Tredwell[5,6], Anja Wilmes[1,2], Lydia Aschauer[2,7], Alexandros P. Siskos[5], Agapios Sachinidis[8], Hector C. Keun[5], Annette Kopp-Schneider[4], Theo M. de Kok[3], Jos C. S. Kleinjans[3] and Paul Jennings[1,2]**

[1] Division of Molecular and Computational Toxicology, Amsterdam Institute for Molecules, Medicines and Systems, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [2] Division of Physiology, Department of Physiology and Medical Physics, Medical University of Innsbruck, Innsbruck, Austria, [3] Department of Toxicogenomics, GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, Netherlands, [4] Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany, [5] Division of Cancer, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Hammersmith Hospital, London, United Kingdom, [6] Department of Applied Mathematics, Research School of Physics and Engineering, Australian National University, Canberra, ACT, Australia, [7] Brookes Innovation Hub, Orbit Discovery, Oxford, United Kingdom, [8] Institute of Neurophysiology and Center for Molecular Medicine Cologne (CMMC), University of Cologne (UKK), Cologne, Germany

OPEN ACCESS

Edited by:
Pradyumna Kumar Mishra, ICMR-National Institute for Research in Environmental Health, India

Reviewed by:
Venkata Raghuram Gorantla, Tata Memorial Hospital, India
Ratnakar Tiwari, CSIR-Indian Institute of Toxicology Research, India

***Correspondence:**
Alice Limonciel
a.limonciel@vu.nl
Paul Jennings
p.jennings@vu.nl

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Genetics

Received: *28 May 2018*
Accepted: *31 October 2018*
Published: *03 December 2018*

Citation:
Limonciel A, van Breda SG, Jiang X, Tredwell GD, Wilmes A, Aschauer L, Siskos AP, Sachinidis A, Keun HC, Kopp-Schneider A, de Kok TM, Kleinjans JCS and Jennings P (2018) Persistence of Epigenomic Effects After Recovery From Repeated Treatment With Two Nephrocarcinogens. Front. Genet. 9:558. doi: 10.3389/fgene.2018.00558

The discovery of the epigenetic regulation of transcription has provided a new source of mechanistic understanding to long lasting effects of chemicals. However, this information is still seldom exploited in a toxicological context and studies of chemical effect after washout remain rare. Here we studied the effects of two nephrocarcinogens on the human proximal tubule cell line RPTEC/TERT1 using high-content mRNA microarrays coupled with miRNA, histone acetylation (HA) and DNA methylation (DM) arrays and metabolomics during a 5-day repeat-dose exposure and 3 days after washout. The mycotoxin ochratoxin A (OTA) was chosen as a model compound for its known impact on HA and DM. The foremost effect observed was the modulation of thousands of mRNAs and histones by OTA during and after exposure. In comparison, the oxidant potassium bromate ($KBrO_3$) had a milder impact on gene expression and epigenetics. However, there was no strong correlation between epigenetic modifications and mRNA changes with OTA while with $KBrO_3$ the gene expression data correlated better with HA for both up- and down-regulated genes. Even when focusing on the genes with persistent epigenetic modifications after washout, only half were coupled to matching changes in gene expression induced by OTA, suggesting that while OTA causes a major effect on the two epigenetic mechanisms studied, these alone cannot explain its impact on gene expression. Mechanistic analysis confirmed the known activation of Nrf2 and p53 by $KBrO_3$, while OTA inhibited most of the same genes, and genes involved in the unfolded protein response. A few miRNAs could be linked to these effects of OTA, albeit without clear contribution of epigenetics to the modulation of the pathways at large. Metabolomics revealed disturbances in amino acid balance, energy

catabolism, nucleotide metabolism and polyamine metabolism with both chemicals. In conclusion, the large impact of OTA on transcription was confirmed at the mRNA level but also with two high-content epigenomic methodologies. Transcriptomic data confirmed the previously reported activation (by $KBrO_3$) and inhibition (by OTA) of protective pathways. However, the integration of omic datasets suggested that HA and DM were not driving forces in the gene expression changes induced by either chemical.

Keywords: recovery, persistence, epigenomics, stress responses, ochratoxin A, potassium bromate, nephrotoxicity, metabolomics

INTRODUCTION

Ochratoxin A (OTA) is a food contaminating mycotoxin, a nephrotoxin and a suspected renal carcinogen (Limonciel and Jennings, 2013). In Europe, the average daily intake of OTA has been estimated at 1 $ng.kg^{-1}$ b.w. but exposures up to eight times higher have been reported (Schaaf et al., 2002; Clark and Snedeker, 2006). Its mechanism of toxicity remains elusive, with some studies suggesting genotoxicity, others suggesting epigenetic effects and yet others showing OTA-induced disturbances in the Nrf2 response to oxidative stress (Limonciel and Jennings, 2013; Vettorazzi et al., 2013). One striking effect of OTA is its very large impact on the transcriptome, affecting the expression of thousands of genes in both *in vitro* and *in vivo* settings (Jennings et al., 2012). Networks affected include genes involved in cytoskeleton organization, nucleosome regulation, transcription and translation, ubiquitination and cell cycle regulation. However, from the initiation of gene transcription to the splicing and maturation of mRNAs, a multitude of steps can alter gene expression and result in a disturbance of cellular homeostasis. Targeted mechanistic investigations have revealed that OTA perturbs the acetylation of proteins in general and of histones in particular. More specifically, OTA inhibited histone acetyltransferases (HATs) *in vitro* (Czakai et al., 2011) and enhanced the activity of histone deacetylases (HDACs) (Marin-Kuan et al., 2006), suggesting a global deacetylating effect. In rats, this toxin impacted the maturation of microRNAs (miRNAs) via a down-regulation of the expression of the genes encoding Dicer1 and Drosha (Dai et al., 2014). Thus, the large impact of OTA on gene expression could be due to a combination of factors including epigenetic modifications and differential miRNA regulation.

$KBrO_3$ is an oxidiser historically manufactured for primary use in bread preparations and hair products (International Agency for Research on Cancer, 2018). While bromate is not known to form in nature, it has been shown to occur during drinking water ozonation. Numerous cases of acute human exposures have been reported, usually following voluntary ingestions or after accidental contamination of bread preparations with excessive amounts of $KBrO_3$, causing nephrotoxicity and ototoxicity in children and adults (Campbell, 2006). The IARC classified KBrO3 as a possible carcinogen to humans as a consequence of the evidence found in rodents but in the absence of chronic exposure data in humans. In rodents, $KBrO_3$ exposure resulted in reactive oxygen species production and a depletion of glutathione, involved in the protection against oxidative stress (Sai et al., 1992; Zhang et al., 2010) as well as DNA damage (Ballmaier and Epe, 2006) involving the formation of 8-OHdG (Kasai et al., 1987; Cho et al., 1993), micronuclei (Hayashi et al., 1988) and chromosomal aberrations (Ishidate et al., 1984; Fujie et al., 1988).

In the current study, we investigated the effects of OTA and $KBrO_3$ on epigenetic modifications and miRNAs, and their potential link to the transcriptomic effects caused by the test chemicals. To this end, the global effects on mRNA and miRNA expression and epigenetic modifications (DNA methylation (DM) and histone acetylation (HA)) were integrated and compared in a human renal proximal tubule cell line (RPTEC/TERT1) exposed to the chemicals in a repeat-dose testing regime and after a recovery period of 3 days after treatment. In addition, we investigated the metabolomic profile of these cells during and after exposure to identify downstream dysfunctions in homeostasis regulation. The modulation of stress response pathways was also addressed within the mechanistic investigation, with a particular focus to the Nrf2 response to oxidative/alkylating stress and the activation of p53, another transcription factor widely known as a tumor suppressor for its role in the maintenance of DNA integrity in the presence of carcinogens, for which $KBrO_3$ served as a positive control (Limonciel et al., 2012).

MATERIALS AND METHODS

Chemicals

The two chemicals in study were purchased from Sigma-Aldrich (OTA, O1877 and $KBrO_3$ P7332). All chemicals unless otherwise stated were purchased from Sigma and were of the highest grade available.

Cell Culture

Under routine conditions, human proximal tubule RPTEC/TERT1 cells (Wieser et al., 2008) were cultured at 37°C in a 5% CO_2 humidified atmosphere, fed 3 times a week and sub-cultured by trypsinisation. RPTEC/TERT1 cells were seeded onto 96-well cell culture plates (655180, Greiner) for concentration screening, PET 96-well E-plate VIEW cell culture plates (300600910, ACEA) for impedance measurements, 6-well cell culture plates (657160, Greiner) for mRNA and miRNA sample preparation, and on 10-cm cell culture dishes (831802,

Sarstedt) for all other measurements. Cells were grown in hormonally defined medium (HDM) as previously described (Aschauer et al., 2013). Briefly, after confluence was reached, the cells were allowed to stabilize and form a contact-inhibited monolayer for ten days with feeding every 2–3 days. HDM consisted of a 1:1 mixture of Dulbecco's modified Eagle's medium (DMEM, Invitrogen, cat. no. 11966) and Ham's F-12 nutrient mix (Invitrogen, cat. no. 21765) supplemented with 2 mM glutamax (Invitrogen, cat. no. 35050-038), 5 μg/mL insulin, 5 μg/mL transferrin and 5 ng/mL sodium selenite, 100 U/mL penicillin and 100 μg/mL streptomycin, 10 ng/mL epithelial growth factor and 36 ng/mL hydrocortisone.

Stocks of the test chemicals were prepared as follows. Five milligrams OTA were dissolved in DMEM/F-12 medium (without additives) to a 2.48 mM stock and further diluted to a 50× stock (6.5 μM) in DMEM/F-12. $KBrO_3$ was directly dissolved to a 50X stock (40 mM) in the same aqueous solvent.

Cell Viability and Cell Stress

Test concentrations were 130 nM OTA and 0.8 mM $KBrO_3$ for all omic experiments. These concentrations were sub-cytotoxic, but induced cellular stress, as shown in preparatory experiments (**Figure 1A**). Impedance was measured in the xCELLigence device from ACEA. Cells were seeded onto E-plates in 60 μL medium and differentiated. Impedance reflects the attachment of the cells to the growth support and can therefore be used as a cell viability endpoint in contact-inhibited cell monolayers. Cell index (CI) was measured every 24h and normalized for each treatment condition to the average T1 value ($n = 3$). Decreased CI corresponds to a decrease in cell viability (Limonciel et al., 2018). Increased CI can be seen as a marker of cellular stress, possibly linked to the collapse of dome structures on solid plastic support. Supernatant lactate was quantified using a biochemical assay (Limonciel et al., 2011). Statistical analysis was performed using a two-way ANOVA with a Bonferroni multiple comparisons posttest using GraphPad Prism v6.01 for each dataset ($^*p < 0.05$).

Cell Treatment

Treatments were applied in a bi-phasic regime where the cells were exposed to either OTA, $KBrO_3$ or HDM (vehicle control) for 5 days and allowed a 3-day recovery period post-treatment where all cells were exposed to HDM. For both the treatment and recovery phases, cell culture medium was renewed every 24 h. Cell lysates were prepared for omic investigations and OTA quantification after 1, 3 and 5 days of treatment (T1, T3 and T5, respectively) and at the end of the recovery period (R3, day 8 of experiment) (**Figure 1B**). Supernatant medium was collected for OTA quantification at the same time points. All omic endpoints were measured in three biological replicates.

RNA Preparation

At each lysis time point, medium was removed and cells were harvested in Qiazol (Qiagen). Total RNA was isolated using a miRNeasy Mini Kit (Qiagen, 217004) according to the manufacturer's protocol and followed by DNase I (Qiagen) treatment. Upon purification, RNA concentrations were measured with a NanoDrop® ND-1000 spectrophotometer (Thermo Scientific) at 260 and 280 nm. RNA quality and integrity were assessed by automated gel electrophoresis on an Agilent 2100 Bioanalyzer system (Agilent Technologies). Only RNA samples which showed clear 18S and 28S peaks and with an RNA integrity number (RIN) higher than 8 were used. Samples were stored at −80°C until RNA hybridization.

mRNA Microarrays

The DNA array platform used was the Affymetrix Human Genome U133 plus 2.0 array. CEL files were loaded into BRB array[1] and normalized using the RMA method. Differences in gene expression measurements under any condition compared to time-matched control were summarized for each probe as log2 fold change (LFC) value. The associated p-value was computed using the moderated (unpaired) t-test (Smyth, 2004) and was corrected for multiple testing by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Calculations were performed using the R package "limma" (Smyth, 2005). All transcriptomic data was deposited at Array Express under the accession number E-MTAB-7048.

miRNA Microarrays

Profiling of miRNA expression was performed using Agilent Sureprint G3 Unrestricted Human miRNA V19 8 × 60 K microarrays. Hybridization was performed following standard protocols, after which the microarray slides were washed and scanned using a DNA microarray scanner (Agilent Technologies). The scanned images were converted into TXT files using the Feature Extraction Software v10.7.3.1 from Agilent Technologies, which were imported in R 2.15.3[2] for quality control with an in-house developed pipeline (Coonen et al., 2015). Filtering and normalization was performed using AgiMicroRna (López-Romero, 2011). Total gene signals were log2-transformed and quantile-normalized. Differentially expressed microRNAs with an FDR adjusted p-value < 0.05 were considered statistically significant.

DNA Methylation

Cells were washed with HBSS, trypsinised and lysed with a digestion buffer containing 1 mM EDTA, 50 mM Tris–HCl, 5% SDS and 1 mg/mL proteinase K. DNA was extracted and processed for methylated DNA immunoprecipitation (MeDIP) before hybridisation onto Human 2.1 M Deluxe Promoter arrays (Roche NimbleGen, Basel, Switzerland). Detailed procedures are available in the **Supplementary Material**.

Signal intensity was extracted from images using NimbleScan v2.6 and differentially methylated regions (DMRs) compared to control were identified following analysis with a probe sliding window ANOVA algorithm (sliding window of 750 bp comprising 7 probes and a FDR corrected p-value < 0.01). Detail of the analysis can be found in the **Supplementary Material** and (van Breda et al., 2014). Log2 ratios > 0 indicate hypermethylation and log2 ratios < 0 indicate hypomethylation.

[1] http://linus.nci.nih.gov/BRB-ArrayTools.html

[2] http://www.r-project.org

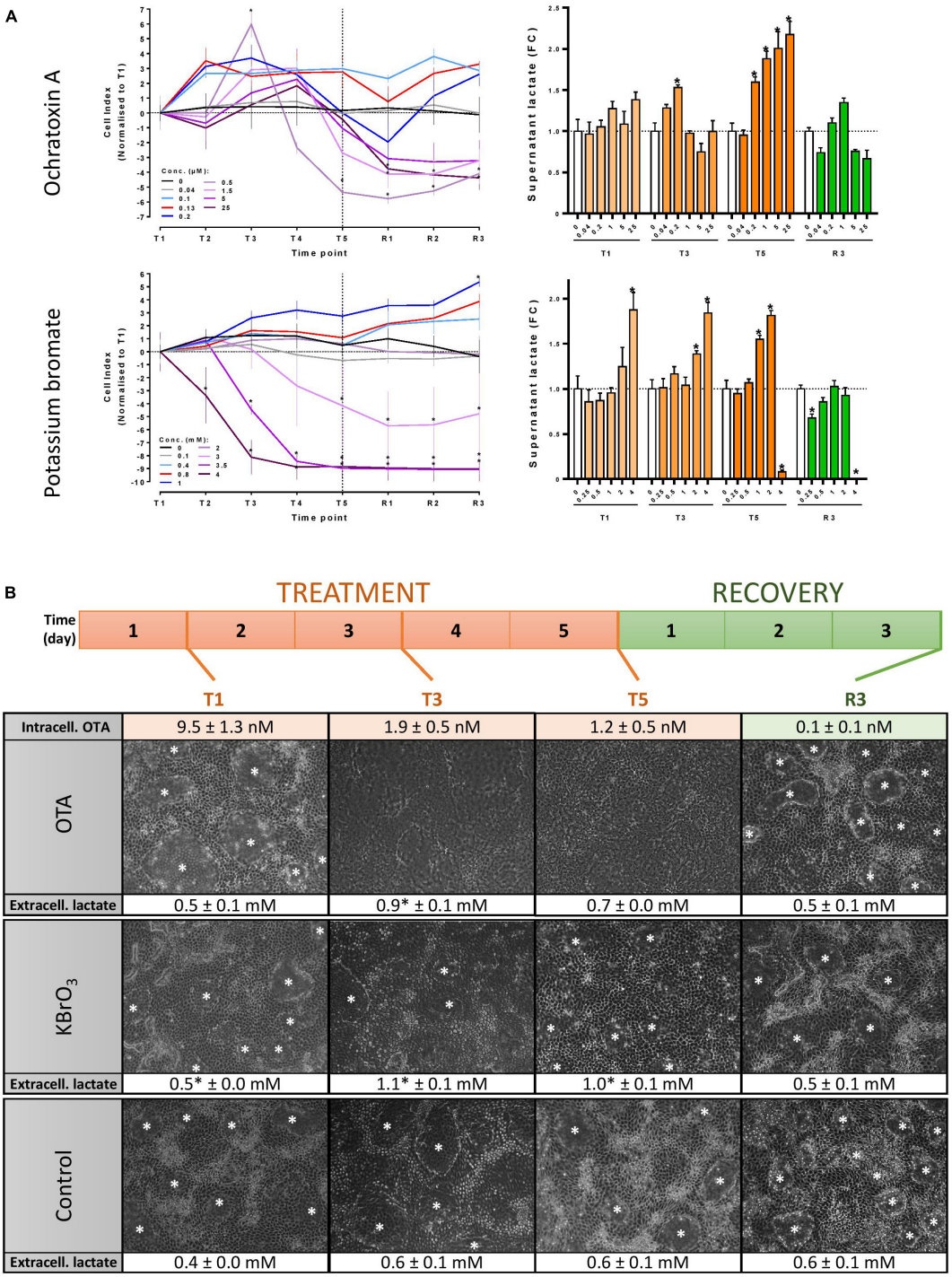


FIGURE 1 | Experimental design optimisation. **(A)** Concentration range testing in 96-well plates: Mature RPTEC/TERT1 cells were exposed to up to 25 μM OTA, 4 mM $KBrO_3$ or a vehicle control for 5 days (T5) and fed a further 3 days with medium (R3). Cell viability was assessed by measuring impedance and calculating cell index (CI) with the xCELLigence device. Cellular stress was assessed by measuring supernatant lactate. Significance of change compared to time-matched control in a minimum of 3 biological replicates was assessed with a two-way ANOVA followed by a Bonferroni posttest. *$p \leq 0.05$. T1, T2, T3, T4, T5: days 1, 2, 3, 4, and 5 of treatment, respectively. R1, R2, R3: days 1, 2, and 3 of recovery post-treatment. **(B)** Experimental design for the omic study. Mature RPTEC/TERT1 cells monolayers were treated every 24 h for 5 days with 130 nM OTA, 0.8 mM $KBrO_3$ or a vehicle control. The treatment was followed by a 3-day recovery period with 24h-feeding cycles with cell culture medium for all three conditions. Cells were lysed for omic investigations at the 4 indicated time points. Intracellular OTA levels were measured by UPLC-MS at the 4 time points. Cell viability and cell stress were monitored throughout the experiment by microscopic examination (original magnification 100×) and lactate measurements in cell culture supernatants. Lactate levels in treated cells were compared to time-matched controls with a two-tailed unpaired Student's *t*-test (*$p \leq 0.05$, $n = 3$). White asterisks (*) show domes indicative of vectorial transport.

Histone Acetylation Analyses Using Chip-on-Chip

Chromatin immunoprecipitation was performed using the SimpleChIP® Enzymatic Chromatin IP Kit (Magnetic Beads) (Cell Signaling Technology) as detailed in the **Supplementary Material**. H3K9 acetylation of human promoters was studied using the Human 2.1 M Deluxe Promoter Array (Roche NimbleGen, Basel, Switzerland). Labelling, hybridization and washing of arrays was performed according to the manufacturer's protocol as described in the DNA methylation section. Data analysis and selection of differentially acetylated genes was performed with the same workflow as for the MeDiP-chip data.

Stress Response Pathway Analysis

Pathway genes were chosen based on the lists of target genes previously generated in our group (Limonciel et al., 2015). Genes were ranked based on their average log2 fold over control (LFC) across the three time points of treatment with $KBrO_3$, the positive control for Nrf2 and p53 activation (Limonciel et al., 2012). Genes showing no significant modulation by $KBrO_3$ were removed from the original list. The remaining genes are ranked based on the average LFC during treatment with OTA.

GC-MS Metabolomics

For metabolomics, the cells were lysed in ice-cold methanol (MeOH). Cell lysate samples were derivatised for GC-MS by a two-step methoximation/silylation derivatization procedure (Kind et al., 2009). The following derivatization standards were added to the samples: ^{13}C-Serine (20 μL, 1 mM), U-^{13}C-Glucose (20 μL, 1 mM) and myristic acid d27 (10 μL, 1.5 mg/mL). The dried samples were first methoximated with a solution of 20 mg/mL methoxyamine hydrochloride in anhydrous pyridine (20 μL) and incubated at 30°C for 90 min. Samples were then silylated by adding 80 μL MSTFA (with 1% TMCS) (Thermo) and incubating at 37°C for 30 min. Following derivatization, 2-fluorobiphenyl in anhydrous pyridine (10 μL, 1 mM) was added as an injection standard and the samples were transferred to deactivated glass vial inserts. GC-MS analysis was performed on an Agilent 7890 GC equipped with a 30 m DB-5MS capillary column with a 10 m Duraguard column connected to an Agilent 5975 MSD operating under electron impact (EI) ionization (Agilent Technologies UK Ltd.). Samples were injected with an Agilent 7693 autosampler injector into deactivated splitless liners according to the method of Fiehn and colleagues (Kind et al., 2009) using helium as the carrier gas. One sample was used as a quality control (QC) and injected repeatedly throughout the run to monitor system performance. Metabolites were assigned using the Fiehn Library with the deconvolution program AMDIS (Stein, 1999), and Matlab program GAVIN, developed in-house, was used to integrate metabolite peak areas for all samples (Behrends et al., 2011). Data was normalized by the QC-RLSC method described by Dunn et al. (2011). Statistical significance of the change induced by OTA or $KBrO_3$ was assessed using a two-way ANOVA with a Sidak posttest in GraphPad Prism v6.05. Significant changes ($p \leq 0.05$) are indicated in bold in **Figure 7**.

OTA Quantitation by UPLC-MS

For intracellular extracts, 150 μL aliquots of the MeOH extracts were dried under reduced pressure in a speedvac, and resuspended in 1:9 acetonitrile:water (100 μL) using UPLC grade solvents (Romil LTD, Code H949, Cambridge, United Kingdom). Sample solutions were then transferred to high recovery chromatography vials (Waters Corporation, Milford, MA, United States). For cell culture medium samples, 100 μL aliquots of the media were added to 300 μL MeOH. Samples were vortexed and then centrifuged at 16000 *g* for 5 min. Supernatants were transferred to high recovery chromatography vials and concentrated under reduced pressure in a speedvac, before resuspension in 1:9 acetonitrile:water (100 μL). Standard solutions of OTA ranging from 100 to 0.001 ng/mL (248 to 0.002 nM) were prepared in 1:9 acetonitrile:water (100 μL) and transferred to high recovery chromatography vials. Reversed-phase chromatographic separation of the cell lysates was conducted using an Acquity UPLC system (Waters Corporation, Milford, MA, United States) on an Acquity HSS T3 C18 column 10 mm × 2.1 mm, 1.8 um (Waters) and a binary gradient elution comprising water +0.1% formic acid (Sigma) and acetonitrile +0.1% formic acid, with an injection volume of 15 μL. Mass spectrometric analysis of the chromatographic eluent was performed using a quadrupole time-of-flight (QtoF-Ultima) spectrometer (Waters) with data collected in centroid mode in the m/z range 70–1000. Analysis was performed in positive ion mode electrospray ionization. The elution gradient was as follows: 99.5% A at 0 min-99.5% at 3 min to 99.5% B at 19 min- 99.5% B at 23 min, 99.5% A at 23.1 min, 99.5% A at 27 min. The column was kept at 50°C and the auto-sampler at 4°C. Limit of detection (LOD) was 0.12 nM and limit of quantitation (LOQ) was 0.32 nM.

Statistical Analysis

For all omics and OTA quantification, three biological replicates were produced. Statistical analysis is reported in the respective section for each method. For correlation of epigenetic modifications with gene expression levels (**Figure 3A**), the correlation of differentially expressed genes with epigenetic modifications was tested with Spearman's rank correlation coefficient (ρ) at each time point. Strong correlation of HA or DM with the direction of gene expression changes renders a coefficient close to 1, strong anti-correlation is represented by a coefficient close to -1. A *p*-value for statistical significance of the association was also calculated and is reported by asterisks when significant ($^{*}p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$).

RESULTS

Concentration Range Testing and OTA Uptake

The concentrations of OTA and $KBrO_3$ used for the omic investigations were chosen after rigorous concentration range testing in RPTEC/TERT1 cells to cause a minimal decrease in cell viability (impedance/CI) and an increase in cellular

stress (extracellular lactate) (**Figure 1A**). Based on these results, three biological replicates of differentiated RPTEC/TERT1 cells were exposed to 130 nM OTA, 0.8 mM $KBrO_3$ or a vehicle control in a 5-day repeat-dose exposure regime followed by a 3-day recovery period with compound washout and cell culture medium renewal every 24 h (**Figure 1B**). There was no significant cell death throughout the omic experiment with either chemical. However, both caused an increase in the stress marker lactate during treatment. OTA also caused the disappearance of dome structures indicative of vectorial transport of water and solutes in proximal tubule cells cultured on solid support (Wilmes et al., 2014). The cells were lysed for transcriptomic, epigenomic (histone acetylation and DNA methylation arrays), miRNA and metabolomic investigations at 4 time points: after 1, 3 and 5 days of treatment (T1, T3 and T5, respectively) and after 3 days of recovery post-treatment (R3).

Ochratoxin A itself was quantified in cell lysates at the 4 time points by UPLC-MS, revealing that after 24 h of exposure, 9.5 ± 1.3 nM OTA were present in the cells (**Figure 1B**). Intracellular levels were lower in the following days, demonstrating a lack of accumulation of the parent compound in spite of 5 consecutive exposures between T1 and T5. In the supernatant OTA remained close to 90 nM at all treatment time points and was below 0.3 nM (LOQ) at R3 (data not shown). At R3, cellular OTA was negligible (less than 0.12 nM), demonstrating an effective washout of the chemical. Interestingly, a product of hydrolysis of OTA, OTAα, was also detected with a peak at T3. Low intracellular levels of OTAα were still detectable after recovery. Taken together, these data suggest that the nominal concentration is close to the actual treatment concentration, that OTA enters the cell, where it is hydrolysed to OTAα and that a 3-day washout effectively reduces the internal concentration of OTA and its metabolite.

Quantitative Impact on Gene Expression and Regulatory Mechanisms

The epigenomic and transcriptomic datasets revealed a very large impact of OTA on histone acetylation (HA) and mRNA expression and a milder effect on DNA methylation (DM) and miRNA expression (**Figure 2**). At T5, OTA had induced significant changes compared to time-matched controls on 11047 genes for HA, 5793 genes for mRNA, 639 genes for DM and 10 miRNAs. Interestingly, the largest impact on mRNA expression occurred after the first 24h of exposure (9102 genes), likely through a direct impact on transcription or mRNA processing, while the impact on HA was strongest on the last day of treatment (T5) and after recovery (R3). In comparison, $KBrO_3$ had a much smaller quantitative impact on epigenetics and mRNA expression, but modulated more miRNAs than OTA, especially at T5 (**Figure 2**).

A global analysis of the correlation of HA/DM status with modulated gene expression (GE) at each time point for both compounds was conducted (**Figure 3A**). The results show that the strongest correlation was between HA and GE in $KBrO_3$-treated cells with a maximum at T1 (correlation coefficient $\rho = 0.75$). Although OTA had a dramatically stronger effect on both HA and GE in terms of number of impacted genes, the correlation between HA and GE in OTA-treated cells was much weaker, suggesting a weak contribution of HA to GE modulation in spite of a very high number of differentially acetylated histones. For DM, where anti-correlation with GE would be expected, neither compound showed a strong anti-correlation.

Effects of OTA on HA regulation have already been reported, notably through inhibition of HATs and activation of HDACs, which would favor a global decrease in histone and overall protein acetylation (Marin-Kuan et al., 2006). While in this study HA was strongly impacted in presence of OTA and after compound washout, **Figure 3B** shows that the majority of epigenetic modifications observed at R3 were new and not conserved from treatment. The most conserved modifications were hypoacetylations with 28% of R3 hypoacetylations already present at T5. However, when cross-analyzing persistent hypoacetylations with significantly modulated mRNAs in the OTA dataset (set of 85 genes-**Figure 4**), 44 genes showed a directionality of gene expression consistent with histone modifications, vs. 41 genes with histone modifications that would favor the opposite effect on gene expression. In the $KBrO_3$ dataset, 43 genes had persistently differentially acetylated histones from T5 to R3. Only 3 of those genes were also differentially expressed at R3 compared to control: CCL2, TGFB2 and SRRM2. All three genes were down-regulated at R3, CCL2 and TGFB2 were hypo-acetylated and SRRM2 was hyper-acetylated. Thus, even with a focus on persisting histone modifications, there was no global correlation between HA and GE with either chemical, suggesting a marginal effect of HA modulation *alone* on mRNA expression on a global scale.

In the miRNA dataset, while $KBrO_3$ induced the most deregulations at T5, very few miRNAs overlapped with any other condition, suggesting a peak of miRNA production that is absent in the OTA dataset. In contrast, while only a maximum of 10 miRNAs were affected at any time point with OTA, most of the up-regulated miRNAs were impacted at several time points, with four of them still up-regulated at R3: miR-3065-3p, miR-141, miR-542-3p and miR-542-5p. In contrast, miR-450a and miR-219-5p were up-regulated and miR-1226* and miR-370 were down-regulated in recovery exclusively (**Figure 5A**). These last two miRNAs were also down-regulated in the $KBrO_3$ dataset, after recovery only (**Figures 5B,C**). miR-132 was heavily induced by OTA treatment only. **Figure 5C** compares the changes in miRNAs in both treatments based on log2 fold change (LFC). Two miRNAs (miR-23b, miR-29b-1) were similarly down-regulated by both treatments after repeated exposures (T3 and T5). Five miRNAs were down-regulated by OTA but up-regulated by $KBrO_3$ treatment at T1 (miR-21-3p, miR-1181, miR-134, miR-3663-3p and miR-4271). Interestingly, the two chemicals had opposite effects on miR-542-5p expression, both at T5 and R3, where its levels were increased by OTA and decreased by $KBrO_3$.

Impact on Stress Response Pathways

The impact of both chemicals on the expression on Nrf2, p53 and unfolded protein response (UPR) related genes, as well as the epigenetic modifications identified on the transcription factors

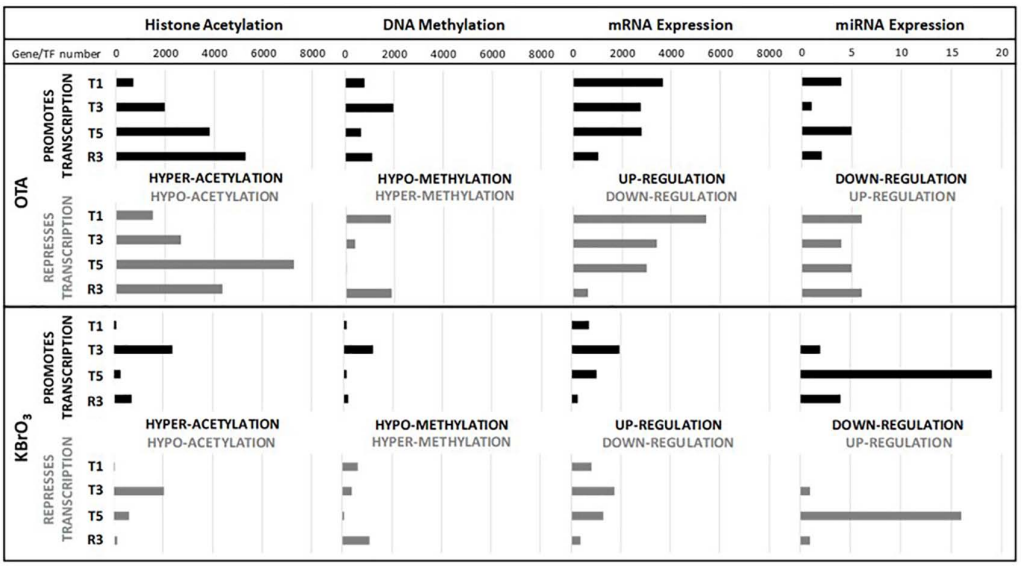


FIGURE 2 | Quantitative impact on gene expression and epigenetic modifications. The number of hypo- and hyper- acetylated histones/methylated genes was based on the sign of the difference of the median treatment and the median control values. For mRNA, the probe list was reduced per time point and treatment with a cut off on *p*-value of change of 0.001. When several probes for a given gene remained, the probe with the highest variation in the condition was chosen. No cut off on the intensity of change was applied. For miRNA, the probe list was reduced per time point and treatment with a *p*-value cut off of 0.05. For each endpoint, the number of changes is represented on the *X* axis and the different time points (T1, T3, T5, R3) on the *Y* axis.

and their target genes are shown in **Figure 6**. $KBrO_3$ induced the up-regulation of Nrf2 and p53 target genes, yet with very few corresponding HA and DM modifications. In contrast, $KBrO_3$ did not alter the expression of the UPR targets studied here.

In OTA-treated cells, several Nrf2 and p53 targets were down-regulated during treatment. In line with the large impact of OTA on HA, most of the genes studied in **Figure 6** showed HA modifications at at least one time point, however there was no sign of a consistent epigenetic modulation with a lasting effect on gene expression. Only two miRNAs impacted by OTA had validated targets from this list: miR-1285-3p targeting TP53 and miR-132-3p targeting CDKN1A (p21). The transcripts of two of the UPR-driving transcription factors (ATF4 and XBP1) were down-regulated in OTA-treated cells. Several of their target genes were also down-regulated, with the notable exception of HSPA5 (up-regulated from T1 to T5), which encodes the protein BiP responsible for protein misfolding sensing in the ER and the activation of all three branches of the UPR. Two elongation factors (EIF2S2 and EIF1), involved in translation, were also consistently up-regulated during OTA treatment and recovered after washout.

Altogether, these results suggest an inhibition of the Nrf2, p53 and unfolded protein responses by OTA at gene expression level, which does not appear to be driven primarily by epigenetic mechanisms (HA, DM or miRNA).

Metabolic Impact

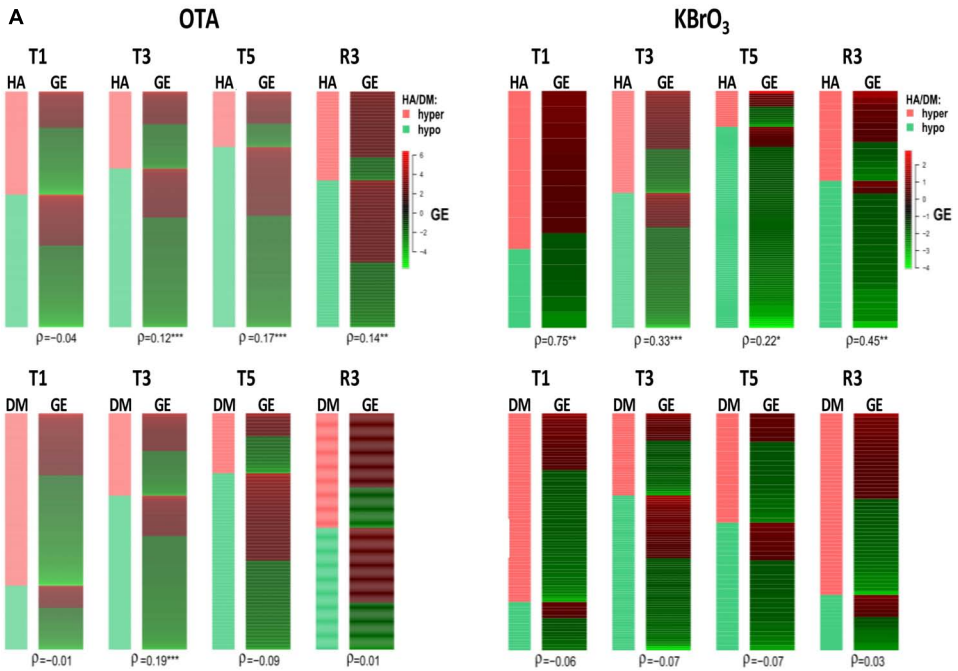
Amino Acids

Metabolomic analysis of the intracellular contents of OTA-treated cells revealed an increase in both essential (histidine, isoleucine, leucine, valine, threonine, methionine, tryptophan phenylalanine and lysine) and non-essential (alanine, proline, glutamine, glutamic acid, serine, tyrosine) amino acids after the first 24 h of treatment, with the notable exception of the non-essential amino acids aspartic acid, cysteine and glycine that are all important in the synthesis of the antioxidant glutathione (**Figure 7**). This early response was not sustained at later treatment time points, where most amino acid deregulations were toward a depletion, with the most striking effects on aspartic acid, cysteine and glycine at T3. After OTA washout and recovery (R3), the depletion of several essential (isoleucine, leucine, valine, phenylalanine) and non-essential (asparagine, alanine, proline, glutamine, glutamic acid, serine, tyrosine) amino acids could still be measured, while aspartic acid levels were increased compared to time-matched controls.

Upon exposure to $KBrO_3$, the metabolomic profile was very different, with a mild impact only on non-essential amino acids at T1, but which was sustained until T5 (increased levels of alanine, glutamic acid, aspartic acid). In addition, serine and tyrosine levels were decreased at T3, but returned to control level at T5, while glutamine was depleted at both time points. Essential amino acids were depleted at the late treatment time points only: histidine, methionine, tryptophan, phenylalanine and lysine at T3; methionine, phenylalanine and lysine at T3 and T5. After $KBrO_3$ washout and recovery, amino acid levels were still not back to control levels. In particular, isoleucine, leucine, methionine and aspartic acid levels were still elevated compared to time-matched control (**Figure 7**).

Metabolic Pathways

Ochratoxin A also largely affected metabolites related to cellular energy production and nucleotide biosynthesis and degradation



B

		Hyper-acetylated	Hypo-acetylated	Hyper-methylated	Hypo-methylated
OTA	*# genes at R3*	694	1232	2	45
	% of R3 total	13.2	28.4	1.1	4.1
$KBrO_3$	*# genes at R3*	35	8	7	0
	% of R3 total	4.8	6.0	0.6	0.0

FIGURE 3 | Correlation of epigenetic modifications to mRNA expression changes and persistence after washout and recovery in OTA- and $KBrO_3$-treated cells. **(A)** - Correlation between epigenetics and gene expression was tested at each time point with Spearman's rank correlation coefficient (ρ). Strong correlation between histone acetylation (HA) or DNA methylation (DM) and mRNA gene expression (GE) renders a coefficient close to 1, strong anti-correlation renders a coefficient close to -1. A *p*-value for statistical significance of the correlation was also calculated and is indicated by $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. **(B)** Persisting epigenetic modifications after compound washout and recovery period. Here, epigenetic modifications with sustained directionality at T5 and R3 for a given gene were considered persisting after treatment. Percentage of persistent modifications was calculated as the percentage of modifications in the R3 datasets that were already present at T5 and modified in the same direction compared to time-matched control; e.g., for OTA, 694 genes corresponding to 13.2% of hyper-acetylated genes at R3 were already hyper-acetylated at T5.

(**Figure 7**). Several intermediates of glycolysis and the TCA cycle were significantly impacted by OTA at T1, with glucose, glucose-6-phosphate, 3-phosphoglycerate, lactic acid, citric acid, alpha ketoglutaric acid, fumaric acid and malic acid levels all increased after 24 h. At later exposure time points, succinic acid (decreased) was the only modulated metabolite of these two pathways. After OTA washout and recovery, the first (glucose-6-phosphate) and last (lactic acid) metabolites of glucose degradation through glycolysis were decreased, as well as citric acid. Interestingly, fructose (among the polyols) was consistently increased throughout OTA exposure (strongest increase in the metabolomic dataset) but not at R3. $KBrO_3$ did not affect fructose levels, but consistently caused a depletion in glucose in the cells, which was recovered at R3. In addition, $KBrO_3$ induced an increase at T1 in all TCA intermediates measured except citric acid. These higher levels were sustained until T3 for alpha ketoglutarate and malic acid and until T5 for succinic acid. At R3, fumaric and malic acid levels were still above control levels after $KBrO_3$ washout.

Metabolomics also revealed an impact on the pentose phosphate pathway (PPP), which can branch from glycolysis (C6 metabolism) to provide C5 ribose derivatives, notably for nucleotide biosynthesis. OTA caused an increase in 6-phosphogluconic acid (T1), D-ribose (T1, T3, and T5) and ribose-5-phosphate (T3). $KBrO_3$ depleted the levels of glucuronic acid from T1 to T5 and of ribose-5-phosphate at T1 and T3 only, without a significant impact on ribose itself. Both chemicals caused an increase at T1 and a decrease at T3 of the intracellular levels of orotic acid, a downstream metabolite involved in the early steps of pyrimidine biosynthesis. An increase was also measured after recovery from $KBrO_3$ but not from OTA. Cytosine and CMP levels were not impacted by either chemical, while uracil and UMP levels were affected at the early time points by OTA only. The degradation product 3-aminoisobutyric acid was consistently increased by $KBrO_3$ during exposure and after washout, while OTA's effects followed the pattern of other pyrimidine metabolites with an increase at T1 and a decrease at later treatment time points.

Entrez ID	Gene symbol	T1 T3 T5 R3	HA	Corr.	Gene name
6781	STC1		HA-	NO	stanniocalcin 1
81578	COL21A1		HA-	NO	collagen, type XXI, alpha 1
56934	CA10		HA+	YES	carbonic anhydrase X
54443	ANLN		HA-	NO	anillin, actin binding protein
3758	KCNJ1		HA+	YES	potassium inwardly-rectifying channel, subfamily J, member 1
51108	METTL9		HA-	NO	methyltransferase like 9
2202	EFEMP1		HA-	NO	EGF containing fibulin-like extracellular matrix protein 1
5734	PTGER4		HA+	YES	prostaglandin E receptor 4 (subtype EP4)
1906	EDN1		HA+	YES	endothelin 1
55824	PAG1		HA-	NO	phosphoprotein associated with glycosphingolipid microdomains 1
257019	FRMD3		HA-	NO	FERM domain containing 3
388650	FAM69A		HA-	NO	family with sequence similarity 69, member A
55086	CXorf57		HA-	NO	chromosome X open reading frame 57
26034	IPCEF1		HA-	NO	interaction protein for cytohesin exchange factors 1
2115	ETV1		HA-	NO	ets variant 1
64756	ATPAF1		HA-	NO	ATP synthase mitochondrial F1 complex assembly factor 1
23179	RGL1		HA-	NO	ral guanine nucleotide dissociation stimulator-like 1
83872	HMCN1		HA-	NO	hemicentin 1
5570	PKIB		HA-	NO	protein kinase (cAMP-dependent, catalytic) inhibitor beta
64343	AZI2		HA-	NO	5-azacytidine induced 2
7477	WNT7B		HA+	YES	wingless-type MMTV integration site family, member 7B
56898	BDH2		HA-	NO	3-hydroxybutyrate dehydrogenase, type 2
26249	KLHL3		HA-	NO	kelch-like 3 (Drosophila)
51528	JKAMP		HA-	NO	JNK1/MAPK8-associated membrane protein
54414	SIAE		HA-	NO	sialic acid acetylesterase
23111	SPG20		HA-	NO	spastic paraplegia 20 (Troyer syndrome)
51302	CYP39A1		HA-	NO	cytochrome P450, family 39, subfamily A, polypeptide 1
2104	ESRRG		HA-	NO	estrogen-related receptor gamma
7707	ZNF148		HA-	NO	zinc finger protein 148
2517	FUCA1		HA-	NO	fucosidase, alpha-L- 1, tissue
29114	TAGLN3		HA-	NO	transgelin 3
8863	PER3		HA-	NO	period homolog 3 (Drosophila)
1738	DLD		HA-	NO	dihydrolipoamide dehydrogenase
79366	HMGN5		HA-	NO	high-mobility group nucleosome binding domain 5
30820	KCNIP1		HA-	NO	Kv channel interacting protein 1
5547	PRCP		HA-	NO	prolylcarboxypeptidase (angiotensinase C)
1486	CTBS		HA-	NO	chitobiase, di-N-acetyl-
987	LRBA		HA+	YES	LPS-responsive vesicle trafficking, beach and anchor containing
201456	FBXO15		HA-	NO	F-box protein 15
3707	ITPKB		HA+	YES	inositol 1,4,5-trisphosphate 3-kinase B
22998	LIMCH1		HA-	NO	LIM and calponin homology domains 1
100132167	LOC100132167		HA+	NO	hypothetical LOC100132167
64778	FNDC3B		HA-	YES	fibronectin type III domain containing 3B
8878	SQSTM1		HA-	YES	sequestosome 1
7456	WIPF1		HA-	YES	WAS/WASL interacting protein family, member 1
11282	MGAT4B		HA+	NO	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme B
55632	G2E3		HA-	YES	G2/M-phase specific E3 ubiquitin protein ligase
9328	GTF3C5		HA+	NO	general transcription factor IIIC, polypeptide 5, 63kDa
57062	DDX24		HA-	YES	DEAD (Asp-Glu-Ala-Asp) box polypeptide 24
9448	MAP4K4		HA-	YES	mitogen-activated protein kinase kinase kinase kinase 4
7748	ZNF195		HA-	YES	zinc finger protein 195
4853	NOTCH2		HA-	YES	notch 2
79807	GSTCD		HA-	YES	glutathione S-transferase, C-terminal domain containing
9898	UBAP2L		HA-	YES	ubiquitin associated protein 2-like
4430	MYO1B		HA-	YES	myosin IB
9813	KIAA0494		HA-	YES	KIAA0494
860	RUNX2		HA-	YES	runt-related transcription factor 2
79627	OGFRL1		HA-	YES	opioid growth factor receptor-like 1
3845	KRAS		HA-	YES	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
55109	AGGF1		HA-	YES	angiogenic factor with G patch and FHA domains 1
64761	PARP12		HA-	YES	poly (ADP-ribose) polymerase family, member 12
5229	PGGT1B		HA-	YES	protein geranylgeranyltransferase type I, beta subunit
23529	CLCF1		HA+	NO	cardiotrophin-like cytokine factor 1
55610	CCDC132		HA-	YES	coiled-coil domain containing 132
23301	EHBP1		HA-	YES	EH domain binding protein 1
55320	MIS18BP1		HA-	YES	MIS18 binding protein 1
162394	SLFN5		HA-	YES	schlafen family member 5
1457	CSNK2A1		HA-	YES	casein kinase 2, alpha 1 polypeptide
10018	BCL2L11		HA+	NO	BCL2-like 11 (apoptosis facilitator)
6734	SRPR		HA-	YES	signal recognition particle receptor (docking protein)
4141	MARS		HA-	YES	methionyl-tRNA synthetase
7803	PTP4A1		HA-	YES	protein tyrosine phosphatase type IVA, member 1
57509	MTUS1		HA-	YES	microtubule associated tumor suppressor 1
92689	FAM114A1		HA-	YES	family with sequence similarity 114, member A1
827	CAPN6		HA-	YES	calpain 6
83856	FSD1L		HA-	YES	fibronectin type III and SPRY domain containing 1-like
2058	EPRS		HA-	YES	glutamyl-prolyl-tRNA synthetase
255488	RNF144B		HA-	YES	ring finger protein 144B
4026	LPP		HA-	YES	LIM domain containing preferred translocation partner in lipoma
25800	SLC39A6		HA-	YES	solute carrier family 39 (zinc transporter), member 6
2335	FN1		HA-	YES	fibronectin 1
467	ATF3		HA-	YES	activating transcription factor 3
90102	PHLDB2		HA-	YES	pleckstrin homology-like domain, family B, member 2
7128	TNFAIP3		HA-	YES	tumor necrosis factor, alpha-induced protein 3
7498	XDH		HA-	YES	xanthine dehydrogenase

FIGURE 4 | Modulated genes with persistent histone acetylation (HA) modification after OTA washout. These 85 genes had (1) significantly altered mRNA levels at R3 (± 0.58 LFC) and (2) persistent HA modifications from T5 to R3. Within this list, 44 genes had HA modification consistent with the direction of GE (correlation: YES, i.e., up-regulated GE/HA+ or down-regulated GE/HA-) and 41 had inconsistent HA modifications (correlation: NO). Most genes had hypo-acetylated histones. Red corresponds to up-regulated GE. Green corresponds to down-regulated GE.

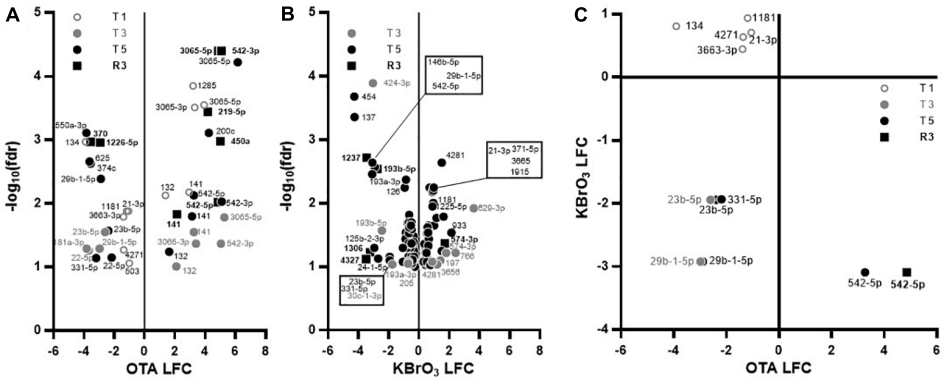


FIGURE 5 | Effects on miRNA expression in RPTEC/TERT1 cells. **(A,B):** Differentially expressed miRNAs in OTA- **(A)** and $KBrO_3$- **(B)** treated cells. Significance of change compared to time-matched control was set to $p \leq 0.1$ based on false discovery rate (fdr) and plotted as $\log_2$ fold change (LFC) vs $-\log_{10}$(fdr). There was no significant change at T1 in $KBrO_3$-treated cells. **(C):** Relative expression of miRNAs significantly impacted by $KBrO_3$ at T5 and by OTA at the indicated time point, regardless of the direction. Changes occurring with one chemical only are not represented.

Regarding purine metabolites, OTA caused a depletion in adenosine (T1, T3), AMP (T3), guanosine (T3) and 5′-methylthioadenosine (5′-MTA; T1, T3, T5, also in the polyamine pathway). OTA exposure also resulted in an accumulation of purine degradation products (hypoxanthine and xanthine) at T5, while the levels of both were back to control level after washout. The levels of adenine were not affected by either chemical. $KBrO_3$ decreased the levels of adenosine and guanosine at T3 and caused a mild increase in AMP and inosine after washout. Contrary to OTA, $KBrO_3$ caused a decrease in xanthine levels at T3 and T5.

A set of metabolites involved in the urea cycle was particularly impacted by OTA treatment, as well as after recovery (**Figure 7**). OTA decreased the levels of ornithine, aspartic acid and fumaric acid without affecting the levels of urea itself. In addition, the levels of putrescine, a polyamine metabolite, were alternately increased at T1, strongly decreased at T3 and T5 and again increased at R3. $KBrO_3$ had a different effect on urea cycle metabolites, affecting primarily citrulline (decreased at T3 and T5) and urea itself (increased at T5 and R3). $KBrO_3$ also impacted the levels of putrescine, although to a lower extent. The metabolite 5′-MTA, a product of spermidine and spermine metabolism, was decreased by OTA at all exposure time points and levels were recovered after OTA washout. $KBrO_3$ did not impact the levels of 5′-MTA.

DISCUSSION

Renal proximal tubule cells are the primary target of OTA toxicity, likely due to basolateral organic anion transport at this site (Tsuda et al., 1999). Previous studies have shown that OTA induces a severe alteration of gene expression *in vitro* in proximal tubule cell models and *in vivo* in the rat renal cortex (Jennings et al., 2012). However, despite a strong impact on the transcriptome, the common toxicologically relevant pathways were not directly impacted with exception of an unusual suppression of the Nrf2 pathway (Limonciel and Jennings, 2013). Thus, transcriptomics alone, particularly in single dose applications, does not reveal a clear mechanism of OTA induced nephrotoxicity and/or carcinogenicity. Here we investigated molecular perturbations induced by OTA at the epigenetic and metabolic levels in repeat dose exposures and in recovery experiments. The effects of OTA were compared and contrasted to those induced by $KBrO_3$, a well described nephrotoxin and renal carcinogen with a firm mechanism of toxicity based on oxidative stress and genotoxicity (Limonciel et al., 2012).

Ochratoxin A exhibited cytotoxicity in repeat dose exposures at 5 μM and above at T5 and was even more cytotoxic at these concentrations after recovery (R3). Cellular stress, as measured by increased lactate production (Limonciel et al., 2011), occurred at T3 at the chosen concentration of 130 nM. This concentration, while non-cytotoxic, also inhibited dome formation, an indicator of vectorial transport of water and solutes in proximal tubule cells grown on solid support (Wilmes et al., 2014) by T3. Transport function fully recovered at R3, as evidenced by the reappearance of domes. Measurement of OTA concentrations within the cells, showed a peak of the parent compound at T1, decreasing at T3 and T5. Approx. 1 % of the T1 peak was detected at R3. It is likely that OTA is metabolized quickly to OTAα, which peaked intracellularly at T3 and was only slightly above the limit of detection after recovery. For comparison, a 0.8 mM non-cytotoxic concentration of $KBrO_3$ was chosen for the omic studies. This concentration exhibited similar effects on the cells, including elevated supernatant lactate and mildly decreased transport capacity at T3, with full apparent recovery at R3.

Ochratoxin A at 130 nM exhibited a more severe effect on gene expression, histone acetylation and DNA methylation than $KBrO_3$. While gene expression somewhat recovered after removal of OTA, both histone hyper-acetylation and DNA hyper-methylation peaked. For both compounds, there was a correlation of histone acetylation to gene expression, however, the correlation was much weaker for OTA. The strongest correlation of the data set was $KBrO_3$ at T1, which gradually decreased with repeated dose and increased again in recovery. The opposite

FIGURE 6 | Epigenetic and mRNA changes associated with stress response pathways. Three stress response pathways are described: the Nrf2 response to oxidative and alkylating stress, the p53 response to DNA damage and the unfolded protein response. Gene lists include the governing transcription factors Nrf2 (NFE2L2), p53 (TP53), ATF4, ATF6 and XBP1, followed by some of their target genes. For HA and DM, "+" means hyper- and "–" means hypo-modifications compared to time-matched control (in all 3 replicates). Red symbolizes modifications that classically promote gene expression for HA and DM / mRNA up-regulation. Green symbolizes modifications that classically reduce gene expression for HA and DM / mRNA down-regulation. Heatmap for mRNA expression ranges from 4 to -2 and is centered to 0 (white). "T average" is the average log2 fold change compared to time-matched control. For each pathway, genes are sorted in descending order of T average in the OTA dataset.

was true for OTA, with a poor correlation at T1 that increased at T3 and T5. This analysis suggests that histone acetylation is not the driving force for OTA-induced gene transcription. This disassociation of histone acetylation and gene expression may point to a mechanism of toxicity of OTA. Regarding DNA methylation, neither compound showed a consistent anti-correlation with gene expression.

Ochratoxin A exposure exhibited a minor effect on miRNA expression, affecting 10 miRNAs altogether, whereas $KBrO_3$ exposure resulted in a differential expression of 35 miRNAs at T5. A possible explanation for OTA's imbalance in mRNA / miRNA alterations is an inhibition of the miRNA maturation machinery, as previously reported (Dai et al., 2014; Zhao et al., 2017). Previous reports on the effect of OTA on miRNA expression in HEK293 and HepG2 cells (Zhao et al., 2017) and in GC-2 cells (Chen et al., 2015) have no overlaps with our study. However, the study by Hennemeier et al. (2014) demonstrates the implication of miR-29b (down-regulated by both OTA and $KBrO_3$ in our study) in collagen formation in HEK293 cells. In addition, in our study OTA induced the expression of several

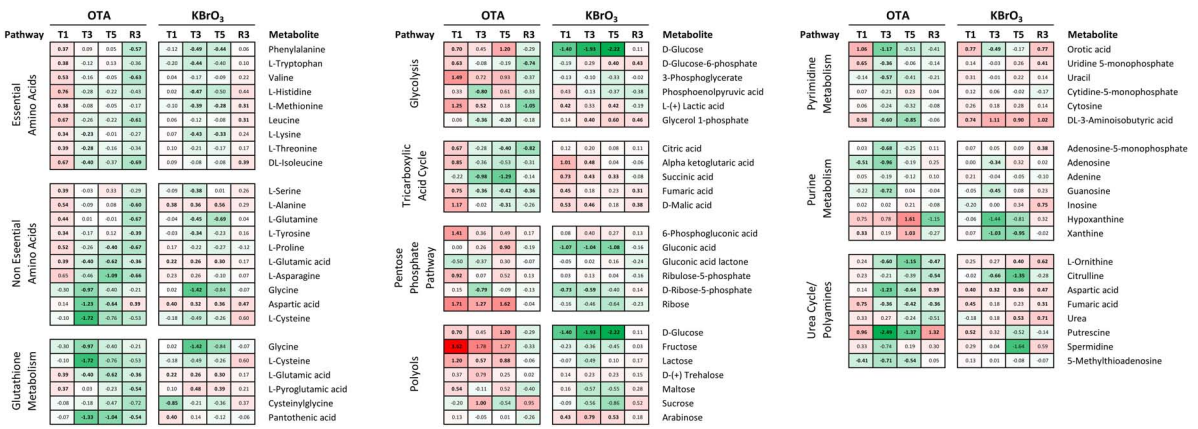


FIGURE 7 | Effects on metabolite levels in RPTEC/TERT1 cell lysates. Intracellular metabolites were measured in methanol lysates by GC-MS at the four indicated time points. Data is shown as log2 fold over time-matched control (LFC). Metabolites are grouped per pathway and appear multiple times if shared between pathways. Heatmap ranges from 3.62 (fructose / OTA T1, red) to –2.22 (glucose / $KBrO_3$ T5) and is centered to 0 (white). Bold LFCs indicate LFC significantly different from control, as determined by 2-way ANOVA with Sidak posttest ($p \leq 0.05$).

miRNAs at all treatment time points including miR-3065-5p, miR-141, miR-542-5p, miR-542-3p and miR-132. miR-132 is of potential mechanistic interest as it has previously been shown to be involved in the suppression of Nrf2 genes, with a miR-132 antigomir being capable of preventing OTA-induced Nrf2 mRNA depletion in LLC-PK1 cells (Stachurska et al., 2013).

Indeed, analysis of the OTA-induced transcriptome alterations demonstrated that Nrf2 target genes HMOX1, NQO1, GCLM, TXNRD1 and SRXN1 were severely attenuated whereas all were robustly induced by the oxidant and Nrf2 activator $KBrO_3$. It has been well-documented that OTA can induce reactive oxygen species (ROS) (Schaaf et al., 2002; Costa et al., 2016). Thus, it is counter-intuitive that the Nrf2 response that protects the cell against oxidative stress be attenuated by a ROS-inducing chemical. However, this is a striking finding in this study and has been reported by us and several other groups (Limonciel and Jennings, 2013). The mechanism of OTA-induced Nrf2 response inhibition is not clear and is potentially based on inhibition of Nrf2 translocation, induction of miR-132, inhibition of protein acetylation through HDAC activation and HAT inhibition or combinations of all. In any case, it is plausible that OTA-induced Nrf2 inhibition renders the cell defenseless to oxidative injury, potentially leading to increased cell death rates and cancer.

Within the UPR pathway, OTA induced ATF6 and HSPA5 (aka BiP) transcription, but suppressed ATF4 and many of its target genes, including tRNA synthetases (YARS, AARS, LARS, SARS, VARS, TARS, EPRS, GARS, CARS, NARS, WARS, IARS) and amino acid transporters (SLC1A5, SLC1A4, SLC6A9, SLC7A11). All of these may point to a general increase in protein turnover. Indeed OTA's strong affinity for serum albumin is responsible for its high plasma half-life of up to 35 days (Studer-Rohr et al., 2000). It is conceivable that OTA also binds to cytosolic and cytoskeletal proteins with high affinity initiating proteasomal degradation and autophagy.

We have previously demonstrated an interaction of the Nrf2 and ATF4 pathways in the maintenance of glutathione levels after oxidative injury (Wilmes et al., 2013). Since Nrf2 also induces mRNA expression of ATF4, it is possible that inhibition of the Nrf2 pathway also suppresses the ATF4 branch of the UPR. In the UPR, ATF4 primarily orchestrates the expression of amino acid transporters and aminoacyl-tRNA synthetase enzymes that attach amino acids to their specific tRNAs for inclusion in newly translated proteins (Jennings et al., 2013). Metabolomic analysis showed an OTA-induced increase in all essential amino acids at T1, which could result from an abrupt interruption of translation at the beginning of exposure or an increase in amino acid transport from the cell culture medium that was not sustained at later time points. For non-essential amino acids, however, while most metabolites were increased at T1, many were strongly decreased at all the other time points, including R3. In particular at T3, the glutathione building blocks glycine and cysteine were decreased, suggesting an impact on the capacity of the cells for *de novo* glutathione synthesis.

$KBrO_3$ exhibited a strong induction of genes in the p53 pathway. OTA caused a weaker response although some p53 genes were robustly induced, including CDKN1A (p21) and GADD45A. The p53 pathway is an important regulator of many processes including DNA damage and glycolysis. Although both chemicals increased lactate production, OTA and $KBrO_3$ had very different impacts on other metabolites involved in glucose metabolism through glycolysis (of which lactate is the final metabolite), the tricarboxylic acid (TCA) cycle, the PPP and the polyol pathway. In the latter, glucose is converted to sorbitol by aldo keto reductases (AKR1B1, AKR1B10) and then to fructose by sorbitol dehydrogenase (SORD). In OTA-treated cells, fructose was consistently increased, suggesting an activation of the polyol pathway, as fructose is not present in the cell culture medium used. In addition, SORD, encoding the enzyme that converts sorbitol to fructose, was amongst the strongest up-regulated genes in the OTA gene expression dataset. Its log2 FC was 5.0 at T1, 5.3 at T3, 5.0 at T5 and still 1.0 at R3 (2 folds above time-matched control).

Another interesting aspect of the metabolomic dataset was the impact on nucleotide metabolism that could interfere with the availability of nucleotides for mRNA synthesis or reflect an attempt at *de novo* nucleotide biosynthesis, possibly to support DNA repair mechanisms.

The polyamine metabolism pathway was also particularly affected by OTA. Putrescine, a metabolite of ornithine, was impacted by OTA at all time points. However, putrescine is available from the cell culture medium, thus the increases measured at T1 and R3 could be the result of increased uptake by the cells. The decreases in putrescine levels at T3 and T5, however, suggest an interference with its further degradation to spermidine, of which the levels were not significantly changed by OTA. 5′-MTA, a by-product of spermidine and spermine synthesis, on the other hand was consistently decreased during OTA exposure. 5′-MTA has many roles such as being an inhibitor of polyamine synthesis (Evans et al., 2004), as a starting point for the purine and methionine salvage pathway (Williams-Ashman et al., 1982) and as a methyl donor for methylation of other molecules (Avila et al., 2004). This last property is of particular interest in the context of DNA methylation, as OTA has been shown to cause a global hypomethylation of DNA in HepG2 cells (Zheng et al., 2013).

As a food contaminant of great concern, OTA has been at the center of several studies focusing on the use of metabolomics to identify new biomarkers of exposure in blood and urine. Male rats exposed to up to 210 μg OTA/kg bw by gavage for up to 90 days had elevated glucose, lactate, alanine and glycine levels in the urine, while citrate and oxoglutarate levels were decreased compared to control (Sieber et al., 2009). Another study with similar exposure up to 26 weeks found elevated levels of alanine and threonine in the rats' blood associated with low blood glucose and high lactate (Xia et al., 2014). This study also found high levels of fumarate, malate (increased at T1 in our study), ribose (increased throughout treatment), uridine, sorbitol, fructose (increased at T1), aspartic acid (decreased at late treatment time points, increased at R3), leucine, serine, proline (all 3 increased at T1) and ornithine (decreased at late treatment time points and R3) in the biofluids analyzed. A single dose of OTA administered to male rats (6.25 mg OTA) was also shown to cause a modulation in the levels of citrate and an increase in oxoglutarate, lactate, glucose and succinate levels in the urine (Mantle et al., 2011). Although these studies focus on extracellular fluids, our analysis of the intracellular contents shows an overlap for many of the features identified as potential biomarkers of OTA exposure *in vivo*. Metabolomic analysis was previously performed both in cell lysates and supernatants of RPTEC/TERT1 exposed to both chemicals for up to 3 days (1mM $KBrO_3$ / 300 nM OTA) (Ellis et al., 2011). After a bolus exposure, extracellular lactate and pyruvate levels were increased (intracellular levels were unchanged), while glucose was depleted from the medium with both chemicals. Intracellular betaine was decreased by both chemicals. In addition, $KBrO_3$ caused an increase in alanine, glycerophosphocholine and total glutathione within the cells. Thus the features identified in the Ellis et al. study further support the deep interference with energy metabolism identified for both chemicals in our study.

Taken together these results support previous reports of the effect of OTA on metabolic processes related to protein synthesis (amino acid availability), nucleotide synthesis and energy metabolism, as well as the effect of both chemicals on stress responses to oxidative stress and DNA damage (activated by $KBrO_3$, inhibited by OTA). The exhaustive metabolomic investigation is concordant with previous reports of the effects of OTA and brings further insights on the effects of $KBrO_3$ on the metabolome. While the large effect of OTA on gene expression and epigenomic regulation had been previously reported, we show here that the effects on histone acetylation and DNA methylation, do not appear to be a driving force in the large transcriptional impact of OTA in renal proximal tubule cells. It is likely that OTA uptake into the cell initiates several simultaneous perturbations including proteotoxicity, disturbances of the histone machinery, Nrf2 inhibition, DNA injury and perturbations of glucose catabolism. Further work will be needed to delineate these mechanisms and to uncouple which mechanisms are direct OTA effects and which are compensatory mechanisms.

AUTHOR CONTRIBUTIONS

The experiments were designed by AL and PJ. The cell culture experiments were conducted by AL, AW, and LA. miRNA and epigenomic samples analysis was coordinated by SvB, TdK, and JK. Metabolomic samples analysis was coordinated by GT, APS, and HK. mRNA samples measurement was coordinated by AS. Transcriptomic data normalization and correlation to epigenomics was conducted by XJ and AK-S. Data was analyzed and integrated by AL. Manuscript was primarily written by AL and PJ with contribution from all co-authors.

FUNDING

The study was funded by the 7th Framework project DETECTIVE (grant no. 266838 to PJ, JK, AS, HK, and AK-S), the Long Range Initiative Innovative Science Award of the European Chemical Industry Council (CEFIC, 2015 to AL), the Horizon 2020 project EU-ToxRisk (http://www.eu-toxrisk.eu/ grant no. 681002, to PJ) and the Tiroler Wissenschaftfund (Grant no. UNI-0404/1768, to AW).

ACKNOWLEDGMENTS

The authors would like to thank Margit Henry and Tamara Rothstein, University of Cologne, for the Affymetrix microarray analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00558/full#supplementary-material

REFERENCES

Aschauer, L., Gruber, L. N., Pfaller, W., Limonciel, A., Athersuch, T. J., Cavill, R., et al. (2013). Delineation of the key aspects in the regulation of epithelial monolayer formation. *Mol. Cell. Biol.* 33, 2535–2550. doi: 10.1128/MCB.01435-12

Avila, M. A., García-Trevijano, E. R., Lu, S. C., Corrales, F. J., and Mato, J. M. (2004). Methylthioadenosine. *Int. J. Biochem. Cell Biol.* 36, 2125–2130. doi: 10.1016/j.biocel.2003.11.016

Ballmaier, D., and Epe, B. (2006). DNA damage by bromate: mechanism and consequences. *Toxicology* 221, 166–171. doi: 10.1016/j.tox.2006.01.009

Behrends, V., Tredwell, G. D., and Bundy, J. G. (2011). A software complement to AMDIS for processing GC-MS metabolomic data. *Anal. Biochem.* 415, 206–208. doi: 10.1016/j.ab.2011.04.009

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57, 289–300.

Campbell, K. C. M. (2006). Bromate-induced ototoxicity. *Toxicology* 221, 205–211. doi: 10.1016/j.tox.2005.12.015

Chen, R., Deng, L., Yu, X., Wang, X., Zhu, L., Yu, T., et al. (2015). MiR-122 partly mediates the ochratoxin A-induced GC-2 cell apoptosis. *Toxicol. Vitro* 30, 264–273. doi: 10.1016/j.tiv.2015.10.011

Cho, D. H., Hong, J. T., Chin, K., Cho, T. S., and Lee, B. M. (1993). Organotropic formation and disappearace of 8-hydroxydeoxyguanosine in the kidney of Sprague-Dawley rats exposed to Adriamycin and KBrO3. *Cancer Lett.* 74, 141–145. doi: 10.1016/0304-3835(93)90235-2

Clark, H. A., and Snedeker, S. M. (2006). Ochratoxin A: its cancer risk and potential for exposure. *J. Toxicol. Environ. Health B Crit. Rev.* 9, 265–296. doi: 10.1080/15287390500195570

Coonen, M. L. J., Theunissen, D. H. J., Kleinjans, J. C. S., and Jennen, D. G. J. (2015). MagiCMicroRna: a web implementation of AgiMicroRna using shiny. *Source Code Biol. Med.* 10:4. doi: 10.1186/s13029-015-0035-5

Costa, J. G., Saraiva, N., Guerreiro, P. S., Louro, H., Silva, M. J., Miranda, J. P., et al. (2016). Ochratoxin A-induced cytotoxicity, genotoxicity and reactive oxygen species in kidney cells: an integrative approach of complementary endpoints. *Food Chem. Toxicol.* 87, 65–76. doi: 10.1016/j.fct.2015.11.018

Czakai, K., Müller, K., Mosesso, P., Pepe, G., Schulze, M., Gohla, A., et al. (2011). Perturbation of mitosis through inhibition of histone acetyltransferases: the key to ochratoxin a toxicity and carcinogenicity? *Toxicol. Sci.* 122, 317–329. doi: 10.1093/toxsci/kfr110

Dai, Q., Zhao, J., Qi, X., Xu, W., He, X., Guo, M., et al. (2014). MicroRNA profiling of rats with ochratoxin A nephrotoxicity. *BMC Genomics* 15:333. doi: 10.1186/1471-2164-15-333

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., et al. (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* 6, 1060–1083. doi: 10.1038/nprot.2011.335

Ellis, J. K., Athersuch, T. J., Cavill, R., Radford, R., Slattery, C., Jennings, P., et al. (2011). Metabolic response to low-level toxicant exposure in a novel renal tubule epithelial cell system. *Mol. Biosyst.* 7, 247–257. doi: 10.1039/c0mb00146e

Evans, G. B., Furneaux, R. H., Schramm, V. L., Singh, V., and Tyler, P. C. (2004). Targeting the polyamine pathway with transition-state analogue inhibitors of 5′-methylthioadenosine phosphorylase. *J. Med. Chem.* 47, 3275–3281. doi: 10.1021/jm0306475

Fujie, K., Shimazu, H., Matsuda, M., and Sugiyama, T. (1988). Acute cytogenetic effects of potassium bromate on rat bone marrow cells in vivo. *Mutat. Res. Toxicol.* 206, 455–458. doi: 10.1016/0165-1218(88)90053-5

Hayashi, M., Kishi, M., Sofuni, T., and Ishidate, M. (1988). Micronucleus tests in mice on 39 food additives and eight miscellaneous chemicals. *Food Chem. Toxicol.* 26, 487–500. doi: 10.1016/0278-6915(88)90001-4

Hennemeier, I., Humpf, H. U., Gekle, M., and Schwerdt, G. (2014). Role of microRNA-29b in the ochratoxin A-induced enhanced collagen formation in human kidney cells. *Toxicology* 324, 116–122. doi: 10.1016/j.tox.2014.07.012

International Agency for Research on Cancer (2018). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. Lyon: IARC.

Ishidate, M., Sofuni, T., Yoshikawa, K., Hayashi, M., Nohmi, T., Sawada, M., et al. (1984). Primary mutagenicity screening of food additives currently used in Japan. *Food Chem. Toxicol.* 22, 623–636. doi: 10.1016/0278-6915(84)90271-0

Jennings, P., Limonciel, A., Felice, L., and Leonard, M. O. (2013). An overview of transcriptional regulation in response to toxicological insult. *Arch. Toxicol.* 87, 49–72. doi: 10.1007/s00204-012-0919-y

Jennings, P., Weiland, C., Limonciel, A., Bloch, K. M., Radford, R., Aschauer, L., et al. (2012). Transcriptomic alterations induced by Ochratoxin A in rat and human renal proximal tubular in vitro models and comparison to a rat in vivo model. *Arch. Toxicol.* 86, 571–589. doi: 10.1007/s00204-011-07804

Kasai, H., Nishimura, S., Kurokawa, Y., and Hayashi, Y. (1987). Oral administration of the renal carcinogen, potassium bromate, specifically produces 8-hydroxydeoxyguanosine in rat target organ dna. *Carcinogenesis* 8, 1959–1961. doi: 10.1093/carcin/8.12.1959

Kind, T., Wohlgemuth, G., Lee, D. Y., Lu, Y., Palazoglu, M., Shahbaz, S., et al. (2009). FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* 81, 10038–10048. doi: 10.1021/ac9019522

Limonciel, A., Aschauer, L., Wilmes, A., Prajczer, S., Leonard, M. O., Pfaller, W., et al. (2011). Lactate is an ideal non-invasive marker for evaluating temporal alterations in cell stress and toxicity in repeat dose testing regimes. *Toxicol. Vitro* 25, 1855–1862. doi: 10.1016/j.tiv.2011.05.018

Limonciel, A., Ates, G., Carta, G., Wilmes, A., Watzele, M., Shepard, P. J., et al. (2018). Comparison of base-line and chemical-induced transcriptomic responses in HepaRG and RPTEC/TERT1 cells using TempO-Seq. *Arch. Toxicol.* 92, 2517–2531. doi: 10.1007/s00204-018-2256-2

Limonciel, A., and Jennings, P. (2013). A review of the evidence that ochratoxin A is an Nrf2 inhibitor: implications for nephrotoxicity and renal carcinogenicity. *Toxins* 6, 371–379. doi: 10.3390/toxins6010371

Limonciel, A., Moenks, K., Stanzel, S., Truisi, G. L., Parmentier, C., Aschauer, L., et al. (2015). Transcriptomics hit the target: monitoring of ligand-activated and stress response pathways for chemical testing. *Toxicol. Vitro* 30, 7–18. doi: 10.1016/j.tiv.2014.12.011

Limonciel, A., Wilmes, A., Aschauer, L., Radford, R., Bloch, K. M., McMorrow, T., et al. (2012). Oxidative stress induced by potassium bromate exposure results in altered tight junction protein expression in renal proximal tubule cells. *Arch. Toxicol.* 86, 1741–1751. doi: 10.1007/s00204-012-0897-0

López-Romero, P. (2011). Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. *BMC Genomics* 12:64. doi: 10.1186/1471-2164-12-64

Mantle, P. G., Nicholls, A. W., and Shockcor, J. P. (2011). H NMR spectroscopy-based metabolomic assessment of uremic toxicity, with toxicological outcomes, in male rats following an acute, mid-life insult from ochratoxin A. *Toxins* 3, 504–519. doi: 10.3390/toxins3060504

Marin-Kuan, M., Nestler, S., Verguet, C., Bezençon, C., Piguet, D., Mansourian, R., et al. (2006). A toxicogenomics approach to identify new plausible epigenetic mechanisms of ochratoxin a carcinogenicity in rat. *Toxicol. Sci.* 89, 120–134. doi: 10.1093/toxsci/kfj017

Sai, K., Uchiyama, S., Ohno, Y., Hasegawa, R., and Kurokawa, Y. (1992). Generation of active oxygen species in vitro by the interaction of potassium bromate with rat kidney cell. *Carcinogenesis* 13, 333–339. doi: 10.1093/carcin/13.3.333

Schaaf, G. J., Nijmeijer, S. M., Maas, R. F. M., Roestenberg, P., De Groene, E. M., and Fink-Gremmels, J. (2002). The role of oxidative stress in the ochratoxin A-mediated toxicity in proximal tubular cells. *Biochim. Biophys. Acta Mol. Basis Dis.* 1588, 149–158. doi: 10.1016/S0925-4439(02)00159-X

Sieber, M., Wagner, S., Rached, E., Amberg, A., Mally, A., and Dekant, W. (2009). Metabonomic study of ochratoxin A toxicity in rats after repeated administration: phenotypic anchoring enhances the ability for biomarker discovery. *Chem. Res. Toxicol.* 22, 1221–1231. doi: 10.1021/tx800459q

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, 1–25. doi: 10.2202/1544-6115.1027

Smyth, G. K. (2005). "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber (New York, NY: Springer), 397–420. doi: 10.1007/0-387-29362-0_23

Stachurska, A., Ciesla, M., Kozakowska, M., Wolffram, S., Boesch-Saadatmandi, C., Rimbach, G., et al. (2013). Cross-talk between microRNAs, nuclear factor E2-related factor 2, and heme oxygenase-1 in ochratoxin A-induced toxic effects in renal proximal tubular epithelial cells. *Mol. Nutr. Food Res.* 57, 504–515. doi: 10.1002/mnfr.201200456

Stein, S. E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.* 10, 770–781. doi: 10.1016/S1044-0305(99)00047-1

Studer-Rohr, I., Schlatter, J., and Dietrich, D. R. (2000). Kinetic parameters and intraindividual fluctuations of ochratoxin A plasma levels in humans. *Arch. Toxicol.* 74, 499–510. doi: 10.1007/s002040000157

Tsuda, M., Sekine, T., Takeda, M., Cha, S. H., Kanai, Y., Kimura, M., et al. (1999). Transport of ochratoxin A by renal multispecific organic anion transporter 1. *J. Pharmacol. Exp. Ther.* 289, 1301–1305.

van Breda, S. G. J., Claessen, S. M. H., Lo, K., van Herwijnen, M., Brauers, K. J. J., Lisanti, S., et al. (2014). Epigenetic mechanisms underlying arsenic-associated lung carcinogenesis. *Arch. Toxicol.* 89, 1959–1969. doi: 10.1007/s00204-014-1351-2

Vettorazzi, A., van Delft, J., and López de Cerain, A. (2013). A review on ochratoxin A transcriptomic studies. *Food Chem. Toxicol.* 59, 766–783. doi: 10.1016/j.fct.2013.05.043

Wieser, M., Stadler, G., Jennings, P., Streubel, B., Pfaller, W., Ambros, P., et al. (2008). hTERT alone immortalizes epithelial cells of renal proximal tubules without changing their functional characteristics. *Am. J. Physiol. Renal Physiol.* 295, F1365–F1375. doi: 10.1152/ajprenal.90405.2008

Williams-Ashman, H. G., Seidenfeld, J., and Galletti, P. (1982). Trends in the biochemical pharmacology of 5′-deoxy-5′-methylthioadenosine. *Biochem. Pharmacol.* 31, 277–288. doi: 10.1016/0006-2952(82)90171-X

Wilmes, A., Aschauer, L., Limonciel, A., Pfaller, W., and Jennings, P. (2014). Evidence for a role of claudin 2 as a proximal tubular stress responsive paracellular water channel. *Toxicol. Appl. Pharmacol.* 279, 163–172. doi: 10.1016/j.taap.2014.05.013

Wilmes, A., Limonciel, A., Aschauer, L., Moenks, K., Bielow, C., Leonard, M. O., et al. (2013). Application of integrated transcriptomic, proteomic and metabolomic profiling for the delineation of mechanisms of drug induced cell stress. *J. Proteomics* 79, 180–194. doi: 10.1016/j.jprot.2012.11.022

Xia, K., He, X., Dai, Q., Cheng, W. H., Qi, X., Guo, M., et al. (2014). Discovery of systematic responses and potential biomarkers induced by ochratoxin A using metabolomics. *Food Addit. Contam. Part A Chem. Anal. Control. Expo. Risk Assess.* 31, 1904–1913. doi: 10.1080/19440049.2014.957249

Zhang, X., De Silva, D., Sun, B., Fisher, J., Bull, R. J., Cotruvo, J. A., et al. (2010). Cellular and molecular mechanisms of bromate-induced cytotoxicity in human and rat kidney cells. *Toxicology* 269, 13–23. doi: 10.1016/j.tox.2010.01.002

Zhao, J., Qi, X., Dai, Q., He, X., Dweep, H., Guo, M., et al. (2017). Toxicity study of ochratoxin A using HEK293 and HepG2 cell lines based on microRNA profiling. *Hum. Exp. Toxicol.* 36, 8–22. doi: 10.1177/0960327116632048

Zheng, J., Zhang, Y., Xu, W., Luo, Y., Hao, J., Shen, X. L., et al. (2013). Zinc protects HepG2 cells against the oxidative damage and DNA damage induced by ochratoxin A. *Toxicol. Appl. Pharmacol.* 268, 123–131. doi: 10.1016/j.taap.2013.01.021

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

HYPOTHESIS AND THEORY
published: 21 December 2018
doi: 10.3389/fgene.2018.00661

Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials

Marvin Martens[1], Tim Verbruggen[1], Penny Nymark[2,3], Roland Grafström[2,3], Lyle D. Burgoon[4], Hristo Aladjov[5], Fernando Torres Andón[6,7], Chris T. Evelo[1,8] and Egon L. Willighagen[1]**

[1] Department of Bioinformatics – BiGCaT, NUTRIM, Maastricht University, Maastricht, Netherlands, [2] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden, [3] Department of Toxicology, Misvik Biology, Turku, Finland, [4] U.S. Army Engineer Research and Development Center, Vicksburg, MS, United States, [5] Organisation for Economic Co-operation and Development Environment Directorate, Paris, France, [6] Laboratory of Cellular Immunology, Humanitas Clinical and Research Institute, Rozzano, Italy, [7] Center for Research in Molecular Medicine and Chronic Diseases, University of Santiago de Compostela, Santiago de Compostela, Spain, [8] Maastricht Centre for Systems Biology, Maastricht University, Maastricht, Netherlands

OPEN ACCESS

Edited by:
Paul Jennings,
VU Amsterdam, Netherlands

Reviewed by:
Mark Cronin,
Liverpool John Moores University, United Kingdom
Frederic Y. Bois,
French National Institute for Industrial Environment and Risks, France

***Correspondence:**
Marvin Martens
marvin.martens@maastrichtuniversity.nl
Egon L. Willighagen
egon.willighagen@maastrichtuniversity.nl

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 07 July 2018
Accepted: 04 December 2018
Published: 21 December 2018

Citation:
Martens M, Verbruggen T, Nymark P, Grafström R, Burgoon LD, Aladjov H, Torres Andón F, Evelo CT and Willighagen EL (2018) Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials. Front. Genet. 9:661. doi: 10.3389/fgene.2018.00661

A paradigm shift is taking place in risk assessment to replace animal models, reduce the number of economic resources, and refine the methodologies to test the growing number of chemicals and nanomaterials. Therefore, approaches such as transcriptomics, proteomics, and metabolomics have become valuable tools in toxicological research, and are finding their way into regulatory toxicity. One promising framework to bridge the gap between the molecular-level measurements and risk assessment is the concept of adverse outcome pathways (AOPs). These pathways comprise mechanistic knowledge and connect biological events from a molecular level toward an adverse effect outcome after exposure to a chemical. However, the implementation of omics-based approaches in the AOPs and their acceptance by the risk assessment community is still a challenge. Because the existing modules in the main repository for AOPs, the AOP Knowledge Base (AOP-KB), do not currently allow the integration of omics technologies, additional tools are required for omics-based data analysis and visualization. Here we show how WikiPathways can serve as a supportive tool to make omics data interoperable with the AOP-Wiki, part of the AOP-KB. Manual matching of key events (KEs) indicated that 67% could be linked with molecular pathways. Automatic connection through linkage of identifiers between the databases showed that only 30% of AOP-Wiki chemicals were found on WikiPathways. More loose linkage through gene names in KE and Key Event Relationships descriptions gave an overlap of 70 and 71%, respectively. This shows many opportunities to create more direct connections, for example with extended ontology annotations, improving its interoperability. This interoperability allows the needed integration of omics data linked to the molecular pathways with AOPs. A new AOP Portal on WikiPathways is

presented to allow the community of AOP developers to collaborate and populate the molecular pathways that underlie the KEs of AOP-Wiki. We conclude that the integration of WikiPathways and AOP-Wiki will improve risk assessment because omics data will be linked directly to KEs and therefore allow the comprehensive understanding and description of AOPs. To make this assessment reproducible and valid, major changes are needed in both WikiPathways and AOP-Wiki.

Keywords: adverse outcome pathways, risk assessment, omics, WikiPathways, interoperability

INTRODUCTION

The last decades have seen many developments in risk assessment strategies for an ever-growing number of chemicals and nanomaterials, aiming to reduce the use of animals and cost of risk assessment and to increase the predictive value. In parallel to these changes, experimental approaches in regular toxicology research have also made major steps setting up novel high-throughput technologies for generating large-scale (omics) datasets such as transcriptomics, metabolomics, and proteomics. However, these technologies are not consistently implemented in regulatory risk assessment and there is a need for proper integration of knowledge, testing systems, and analysis tools for these approaches to be of added value over existing methodologies in risk assessment.

To support the paradigm shift toward animal-free, cheap and more effective risk assessments of chemicals, the concept of adverse outcome pathways (AOPs) emerged (Ankley et al., 2010), which integrate mechanistic knowledge of the toxicological effects of chemical compounds and nanomaterials and thereby assist integrated approaches to testing and assessment strategies. AOPs are structured as logical sequences of causally linked and measurable biological events [key events (KEs)] that occur after exposure to a stressor that triggers a biological perturbation, called the molecular initiating event (MIE). These KEs are connected by Key Event Relationships (KERs) and describe the downstream effects on increasing levels of biological organization, from molecular, cellular, tissue, organ, individual, and population responses toward an adverse outcome (AO) (Villeneuve et al., 2014; Leist et al., 2017; Vinken et al., 2017).

The Organisation for Economic Co-operation and Development (OECD) was the first organization to embrace AOPs by launching the AOP Development Programme in 2012 for the establishment of AOPs in a qualitative way and provide guidance material for standardized, structured development of AOPs (Vinken, 2013; Organisation for Economic Co-operation and Development [OECD], 2017). With that, the AOP Knowledge Base (AOP-KB[1]) emerged in 2014 as a collective platform of various tools to assist in the development of AOPs. Its main components are the AOP-Wiki[2], Effectopedia[3] and the AOPXplorer Cytoscape application.

The AOP-Wiki is the result of collaboration between the European Commission's Joint Research Center (JRC) and the United States Environmental Protection Agency (US EPA). It is developed to be a central knowledge-sharing platform which facilitates cooperative development of AOPs and strictly follows the OECD's guidance materials for AOP development. Nowadays, it is the most actively used module of the AOP-KB and with the recent efforts on annotation with ontology tags, it has been aiming for semantic interoperability. This started with the development of the AOP Ontology (Burgoon, 2017) and recently, the addition of various other ontologies to match the various domains described in AOPs, from Gene Ontology for biology annotation toward the Population and Community Ontology for annotation of events on the population level (Ives et al., 2017).

Effectopedia (Watanabe et al., 2018) is another tool from AOP-KB, developed by OECD, dedicated to the collaborative development of quantitative AOPs. The AOP diagram is the focal point of its user interface providing visual means for adding new and navigation through existing AOP elements, offering easy access to their description. In addition to KE and KER, Effectopedia also has an explicit representation of test methods, collected data and executable models. The integration of response data in KER allows the system to predict downstream KEs using measurements or models for upstream KEs that can be measured using *in chemico*, high throughput and or *in vitro* methods. The goal of fully quantified AOPs is to allow the prediction of an adverse outcome in time and magnitude using a minimum number of experimental measurements for KE responses that cannot be adequately modeled by other means.

The third is AOPXplorer, a Cytoscape application, meant for building networks of KEs, forming AOP Networks (AOPNs) and allow data visualization of various types on top of the AOPNs. The goal of AOPXplorer is to help investigators and risk assessors understand how chemical exposures result in information flow throughout the AOPN, allowing them to make defensible stories and inferences about potential adverse outcomes.

It has been postulated that omics technologies can be used for various goals in regulatory toxicology, such as biological read-across based on molecular events to prioritize chemicals for testing, cross-species extrapolation to link to evolutionary biology and the identification of KEs (Hartung, 2016). Although omics approaches have already been used in toxicology to define specific modes of action (Edwards and Preston, 2008) or identifying biomarkers (Grafström et al., 2015), they have not found their way into regulatory acceptance for assessment of chemicals and nanomaterials (Buesen et al., 2017; van Ravenzwaay et al., 2017). There is a need for well-established experimental protocols for data generation, storage, processing,

[1]https://aopkb.oecd.org/index.html

[2]https://aopwiki.org/

[3]https://effectopedia.org

analysis, and interpretation to reach regulatory acceptance. Besides, an integration framework for data interpretation to identify relevant molecular changes and pathways is required, as well as the filling of knowledge gaps that keep risk assessors from causally linking molecular events to an adverse outcome at a higher level of biological organization (Brockmeier et al., 2017; Buesen et al., 2017; Sauer et al., 2017; Vachon et al., 2017; Campos and Colbourne, 2018). Taken together, the level of uncertainties and inconsistencies in experimental design should be minimized to allow omics approaches in risk assessment and AOPs. So far, various ideas have emerged to introduce omics data to the concept of the AOPs, such as a pipeline for KE enrichment (Nymark et al., 2018), workflow for computationally predicted AOPs from public data (Bell et al., 2016) and the Transcriptomics Reporting Framework (Gant et al., 2017).

There is a demand for a consistent, well-defined protocol to analyze and integrate the data in order to describe the molecular effects downstream of an MIE (Brockmeier et al., 2017). Molecular pathway databases and tools exist to analyze omics datasets through pathway analysis, which happens through probability scoring of pathways containing differently expressed genes and thereby reducing the number of dimensions of omics datasets to the number of biological pathways. Various molecular pathway databases exist which could be viable tools for the integration of omics approaches in regulatory risk assessment, such as KEGG (Kanehisa and Goto, 2000), Reactome (Fabregat et al., 2018) and WikiPathways (Slenter et al., 2018).

In this paper, we describe how WikiPathways[4] (Slenter et al., 2018) an open-science molecular pathway database which captures mechanistic knowledge in pathway diagrams, can be a supportive database for AOPs and the analysis and interpretation of omics datasets through pathway analysis. WikiPathways has similar levels of coverage of genes and metabolites as Reactome and KEGG (Kutmon et al., 2016; Slenter et al., 2018) and performs better in covering signaling pathways (Azad et al., 2017). This can be done with PathVisio (Kutmon et al., 2015), a pathway diagram drawing tool that is connected to WikiPathways, in which omics data can be visualized and pathway analysis can be performed. Also, WikiPathways exists as a Cytoscape application, which allows the same pathways to be used for network analysis (Kutmon et al., 2014).

Thanks to the adaptability and accessibility of WikiPathways, communities can collaborate on creating, assessing and improving the understanding of molecular pathways (Pico et al., 2008). Therefore, WikiPathways could be a valuable tool for the risk assessment community. It can provide improved molecular descriptions of early KEs which support biological plausibility. At the same time, it can serve as empirical support to KERs and allow the integration of omics technologies in the concept of AOPs in a systematic manner. As illustrated in **Figure 1**, ideally, all KEs in AOP-Wiki are linked by at least one molecular pathway, which can be highlighted by omics analysis and thereby revealing KEs. However, WikiPathways needs to be integrated with the existing modules in the AOP-KB. Here, we focus on the AOP-Wiki by describing its current implementation of semantic annotations and we will show how we can connect the AOP-Wiki with WikiPathways through identifiers for genes, proteins and metabolites, and ontologies (Bard and Rhee, 2004), which are pre-defined vocabularies used to describe knowledge and assist in the integration of data sources. Furthermore, we will propose a strategy for future work on connecting the two databases, describing the planned work on WikiPathways and suggestions for improving the AOP-Wiki and its contents to allow linkage of databases.

[4]http://wikipathways.org

MATERIALS AND METHODS

Retrieval of AOP-Wiki Data

The AOP-Wiki allows the use of their data for publication purposes, by storing permanent quarterly downloads on the website[5]. For this paper, we used the AOP-Wiki XML file of April 1st, 2018, containing all AOP-Wiki content.

Parsing the AOP-Wiki XML

The AOP-Wiki XML was parsed with Python 3.5 (Python Software Foundation, 2010) and the ElementTree XML API with the ".parse"-function which resulted in an ElementTree wrapper class that represents an entire element hierarchy. The information, that was required for the experiments, was extracted included stressor information, ontology annotations, and information on KEs and KERs. The source code, as well as a brief tutorial on the execution of it, are available on GitHub (Martens, 2018).

BridgeDb Identifier Mapping in R

In order to perform identifier mapping for the chemicals that are stored on AOP-Wiki with CAS Registration Numbers (CAS numbers), we used the BridgeDb, an identifier mapping framework (Van Iersel et al., 2010). The CAS numbers from the AOP-Wiki were saved as plain text file and imported in RStudio (version 1.1.447; R version 3.4.4) (R Core Team, 2013; R Studio Team, 2015), in which the R-package BridgeDbR (Leemans et al., 2018) was utilized to map the CAS numbers to ChEBI identifiers with the BridgeDb metabolite identifier mapping dataset (Slenter, 2018). The R code used for the identifier mapping is available on GitHub along with a tutorial to execute the script (Martens, 2018).

WikiPathways Data

Information from WikiPathways was retrieved using the WikiPathways SPARQL endpoint[6] (Waagmeester et al., 2016), version 20180610. SPARQL is a query language to select specific subsets of data from a collection of RDF, a standard framework for knowledge descriptions. For this manuscript, various queries were performed to request information about WikiPathways' use of ontologies and to retrieve pathways for lists of genes related to KEs.

[5]https://aopwiki.org/downloads

[6]http://sparql.wikipathways.org/

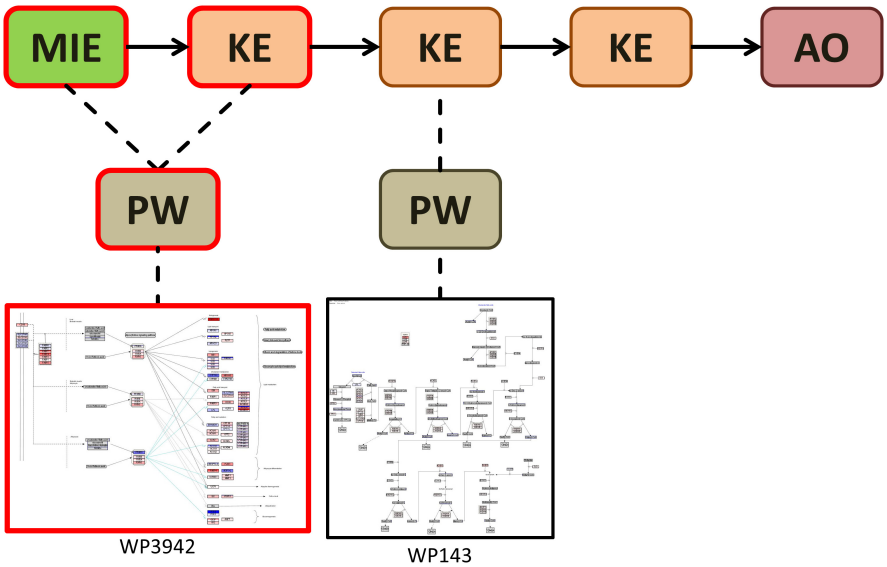


FIGURE 1 | Illustrative description of the linkage of KEs of an AOP with molecular pathways described in WikiPathways and the practical application of transcriptomics. Transcriptomics and pathway enrichment analysis are commonly used to elucidate molecular pathways affected after exposure to a chemical or stress signal. In this illustration, gene expression levels in WP3942 (Adriaens et al., 2018) are significantly changed (red and blue nodes in the pathway diagram, for up- and downregulation). Because this pathway is linked to the MIE and first KE, these are hypothetically affected by the chemical, highlighted with red borders and require validation. WP143 (Hanspers and Slenter, 2017) is not affected by the exposure of this chemical at the same time and dose, and the KE that is linked to this biological pathway is not considered to be affected but could follow later or at a higher dose. AO, adverse outcome; KE, key event; MIE, molecular initiating event; PW, pathway; WP, WikiPathways.

Textual Identifier Mapping for Genes and Proteins

In order to perform identifier mapping on the free-text descriptions of AOP-Wiki, we downloaded a human gene identifier dataset from the HUGO Gene Nomenclature Committee (HGNC) (HUGO Gene Nomenclature Committee [HGNC], 2018) in May 2018 via genenames.org, a curated online repository for HGNC-approved gene nomenclature, gene families and associated resources (Yates et al., 2017). A custom download was performed in which we requested HGNC IDs, approved symbols, approved names, previous symbols, synonyms, and Ensembl IDs. These identifiers were loaded in Python and used to filter the descriptions of KEs for genes, which are filtered for KEs on the molecular, cellular, tissue, and organ level of biological organization. Also, the KERs that connect these KEs were parsed and identifiers were mapped on their descriptions and texts on biological plausibility and empirical support.

Manual Matching of AOP-Wiki KEs to Molecular Pathways on WikiPathways

All AOP-Wiki KE IDs on the molecular, cellular, tissue, and organ level were extracted and their corresponding web pages were opened on aopwiki.org. From the KE titles and descriptive text, pathway names were selected and queried on wikipathways.org via the search-bar for molecular pathways. If results showed up for this initial search, the KE was considered present in WikiPathways. If the KEs did not contain a direct mention of a pathway, the genes and proteins were noted and were queried for their presence in pathways via the WikiPathways SPARQL endpoint. For KEs at the cellular level, at least the majority of the genes and proteins should be present in at least one pathway. However, for molecular KEs that describe only an interaction between two molecules, only the presence of the target molecule in WikiPathways was necessary to consider the KE covered by WikiPathways. This method was meant to give a rough overview of the overlap between the AOP-Wiki and WikiPathways databases. Because it does not include synonyms or ontological similarity, this overview is expected to underestimate the overlap.

RESULTS

For hard linkage of the two databases, meaning explicit identifier matching, we looked at the usage of ontology annotations of the AOP-Wiki and WikiPathways. For the AOP-Wiki we extracted ontology annotations from KEs on the molecular, cellular, tissue and organ level and identified which ontology sources were currently in use for biological processes, biological objects, cell-terms, and organ-terms. As shown in **Figure 2**, a large amount of KEs are not yet annotated with ontology tags. When looking more in detail, one can notice that biological processes are mostly described with Gene Ontology (GO) tags, especially at the molecular and cellular KEs whereas the biological objects are mostly annotated with tags from ChEBI and Protein Ontology (PR). Although AOP-Wiki

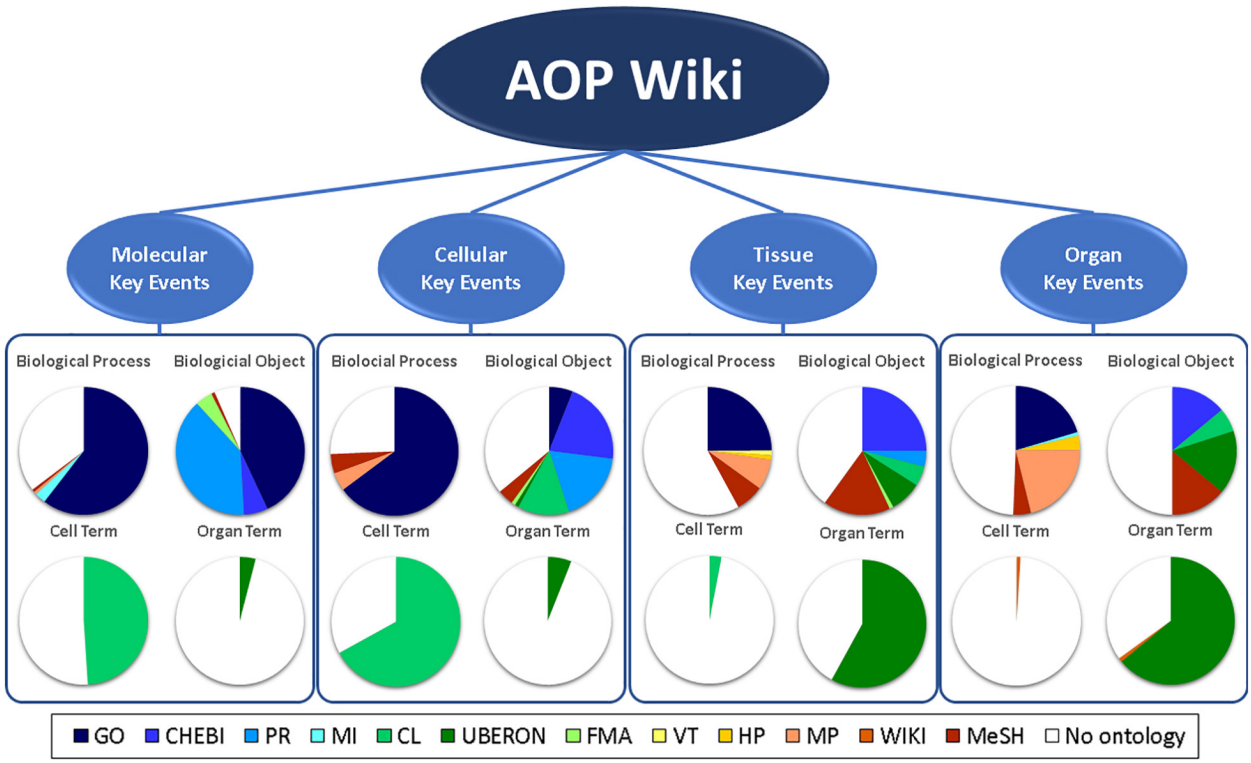


FIGURE 2 | Ontology usage of AOP-Wiki for KEs on the molecular, cellular, tissue, and organ level of biological organization. GO, gene Ontology; CHEBI, chemical entities of biological interest; PRO, protein ontology; MI, molecular interactions; CL, cell ontology; UBERON, uber anatomy ontology; FMA, foundational model of anatomy; VT, vertebrate trait; HP, human phenotype ontology; MP, mammalian phenotype ontology; WIKI, AOP-Wiki; MeSH, medical subject headings.

contains various ontology sources, WikiPathways only uses three: Pathway Ontology (PW), Cell Ontology (CL), and the Disease Ontology (DO) (**Figure 3**). However, apart from the CL for a contextual description of the process, WikiPathways and AOP-Wiki do not share ontologies for other biological elements.

Although no direct mappings through ontologies are possible at the moment of writing this paper, an alternative approach for hard linkage is the mapping of chemicals, metabolites, and genes to WikiPathways. Although we do not expect to find many of the AOP-Wiki stressor chemicals in WikiPathways, we wanted to identify the existing overlap of chemicals between the two databases nevertheless. First, we found all 306 stressors, describing 207 chemicals, which were annotated with 205 CAS numbers. We mapped these CAS numbers to ChEBI IDs in R with BridgeDbR and created a SPARQL query to find all pathways that have any of the metabolites included. This resulted in a total 194 out of 205 CAS numbers mapped to 298 ChEBI IDs, of which 48 mapped to a total of 133 WikiPathways.

As opposed to the hard linkage of the two databases, we also investigated a soft linkage, which entails the indirect linking of these databases through a text-based identifier mapping approach of human genes and performed a similar SPARQL query as for the metabolites (**Figure 4**). After extracting all KE descriptions from the AOP-Wiki, we mapped gene identifiers, symbols, alternative names, and previous names from HGNC to each description, leading to the identification of 523 genes in a total of 234 KE descriptions out of 787 KEs. In total, 70% of these genes were found in the molecular pathways of WikiPathways. Also, identifier mapping was performed on all 874 KERs that connect the KEs on the molecular, cellular, tissue and organ level. This was done on all texts for KER descriptions, biological plausibility, and empirical support, when available, and resulted in the identification of 417 genes, of which 296 are present in pathways on WikiPathways, which is 71%.

Furthermore, to benchmark the hard and soft connections between the AOP-Wiki and WikiPathways through ontologies and identifiers, we performed a full-scale manual check for all KEs on the molecular, cellular, tissue, and organ level of biological organization. This showed us that at least 2/3rd of all KEs can be mapped to molecular pathways on WikiPathways.

DISCUSSION

In this paper we explored possibilities for the integration of WikiPathways in the AOP-KB through ontologies, identifiers and manual judgment, to support AOPs and become a valuable tool in regulatory risk assessment. We looked at hard and soft linkages between the AOP-Wiki, the most actively used AOP module of the AOP-KB, and WikiPathways. We did this by extracting different types of information from the AOP-Wiki,

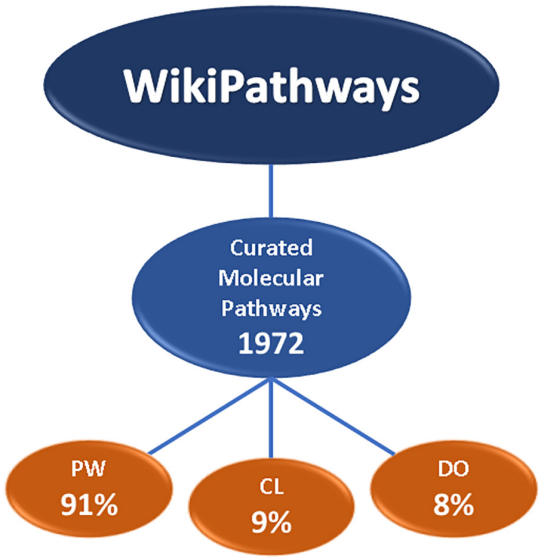


FIGURE 3 | WikiPathways statistics. The total number of molecular pathways in the WikiPathways database, and the level of pathway annotations with onology tags. PW, pathway ontology; CL, cell ontology; DO, disease ontology.

such as chemical CAS numbers, KE and KER descriptions, and ontology annotations, and we performed a manual judgment of the linkage.

We found that the AOP-Wiki uses various ontologies to describe the different elements of KEs. To link the underlying molecular pathways to these KEs, we are mainly interested in the Biological Process that is annotated in the KEs, which describe the biology of the KEs. However, the ontologies currently used in the AOP-Wiki do not directly connect with the ontologies that describe the molecular pathways of WikiPathways. Consequently, manual effort is currently required to make this mapping, which negatively impacts the scalability.

Furthermore, we focused on the metabolites and genes/proteins described on the AOP-Wiki. For the metabolites, we parsed all CAS numbers, mapped these to ChEBI identifiers, and found that only 16% of these are found in WikiPathways. This is not unexpected, because most toxicological effects are caused by exogenous compounds, whereas WikiPathways mostly stores biological pathways containing endogenous metabolites. In fact, most WikiPathways that contain such a stressor do so because the pathway described the biotransformation of the toxic compound.

On the other hand, gene/protein identifiers that we obtained through mapping with an HGNC dataset did show high coverage by WikiPathways (70%). However, with the gene/protein identifier mapping, we only focused on human variants, although KE descriptions on the AOP-Wiki cover a variety of species. The taxonomic information is absent in most KEs and if it is available, the taxonomy identifiers are inconsistent, so we were not able to take this into account in our experiment of identifier mapping. Although species specification with ontologies does exist on the

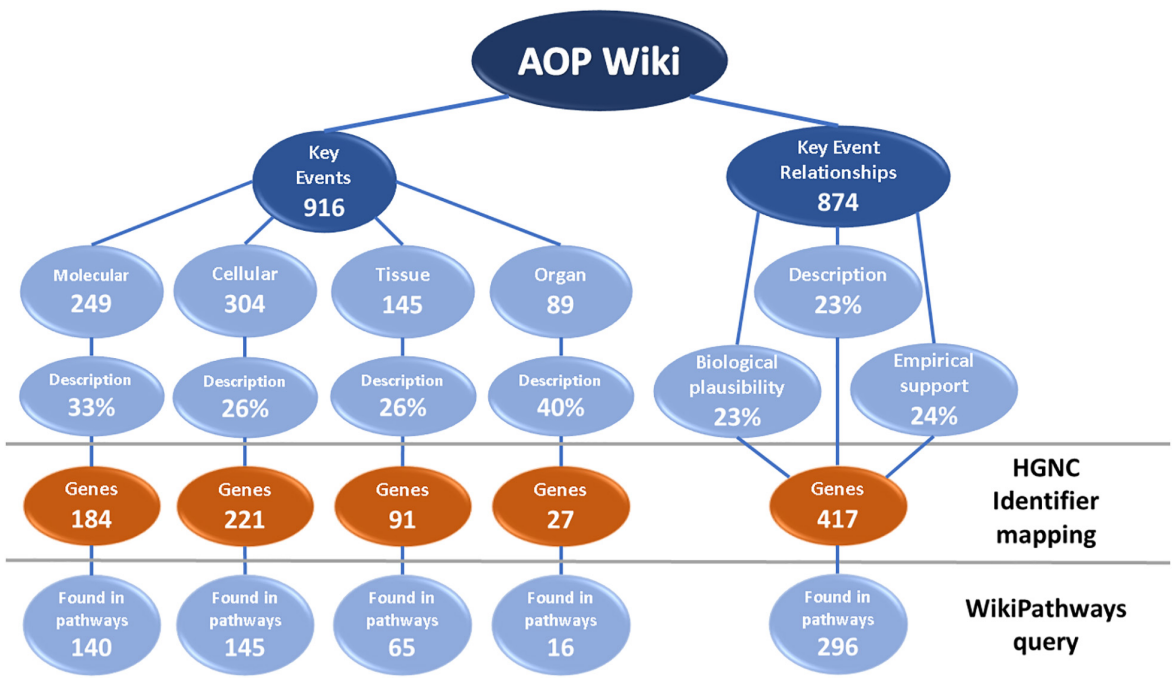


FIGURE 4 | AOP-Wiki statistics on KEs and KERs, identifier mapping with HGNC identifiers and links to molecular pathways in WikiPathways. The KEs on the molecular, cellular, tissue, and organ level of biological organization and the KERs that connect them were parsed for texts of descriptions, on the biological plausibility and on the empirical support. HGNC Identifier mapping was performed to find all human genes described in the key event descriptions, after which these genes were queried on WikiPathways to find pathways that contain these genes.

AOP-Wiki, the number of annotations and the consistency in reporting should increase for it to become a useful piece of data.

Apart from the automated linkages, we performed a manual check, which indicated that the majority of the processes in the AOP-Wiki KEs are covered by the WikiPathways database, either completely, as a part of a pathway or, in case of molecular interactions, the target molecule is part of a molecular pathway. This indicates us that there is potential in the interoperability of AOP-Wiki and WikiPathways to describe KEs. However, there is no one-to-one mapping of biological pathways possible. For example, molecular-level KEs currently often describe a single interaction between a list of stressors and a molecule, which would only be a part of a biological pathway on WikiPathways, besides the downstream cascade of molecular effects. Also, KEs on the tissue- and organ-level of biological organization are often non-specific. This could lead to the mapping of multiple molecular pathways to a single AOP-Wiki KE, even with the current WikiPathways content.

Besides the identification of connections between the AOP-Wiki and WikiPathways for improved descriptions of KEs, we aim for the possibility to introduce omics data analysis in the concept of AOPs. However, one concern mentioned in literature in the implementation of transcriptomics data in the concept of AOPs is the difference in the *causal* and *reactive* pathways (Sturla et al., 2014). Transcriptomics studies, for example, do not differentiate in its measurements between these two types of pathways, and by focusing on gene expression fold changes, pathway enrichment may highlight the reactive pathways. However, KEs may describe a causal event or pathway. Therefore, AOP-Wiki KE descriptions would not necessarily overlap with the results from pathway analysis with omics data. This should be taken into account in the descriptions of the molecular responses of KEs as this might impact the usability of omics approaches and their connections to KEs on the AOP-Wiki.

It is expected that omics approaches have great potential in the field of regulatory toxicology (Brockmeier et al., 2017; Buesen et al., 2017). However, there is a demand for well-described protocols and tools for omics data analysis and interpretation. The integration of WikiPathways in the AOP-KB as a data source and as omics data analysis tool allows more detailed descriptions of KEs and consistency in analysis and interpretation of omics data in the concept of AOPs. For that, you would ideally have molecular mechanistic descriptions for all AOP events in WikiPathways. The current analysis shows that useful connections already exist. To prepare for the integration of molecular pathways in the concept of AOPs, we created an AOP Portal on WikiPathways[7], in which all molecular pathways that are linked to AOP-Wiki KEs will be gathered and stored. This portal is meant to bridge the molecular knowledge and expertise of biologists and toxicologists to the framework of AOPs and allows the whole community to contribute to the collection of molecular pathways. This collection will be available for pathway analysis and network analysis with omics data for large-scale hypothesis generation for AOPs in response to a stressor or for biological read-across on the AOP level (Brockmeier et al., 2017). That would allow a more consistent, standardized approach for the integration of omics approaches in AOPs, and thus for regulatory use.

A variety of molecular pathway databases could fill this role as an omics analysis and interpretation tool for toxicological effects, such as KEGG and Reactome. However, molecular pathways can vary across pathway databases due to differences in pathway annotations by focusing on specific cellular contexts, such as diseases or specific cell types (Herwig et al., 2016). Moreover, Reactome and KEGG cannot be tailored like WikiPathways for specific communities or purposes such as described in this paper (Pico et al., 2008; Hanumappa et al., 2013). Besides, the accessibility of WikiPathways, being a community-driven, free-to-use molecular pathway database, fits with the existing AOP-KB modules and meets the requirements identified by the OECD: open access, standardized representation of data, and consistency in reporting (Pilat and Fukasaku, 2007; Ives et al., 2017). Because the AOP-KB is driven by a scientific community to develop, share and discuss AOPs, this community can also describe the molecular processes underlying the AOPs and contribute to WikiPathways and expand the AOP Portal.

Other work on the linkage of data related to the AOP-Wiki is the development of the AOP-DataBase (AOP-DB) (Pittman et al., 2018). This database will soon be publicly available and will contain various types of information linked to gene IDs that is useful for AOPs to provide a standardized, systematic structure for AOP development. Among a large amount of data, biological pathways from databases such as KEGG, Reactome, and ConsensusDB are included based on GO annotations of KEs in AOP-Wiki (Pittman et al., 2018). While the AOP-DB connects pathway databases based on the ontology annotations to of existing AOPs and assisting the identification of putative AOPs, we think that a direct link between KEs and molecular pathways would be valuable and more reliable.

In order to make a connection between AOP-Wiki and WikiPathways, we recommend a couple of improvements in terms of annotations and accessibility of the data. Since January 2018, the AOP-Wiki made available full XML files containing all data, which are stored as permanent downloads, as well as nightly exports of the full database. These files need to be parsed to retrieve the data, as described in this paper. This could be improved by developing an RDF version of the AOP-Wiki, allowing federated SPARQL queries to request all data, enable automatic information sharing, and has the use of ontologies as a core feature.

Furthermore, the current implementation of annotations with ontologies could be improved by annotating more specific elements of the KEs, as the existing KE components describe the KEs in general. More detailed annotations could be performed for many elements. For example, key genes, proteins, and metabolites should be annotated, as well as detection methods and biological assays, which can be annotated with ontologies such as the Chemical Methods Ontology or BioAssay Ontology. Also, when

[7]http://aop.wikipathways.org

biological pathways are described in a KE, annotations with the Pathway Ontology would allow a direct connection to the WikiPathways database including all genes, proteins, and metabolites involved, which are annotated with various databases through BridgeDb in the WikiPathways diagrams.

Besides the ontology annotations, the only molecules annotated on the AOP-Wiki are the chemicals related to stressors, which are identified with CAS numbers. However, not all of these CAS numbers are linked to open structure data that is incorporated in the BridgeDb mapping that we performed. It is essential that these CAS numbers are included in public databases, such as WikiData (Mietchen et al., 2015) or that public database identifiers are used, such as from ChEBI or even Wikidata as an outside database for chemical information. Besides chemicals, nanomaterials, which are extensively investigated for toxicity, also require annotations, for example with the eNanoMapper ontology (Hastings et al., 2015). Also, the free-text descriptions of KEs that describe the biological process can also be improved by more consistent reporting, such as a fixed vocabulary for all genes, proteins, and metabolites involved in the biological processes. For example, listing the most important molecules by HGNC symbols or ChEBI IDs for a KE would improve machine-readability and the automated discovery of new connections between KEs.

On the other hand, WikiPathways will also need to undergo updates to fit the connection as described, with a specific category of KE-related molecular pathways and the need for so-called meta-pathways to create an AOP Network. Also, the AOP Portal will be populated with pathways in a case-study approach, proving the usefulness of the database. Other improvements related to toxicity research is the linkage to kinetics databases, more info on post-translational modifications of proteins, and improved semantic annotations of localizations, for example, specific organelles, cells, or tissues.

Taken together, we claim that a tight integration of WikiPathways and AOP-KB will improve risk assessment because we can link omics data directly to KEs and therefore AOPs. However, to make assessment reproducible and valid, major changes are needed.

DATA AVAILABILITY STATEMENT

The AOP-Wiki dataset analyzed for this study can be found in the AOP-Wiki (aopwiki.org/downloads). The WikiPathways data used for this study can be found in the WikiPathways database (data.wikipathways.org).

AUTHOR CONTRIBUTIONS

MM and EW designed the study and did the analyses. TV performed the manual matching of the databases. MM wrote the first draft of the manuscript. LB and HA wrote sections of the manuscript. PN, FA, EW, and MM contributed to pathways in the WikiPathways AOP Portal. All authors conceptualized the study, revised the drafts, gave the feedback, and approved the final manuscript.

FUNDING

This project has received funding from the European Union's Horizon 2020 (EU 2020) research and innovation program under grant agreement no. 681002 (EU-ToxRisk) and EINFRA-22-2016 program under grant agreement no. 731075 (OpenRiskNet). FA was supported by a Marie Skłodowska-Curie Individual European Fellowship (H2020-MSCA-IF-2014-EF-ST) from the European Commission for the project NANOTAM, No. 658592 and by the Worldwide Cancer Research, United Kingdom. PN and RG are supported by the EU H2020 projects No. 646221 (NanoReg2), No. 686239 (caLIBRAte), and No. 760813 (PATROLS). The sole responsibility of this publication lies with the authors. The European Union is not responsible for any use that may be made of the information contained therein.

REFERENCES

Adriaens, M. E., Willighagen, E. L., Ahles, P., Pico, A. R., Jagers, F., Slenter, D. N., et al. (2018). *Fatty Acid Beta Oxidation (Homo sapiens)*. Available at: wikipathways.org/instance/WP143_r98914

Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., et al. (2010). Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.* 29, 730–741. doi: 10.1002/etc.34

Azad, A. K. M., Lawen, A., and Keith, J. M. (2017). Bayesian model of signal rewiring reveals mechanisms of gene dysregulation in acquired drug resistance in breast cancer. *PLoS One* 12:e0173331. doi: 10.1371/journal.pone.0173331

Bard, J. B. L., and Rhee, S. Y. (2004). Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* 5, 213–222. doi: 10.1038/nrg1295

Bell, S. M., Angrish, M. M., Wood, C. E., and Edwards, S. W. (2016). Integrating publicly available data to generate computationally predicted adverse outcome pathways for fatty liver. *Toxicol. Sci.* 150, 510–520. doi: 10.1093/toxsci/kfw017

Brockmeier, E. K., Hodges, G., Hutchinson, T. H., Butler, E., Hecker, M., Tollefsen, K. E., et al. (2017). The role of omics in the application of adverse outcome pathways for chemical risk assessment. *Toxicol. Sci.* 158, 252–262. doi: 10.1093/toxsci/kfx097

Buesen, R., Chorley, B. N., da Silva Lima, B., Daston, G., Deferme, L., Ebbels, T., et al. (2017). Applying 'omics technologies in chemicals risk assessment: report of an ECETOC workshop. *Regul. Toxicol. Pharmacol.* 91, S3–S13. doi: 10.1016/j.yrtph.2017.09.002

Burgoon, L. D. (2017). The AOPOntology: a semantic artificial intelligence tool for predictive toxicology. *Appl. Vitr. Toxicol.* 3:2017.0012. doi: 10.1089/aivt.2017.0012

Campos, B., and Colbourne, J. K. (2018). How omics technologies can enhance chemical safety regulation: perspectives from academia. *Gov. Ind.*. 37, 1252–1259. doi: 10.1002/etc.4079

Edwards, S. W., and Preston, R. J. (2008). Systems biology and mode of action based risk assessment. *Toxicol. Sci.* 106, 312–318. doi: 10.1093/toxsci/kfn190

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46, D481–D487. doi: 10.1093/nar/gkx1132

Gant, T. W., Sauer, U. G., Zhang, S. D., Chorley, B. N., Hackermüller, J., Perdichizzi, S., et al. (2017). A generic transcriptomics reporting framework

(TRF) for 'omics data processing and analysis. *Regul. Toxicol. Pharmacol.* 91, S36–S45. doi: 10.1016/j.yrtph.2017.11.001

Grafström, R. C., Nymark, P., Hongisto, V., Spjuth, O., Ceder, R., Willighagen, E., et al. (2015). Toward the replacement of animal experiments through the bioinformatics-driven analysis of "omics" data from human cell cultures. *Altern. Lab. Anim.* 43, 325–332.

Hanspers, K., and Slenter, D. N. (2017). *PPAR Signaling Pathway (Homo Sapiens)*. Available at: wikipathways.org/instance/WP3942_r94205

Hanumappa, M., Preece, J., Elser, J., Nemeth, D., Bono, G., Wu, K., et al. (2013). WikiPathways for plants: a community pathway curation portal and a case study in rice and arabidopsis seed development networks. *Rice* 6:14. doi: 10.1186/1939-8433-6-14

Hartung, T. (2016). Making big sense from big data in toxicology by read-across. *ALTEX* 33, 83–93. doi: 10.14573/altex.1603091

Hastings, J., Jeliazkova, N., Owen, G., Tsiliki, G., Munteanu, C. R., Steinbeck, C., et al. (2015). eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. *J. Biomed. Semantics* 6:10. doi: 10.1186/s13326-015-0005-5

Herwig, R., Hardt, C., Lienhard, M., and Kamburov, A. (2016). Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat. Protoc.* 11, 1889–1907. doi: 10.1038/nprot.2016.117

HUGO Gene Nomenclature Committee [HGNC] (2018). *HGNC Database, HUGO Gene Nomenclature Committee (HGNC), European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus*. Hinxton: HUGO Gene Nomenclature Committee.

Ives, C., Campia, I., Wang, R.-L., Wittwehr, C., and Edwards, S. (2017). Creating a structured adverse outcome pathway knowledgebase via ontology-based annotations. *Appl. Vitr. Toxicol.* 3, 298–311. doi: 10.1089/aivt.2017.0017

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kutmon, M., Lotia, S., Evelo, C. T., and Pico, A. R. (2014). WikiPathways app for cytoscape: making biological pathways amenable to network analysis and visualization. *F1000Res.* 3:152. doi: 10.12688/f1000research.4254.2

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., et al. (2016). WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.* 44, D488–D494. doi: 10.1093/nar/gkv1024

Kutmon, M., van Iersel, M. P., Bohler, A., Kelder, T., Nunes, N., Pico, A. R., et al. (2015). PathVisio 3: an extendable pathway analysis toolbox. *PLoS Comput. Biol.* 11:e1004085. doi: 10.1371/journal.pcbi.1004085

Leemans, C., Willighagen, E., Bohler, A., and Eijssen, L. (2018). *BridgeDbR: Code for Using BridgeDb Identifier Mapping Framework From Within R*. Available at: https://github.com/bridgedb/BridgeDbR

Leist, M., Ghallab, A., Graepel, R., Marchan, R., Hassan, R., Bennekou, S. H., et al. (2017). Adverse outcome pathways: opportunities, limitations and open questions. *Arch. Toxicol.* 91, 3477–3505. doi: 10.1007/s00204-017-2045-3

Martens, M. (2018). *marvinm2/AOPWikiXMLparsing: Version 1.0*. Geneva: ZENODO. doi: 10.5281/ZENODO.1306408

Mietchen, D., Hagedorn, G., Willighagen, E., Rico, M., Gómez-Pérez, A., Aibar, E., et al. (2015). Enabling open science: Wikidata for research (Wiki4R). *Res. Ideas Outcomes* 1:e7573. doi: 10.3897/rio.1.e7573

Nymark, P., Rieswijk, L., Ehrhart, F., Jeliazkova, N., Tsiliki, G., Sarimveis, H., et al. (2018). A data fusion pipeline for generating and enriching adverse outcome pathway descriptions. *Toxicol. Sci.* 162, 264–275. doi: 10.1093/toxsci/kfx252

Organisation for Economic Co-operation and Development [OECD] (2017). *Organisation for Economic Co-operation and Development: Revised Guidance Document on Developing and Assessing Adverse Outcome Pathways*. Paris: Organisation for Economic Co-operation, and Development.

Pico, A. R., Kelder, T., Van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008). WikiPathways: pathway editing for the people. *PLoS Biol.* 6:e184. doi: 10.1371/journal.pbio.0060184

Pilat, D., and Fukasaku, Y. (2007). OECD principles and guidelines for access to research data from public funding. *Data Sci. J.* 6, OD4–OD11. doi: 10.2481/dsj.6.OD4

Pittman, M. E., Edwards, S. W., Ives, C., and Mortensen, H. M. (2018). AOP-DB: a database resource for the exploration of adverse outcome pathways through integrated association networks. *Toxicol. Appl. Pharmacol.* 343, 71–83. doi: 10.1016/J.TAAP.2018.02.006

Python Software Foundation (2010). *Python Language Reference, version 2.7*. Available at http://www.python.org

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

R Studio Team (2015). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc.

Sauer, U. G., Deferme, L., Gribaldo, L., Hackermüller, J., Tralau, T., van Ravenzwaay, B., et al. (2017). The challenge of the application of 'omics technologies in chemicals risk assessment: background and outlook. *Regul. Toxicol. Pharmacol.* 91, S14–S26. doi: 10.1016/j.yrtph.2017.09.020

Slenter, D. (2018). *Metabolite BridgeDb ID Mapping Database (20180508)*. Available at: https://github.com/egonw/create-bridgedb-hmdb/issues

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., et al. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* 46, D661–D667. doi: 10.1093/nar/gkx1064

Sturla, S. J., Boobis, A. R., Fitzgerald, R. E., Hoeng, J., Kavlock, R. J., Schirmer, K., et al. (2014). Systems toxicology: from basic research to risk assessment. *Chem. Res. Toxicol.* 27, 314–329. doi: 10.1021/tx400410s

Vachon, J., Campagna, C., Rodriguez, M. J., Sirard, M. A., and Levallois, P. (2017). Barriers to the use of toxicogenomics data in human health risk assessment: a survey of Canadian risk assessors. *Regul. Toxicol. Pharmacol.* 85, 119–123. doi: 10.1016/j.yrtph.2017.01.008

Van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., et al. (2010). The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11:5. doi: 10.1186/1471-2105-11-5

van Ravenzwaay, B., Sauer, U. G., and de Matos, O. (2017). Editorial: applying 'omics technologies in chemicals risk assessment. *Regul. Toxicol. Pharmacol.* 91, S1–S2. doi: 10.1016/j.yrtph.2017.11.017

Villeneuve, D. L., Crump, D., Garcia-Reyero, N., Hecker, M., Hutchinson, T. H., LaLone, C. A., et al. (2014). Adverse outcome pathway (AOP) development I: strategies and principles. *Toxicol. Sci.* 142, 312–320. doi: 10.1093/toxsci/kfu199

Vinken, M. (2013). The adverse outcome pathway concept: a pragmatic tool in toxicology. *Toxicology* 312, 158–165. doi: 10.1016/j.tox.2013.08.011

Vinken, M., Knapen, D., Vergauwen, L., Hengstler, J. G., Angrish, M., and Whelan, M. (2017). Adverse outcome pathways: a concise introduction for toxicologists. *Arch. Toxicol.* 91, 3697–3707. doi: 10.1007/s00204-017-2020-z

Waagmeester, A., Kutmon, M., Riutta, A., Miller, R., Willighagen, E. L., Evelo, C. T., et al. (2016). Using the semantic web for rapid integration of WikiPathways with other biological online data resources. *PLoS Comput. Biol.* 12:e1004989. doi: 10.1371/journal.pcbi.1004989

Watanabe, K. H., Aladjov, H., Bell, S. M., Burgoon, L., Cheng, W.-Y., Conolly, R., et al. (2018). "Big data integration and inference," in *Big Data Prediction Toxicology*, eds D. Neagu and A. Richarz (Cambridge: Royal Society of Chemistry).

Yates, B., Braschi, B., Gray, K. A., Seal, R. L., Tweedie, S., and Bruford, E. A. (2017). Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.* 45, D619–D625. doi: 10.1093/nar/gkw1033

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 May 2019
doi: 10.3389/fgene.2019.00387

Quality Control of Quantitative High Throughput Screening Data

*Keith R. Shockley[1], Shuva Gupta[2], Shawn F. Harris[3], Soumendra N. Lahiri[4] and Shyamal D. Peddada[5]**

[1] Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Durham, NC, United States, [2] Statistics Department, University of Pennsylvania, Philadelphia, PA, United States, [3] Social and Scientific Systems, Durham, NC, United States, [4] Department of Statistics, North Carolina State University, Raleigh, NC, United States, [5] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, United States

Quantitative high throughput screening (qHTS) experiments can generate 1000s of concentration-response profiles to screen compounds for potentially adverse effects. However, potency estimates for a single compound can vary considerably in study designs incorporating multiple concentration-response profiles for each compound. We introduce an automated quality control procedure based on analysis of variance (ANOVA) to identify and filter out compounds with multiple cluster response patterns and improve potency estimation in qHTS assays. Our approach, called **C**luster **A**nalysis by **S**ubgroups using **ANOVA** (*CASANOVA*), clusters compound-specific response patterns into statistically supported subgroups. Applying *CASANOVA* to 43 publicly available qHTS data sets, we found that only about 20% of compounds with response values outside of the noise band have single cluster responses. The error rates for incorrectly separating true clusters and incorrectly clumping disparate clusters were both less than 5% in extensive simulation studies. Simulation studies also showed that the bias and variance of concentration at half-maximal response (AC_{50}) estimates were usually within 10-fold when using a weighted average approach for potency estimation. In short, *CASANOVA* effectively sorts out compounds with "inconsistent" response patterns and produces trustworthy AC_{50} values.

Keywords: ANOVA, clustering, concentration-response, potency, quantitative high throughput screening, toxicological response

OPEN ACCESS

Edited by:
Danyel Jennen,
Maastricht University, Netherlands

Reviewed by:
Matthew Thomas Martin,
Pfizer, United States
Katie Paul Friedman,
National Center for Computational Toxicology (NCCT), United States

****Correspondence:***
Shyamal D. Peddada
sdp47@pitt.edu

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Genetics

Received: *24 May 2018*
Accepted: *10 April 2019*
Published: *09 May 2019*

Citation:
Shockley KR, Gupta S, Harris SF, Lahiri SN and Peddada SD (2019) Quality Control of Quantitative High Throughput Screening Data. Front. Genet. 10:387. doi: 10.3389/fgene.2019.00387

INTRODUCTION

In 1978 the National Toxicology Program (NTP) was established to evaluate the toxicity and carcinogenicity of environmental chemicals. As part of these efforts, the NTP developed a 2-year rodent cancer bioassay to identify potential human carcinogens. After about 40 years conducting such studies, the NTP has conducted evaluations for about 600 chemicals. However, over 80,000 compounds are registered for use in the United States, and that number is increasing by an estimated 2,000 new chemicals each year (U.S. National Toxicology Program [U.S. NTP], 2017). A large number of these chemicals have unknown effects on human health. Therefore, during the previous decade the NTP and other agencies, including the U.S. Environmental Protection Agency (EPA), the National Center for Advancing Translational Sciences (NCATS), and the U.S. Food and Drug Administration (FDA), established quantitative high throughput screening (qHTS) assays simultaneously screen 1000s of compounds and prioritize chemicals for further testing (Tice et al., 2013). The goal of these qHTS assays was not only to achieve the speed of evaluating

1000s of chemicals in a single experiment, but also to substantially reduce the costs of toxicity testing and, eventually, to transform toxicology into a more predictive science (Collins et al., 2008).

Quantitative high throughput screening of 1000s of different compounds at multiple concentrations represents a marked technological advancement that minimizes the frequency of false negative calls compared to single concentration HTS (Inglese et al., 2006). Data generated from qHTS have a prominent role in toxicological assessment and drug discovery (Collins et al., 2008; Roy et al., 2010; Attene-Ramos et al., 2013; Dahlin et al., 2015). For instance, concentration-response data is currently being generated and made publicly available for 100s of toxicologically relevant endpoints in phase II of the Tox21 collaboration among the EPA, NCATS, the FDA and the NTP (Tice et al., 2013). Outcomes from these qHTS experiments can be used for numerous applications, including phenotypic screening (Kleinstreuer et al., 2014), genome-wide association mapping (Abdo et al., 2015) and prediction modeling (Eduati et al., 2015).

A qHTS assay produces one or more concentration-response curves for each tested compound. Here, we refer to a single concentration-response profile as a "repeat" (see section "Materials and Methods"). Each curve is typically evaluated using non-linear regression models. For example, the sigmoidal Hill model (Hill, 1910) is used to estimate the concentration at half-maximal response (AC_{50}), a quantitative measure of chemical potency. Heteroscedastic responses and outliers should be taken into account using robust statistical modeling, such as the preliminary test estimation based methodology proposed by Lim et al. (2013). In addition to other characteristics of the concentration response curve, potency measures are important to determine how toxic or active a chemical is in the assay system. Estimates of compound potency or other response characteristics are extremely important for assessing toxicity in toxicology assessment or bioactivity in drug discovery applications. Recently, there has been considerable controversy in comparing two large-scale qHTS studies (Barretina et al., 2012; Garnett et al., 2012). Haibe-Kains et al. (2013) reported that the drug response data in these two studies were inconsistent with each other based on poor concordance of IC_{50} and area under the curve (*AUC*) measures. This report and an accompanying commentary (Weinstein and Lorenzi, 2013) suggested that differences in laboratory protocols might account for this discordance and raised important questions about the validity and interpretation of current and future qHTS efforts. A number of studies have subsequently investigated the consistency of phamacogenomic drug response and investigated whether analytical assessments of consistency should take into account experimental features such as cell line (Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium, 2015; Geeleher et al., 2016; Haverty et al., 2016; Safikhani et al., 2016a,d) and viability (Bouhaddou et al., 2016; Safikhani et al., 2016c), and suggested standardized assay methods and laboratory conditions (Mpindi et al., 2016; Safikhani et al., 2016b). Accounting for experimental factors during statistical analysis may help to improve the reliability and reproducibility of qHTS results (Ding et al., 2017). Nevertheless, such modeling approaches may require a prohibitively large number of repeated profiles for each chemical, and many experimental factors remain unknown or confounded in qHTS experiments.

Unfortunately, no systematic quality control (Q/C) procedure has yet been established for qHTS data. We believe that the lack of such a Q/C procedure may contribute to the ongoing debate surrounding the consistency of large-scale *in vitro* screening data. In this paper, we take a simple and principled Q/C approach to sort out chemicals with "inconsistent" response patterns so that the researcher may identify and avoid computing AC_{50} values for potentially troublesome chemicals. Conversely, data with "consistent" responses across repeated profiles would produce AC_{50} values that can be trusted and used for downstream analyses.

In the Tox21 initiative, multiple concentration-response curves are obtained for each compound tested in a qHTS study. However, this may not be the case with other qHTS studies, where only a single response curve is obtained for each tested compound. In some cases, the concentration-response patterns in Tox21 Phase II fall into a single cluster where response patterns are "similar" across all experimental repeats (e.g., **Figures 1A,B**, based on data from an estrogen receptor agonist assay). Concentration-response curves corresponding to oxymetholone in **Figure 1A** appear to be in a single cluster with all repeats exhibiting monotonic responses except at the highest concentration tested. Each curve crosses the upper noise bound (horizontal dashed line), suggesting that this compound is a candidate hit that may activate the estrogen receptor. Similarly, concentration-response data corresponding to hydrochlorothiazide in **Figure 1B** comprise one cluster pattern across all repeats since every concentration curve is within the noise limits, indicating that this chemical may not be active under the tested conditions. In examples such as **Figure 1A**, where all response curves are part of a single cluster, a Hill model (Hill, 1910; Shockley, 2015) or other appropriate non-linear model can be fit to the data in order to obtain potency estimates that summarize each curve. These individual potency estimates can then be used to obtain an overall potency estimate for the compound. Since the compound in **Figure 1B** appears to be inactive under the tested conditions, no potency estimate is obtained for this compound.

In the absence of systematic effects and artifacts, concentration-response curves for each chemical should be "similar" or within a single cluster across all experimental repeats of the compound (Hsieh et al., 2015). However, in **Figure 1C** the concentration-response patterns for 2,3,5,6-tetrachloronitrobenzene are split into four different clusters (indicated by different colors) across the experimental repeats. The AC_{50} values for the three clusters with response values extending outside of the noise band range from 3.93×10^{-10} to 19.57 μM, representing a wide variance in potency associated with this compound. Unfortunately, the numerous examples of compounds with multiple response clusters could produce dramatically different potency estimates for the same compound. In such cases it can be very difficult to ascertain the correct concentration-response pattern for the tested compound, and its corresponding potency estimate, from the data alone.

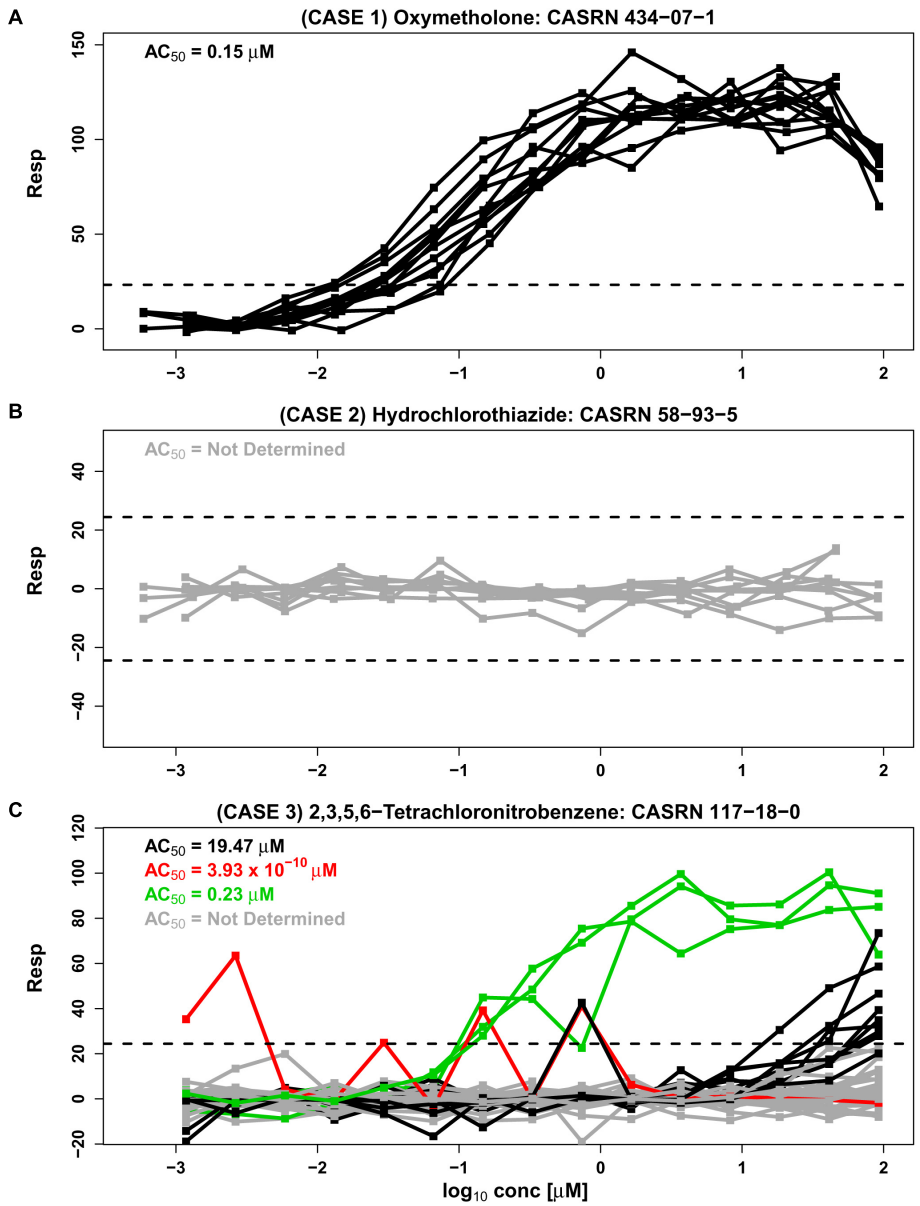


FIGURE 1 | Three separate cases are represented by concentration-response data from the BG1 estrogen receptor agonist assay from phase II of the Tox21 collaboration (*tox21-er-luc-bg1-4e2-agonist-p2*). Responses are shown as a percentage of the assay positive control values after correction by DMSO negative controls (Inglese et al., 2006). The assay detection limits are indicated with dashed lines. An AC_{50} value from the Hill model, calculated using the weighted average approach, summarizes the potency of each cluster (see section "Materials and Methods"). **(A)** Case 1 shows 12 similar response profiles from oxymetholone which extend beyond noise and group together into a single cluster. This case corresponds to two different supplier designations, two library preparation sites and two purities (A and D, representing "good" and "poor" purity, respectively) generated on six different experimental days. **(B)** Case 2 shows nine responses from hydrochlorothiazide which all lie within the noise band and correspond to three supplier sources, three library preparation sites, and a single purity (A) generated in six different experimental days. **(C)** Case 3 is represented by 42 response profiles from 2,3,5,6-tetrachloronitrobenzene corresponding to one supplier, three library preparation sites, one purity designation (A) and seven experimental days. A total of 29 of the 42 repeats lie within the noise band (shown in gray), and other profiles cluster by our proposed methodology *CASANOVA* described in this paper into the three disparate groups of 9, 3, and 1 repeats shown in black, green, and red, respectively. The separation of clusters in Case 3 is not explained by library preparation site or experimental day.

Chemical supplier, institutional site preparing the chemical library (e.g., NTP, FDA, and EPA), concentration-spacing, purity of the compound and other factors can systematically influence response trajectories (Tice et al., 2013). Such experimental factors are associated with different clusters in some instances. However, known design characteristics are not always associated with the observed response groupings.

An important purpose of qHTS assays is to estimate the potencies of active compounds for downstream analyses. In many cases, AC_{50} values and point of departure values estimated

from qHTS assays are used to discriminate between active and inactive compounds. Published studies incorporate AC_{50} potency estimates derived from qHTS assays for predictive cheminformatics (Jamal et al., 2016), *in vivo* activity prediction modeling (Martin et al., 2011; Kleinstreuer et al., 2013; Anthony Tony Cox et al., 2016), screening for therapeutic leads (Martinez et al., 2016; Xu et al., 2016; Chen et al., 2017), drug sensitivity testing (Barretina et al., 2012; Garnett et al., 2012), *in vitro*-to-*in vivo* extrapolation (IVIVE) pharmacokinetic modeling (Rotroff et al., 2010; Wetmore et al., 2012), computational modeling of androgen receptor activity (Kleinstreuer et al., 2017), toxicity testing (Judson et al., 2016; Karmaus et al., 2016) and prioritization for targeted testing (Judson et al., 2010). It is crucial to identify and distinguish compounds that have single cluster response patterns across repeated runs from compounds with multiple cluster response patterns. Otherwise, the potency estimates derived from qHTS assays may not be reliable, as seen for 2,3,5,6-tetrachloronitrobenzene in **Figure 1C** where the potency estimates for different clusters are highly variable. Visual inspection of response profiles and manual curation of "flagged" compounds (Filer et al., 2017) are based on complex rule structures and do not address the quality control issue that is investigated here. Since 1000s of compounds are tested in each assay, there is a need for an automated quality control process to separate compounds with single cluster and multiple cluster response patterns before making activity calls and estimating the potency of biologically responsive agents. Here, we focus on the statistical identification of single cluster and multiple cluster compounds in a data driven framework, and do not address the separate problem of relating the data to pathways of interest (Hsieh et al., 2015).

MATERIALS AND METHODS

Development of the *CASANOVA* Clustering Algorithm

A typical qHTS assay in Tox21 generates concentration-response data multiple times for each compound. Rather than referring to these multiple observations on each compound across concentrations as "replicates" we refer to them as "repeats." In typical experimental designs "replication" refers to repeating the experiment several times under identical experimental conditions. This is not the case with qHTS studies. In qHTS, for a given compound the experiment is often repeated by varying suppliers, laboratories/agencies (sites) preparing the library, chemical purity, etc. In each instance a concentration curve is obtained and these concentrations curves cannot be viewed as conventional replicates.

We developed an automated clustering algorithm called *CASANOVA* to cluster intrachemical responses into single clusters using classical two-way analysis of variance (ANOVA). The workflow for *CASANOVA* is presented in **Supplementary Figure 1**. First, concentration-response repeats having all responses across the concentration range located entirely within the noise band are removed, where the noise band is defined as $\pm$ 3 standard deviations (σ) of the response at the lowest concentration tested in the experiment. qHTS studies typically base the assay detection limit on the variation in the DMSO negative controls (Hsieh et al., 2015), the DMSO controls and the lowest concentration (Huang et al., 2011), or the first two concentrations (Filer et al., 2017). Defining the detection limit based on just the DMSO negative controls could be problematic for antagonist assays in which the response at the lowest tested concentration relies on two different components: the DMSO controls and the agonist response needed to activate a nuclear hormone receptor. To be consistent across assay types and other studies in the literature, we chose to base the assay detection limit on the first tested concentration. In many, but not all, assays the variation in the DMSO negative control wells is very similar to the variation at the lowest tested concentration (**Supplementary Figure 2**).

Here, for each compound with at least two repeats extending beyond the assay detection limit of 3σ (or -3σ), an ANOVA model is fit to all n intrachemical response profiles. If all repeats within a compound lie within the noise band, the compound is designated "Case 2." A grouping factor to divide the concentration space is essential to our approach. In this study, we focus on the 15-point concentration response profiles generated in phase II of Tox21 and use five "3-concentration" bins to define a five-level "concentration" grouping factor termed *CONC*. We consider each concentration-response profile in the experiment to be a "repeat," and *REPEAT* is used as a second factor in the model. Response R_{ijk} for concentration bin i ($CONC_i$), repeat j ($REPEAT_j$), and an interaction term (γ_{ij}) for observation k is modeled using the compound-specific ANOVA model

$$R_{ijk} = \mu + CONC_i + REPEAT_j + \gamma_{ij} + \varepsilon_{ij} \quad (1)$$

where μ is the overall mean and ϵ_{ijk} represents random error for concentration bin i, repeat j and observation k. The γ term is first tested for statistical significance within each compound. If the interaction term is significant at the user specified level of α (H_0: $\gamma_{11} = \gamma_{12} = \ldots = \gamma_{nn}$), then the *REPEAT* term is tested for significance at the α level (H_0: $REPEAT_1 = \ldots = REPEAT_n$). Unless otherwise noted, we used $\alpha = 0.05$ for all analyses presented here. If *REPEAT* is also significant, then repeats are ranked by mean response averaged over all levels of the *CONC* factor and significant pairwise differences between neighboring repeats in the ranked list are used to group repeats into distinct clusters. Subgroup analysis then proceeds by ranking mean response values within the highest *CONC* bin. Significant pairwise differences between neighboring repeats in the ranked list within this bin are used to further divide these clusters into new subclusters. The subgroup analysis proceeds for each *CONC* bin level (from the highest concentration to lowest concentration). If γ is significant, but *REPEAT* is not significant, only the subgroup analysis is performed. If the γ term is not significant, but *REPEAT* is significant, repeats are ranked by mean response averaged over all levels of the *CONC* factor and significant pairwise differences between neighboring repeats in the ranked list are used to group repeats into distinct clusters.

Once the clusters of similar dose profiles have been determined, the mean response values lying above (or below)

the noise band across all concentration bins are compared with the upper (or lower) detection limit using the one sample *t*-test ($\alpha = 0.05$) in order to distinguish between "conclusive" clusters that are statistically separated from the noise band and "inconclusive" clusters that are not statistically different from the noise band detection limit. "Case 1" compounds are composed of *n* single cluster repeats, where *n* refers to all the tested repeats within a compound. "Case 3" compounds each contain multiple cluster response patterns, where one of the clusters can potentially be repeats with all responses located entirely within the noise band. **Supplementary Figure 3** describes the five different classes of possible compound classification outcomes.

Description of Tox21 Phase II Data Sets

Publicly available Tox21 Phase II data was obtained from https://tripod.nih.gov/tox/. This qHTS data involves approximately 10,000 compounds screened for activity related to stress response, nuclear hormone receptor activity, or cell viability. The nuclear receptor hormone assays were performed in agonist and antagonist (or inhibitor) modes and are used to investigate activation or inhibition activities of the given assay. Multiple channel readouts for beta-lactamase gene reporter assays consisted of ch1, ch2 and ratio (ch2/ch1) data, and in those cases we used the ratio data to represent the assay signal. A total of 15 concentrations were evaluated with concentrations typically ranging from approximately 5×10^{-4} μM to about 100 μM (Tice et al., 2013). As part of phase II of Tox21, the library is screened three times with compounds located in different well positions during each experimental run (Tice et al., 2013). The raw plate reads were normalized using the positive and negative control wells and subsequently corrected for row, column, and plate effects using linear interpolation (Inglese et al., 2006). A total of 43 of the 47 publicly available bioassay data sets represented by 72 different readouts from phase II of the Tox21 collaboration were selected for analysis in this study due to their comparable experimental design of 15-point concentration response data generated in triplicate runs. We dropped 4 of the 47 publicly available data sets from our analysis because their study design was not directly comparable with the other 43 data sets; 2 of the assays were conducted as 4- or 8-point concentration-response study designs and 2 additional assays were unreplicated time course experiments.

AC_{50} values, and corresponding standard errors (*SE*), of individual concentration-response curves were estimated from the data using the Hill model after removing outliers as described previously (Shockley, 2012). The AC_{50} from each cluster in a single compound was estimated with a weighted approach using $(1/SE)^2$ as weights and the *weighted.mean()* function in R.

Simulation Studies to Evaluate the *CASANOVA* Algorithm

The performance of *CASANOVA* to correctly cluster similar patterns and separate disjoint patterns, was evaluated in simulation studies conducted across a range of assay noise levels chosen to resemble the characteristics found in Tox21 Phase II qHTS data. A total of 2,000 simulated compounds with at least one response outside of the noise band were generated from either the Hill model (sigmoidal curves) or the gain-loss model ("bell-like" curves) (Shockley, 2016; Filer et al., 2017). The parameters of the simulation study were based on observed data in the Tox21 Phase II data sets. Of the 43 publicly available Tox21 data sets (with 72 readouts) examined here, we chose four assay readouts that span the range of assay noise based on negative control DMSO plates (see **Supplementary Figure 4**) and the lowest tested concentration levels (**Supplementary Figure 5**). These selected readouts come from assays with low noise (data set 1: *tox21-elg1-luc-agonist*), moderate-low noise (data set 2: *tox21-are-bla-p1*), moderate-high noise (data set 3: *tox21-er-luc-bg1-4e2-agonist-p2*), and high noise (data set 4: *tox21-fxr-bla-agonist-p2*). The proportion of chemicals with *N* suppliers ($N = 1, 2, 3, 4$ in the Tox21 Phase II experiments) in each of the selected data sets was calculated (see **Supplementary Table 1**) and used as input probabilities for simulating the number of clusters per compound. Similarly, the proportion of compounds with *N* repeats per supplier ($N = 3$, 6, 9, 12, 42, 45, 48, 51, 54) was determined empirically for the four selected datasets (see **Supplementary Table 2**) and used as input probabilities for simulating the number of repeats per cluster in each compound. An ANOVA model in Eq. (1) was fit to compounds containing at least two repeats with detectable responses as described above. For each chemical, the ANOVA mean squared error (*MSE*), the range defined by maximum observed response – minimum observed response (*ResponseRange*) and the coefficient of variation (*CV*) defined by $\sqrt{MSE}/ResponseRange$ was calculated. These values, presented in **Supplementary Table 3**, were used to similate the data as described in greater detail below.

Simulated concentration-response curves are randomly chosen for each cluster based on a three-parameter Hill equation model or a four-parameter "gain-loss" model. The three-parameter Hill model is described by:

$$E(R_{ij}) = \frac{RMAX_j}{1 + 10^{\{-h_j[log_{10}C_i - log_{10}AC_{50,j}]\}}} \quad (2)$$

where R_{ij} is a normalized response (% of positive control activity) for the *jth* repeat, $RMAX_j$ represents maximal response, h_j is the slope parameter, C_i is the compound concentration, and $AC_{50,j}$ is the concentration for half-maximal activity. Similar to a previous study (Shockley, 2016), the concentrations are based on equivalent $\log_{10}$ concentration spacing from 0.0001 to 100 μM in 15-point concentration-response curves. The "gain-loss" model is given by

$$E(R_{ij}) = RMAX_j \left(\frac{1}{1 + 10^{\{h_j(log_{10}AC_{50(G),j} - log_{10}C_i)\}}} \right) \times \left(\frac{1}{1 + 10^{\{h_j(log_{10}C_i - log_{10}AC_{50(L),j})\}}} \right) \quad (3)$$

where $RMAX_j$ is the shared upper asymptote, both bottom asymptotes are set to zero, h_j is the slope parameter, $AC_{50(G),j}$ is the concentration of half-maximal response in the gain direction and $AC_{50(L),j}$ is the concentration of half-maximal response in the loss direction (Filer et al., 2017).

For each cluster, the mean $RMAX_j$ value (μ_{RMAX}) is selected using random deviates from the uniform distribution on (3σ, *ResponseRange*) and $RMAX_j$ is drawn from $N(\mu_{RMAX}, MSE)$. The slope parameter h_j is drawn from $|N(1,9)|$. For each cluster, mean values (*MEAN*) of $log_{10}AC_{50,j}$ from the Hill model, or $log_{10}AC_{50(G),j}$ from the "gain-loss" model, are randomly selected from (0.0001, 0.001, 0.01, 0.1, 1, 10, 100), or from (0.0001, 0.01, 1, 100) with equal probabilities and without replacement, across clusters for 10-fold AC_{50} spacing and 100-fold AC_{50} spacing, respectively. Mean values of $log_{10}AC_{50(G),j}$ are randomly selected from (0.0001, 0.001, 0.01, 0.1, 1, 10, 100) with equal probabilities where $log_{10}AC_{50(L),j} - log_{10}AC_{50(G),j} \geq 1$ for 10-fold AC_{50} spacing, or from (0.0001, 0.01, 1, 100) with equal probabilities where $log_{10}AC_{50(L),j} - log_{10}AC_{50(G),j} \geq 2$ for 100-fold AC_{50} spacing. If no $log_{10}AC_{50(L),j}$ values within the selected range meet this criterion, $log_{10}AC_{50(L),j}$ is set to 1000. The random realization of the mean $log_{10}AC_{50,j}$ value, or $log_{10}AC_{50(G),j}$, is drawn from $N(MEAN, \sigma)$, where $\sigma = 1/6$ is selected so that ∼99.7% of all AC_{50} values between clusters are separated at least 10- or 100-fold, depending on the simulation scenario. After determining the parameters for each cluster, response data was simulated by adding heteroscedastic noise to ideal curves with $N(0, R_{ij} \times CV)$, where R_{ij} is given from Eq. (2) or Eq. (3) above. Summary statistics for the simulated data are given in **Supplementary Tables 4, 5**.

RESULTS

Applying *CASANOVA* to Tox21 Phase II Data

CASANOVA was applied to publicly available Tox21 Phase II data related to stress response, nuclear receptor signaling and cell viability in order to assess the consistency of intra-chemical response patterns within and between assays. We selected 43 of the 47 publicly available data sets since these data sets were generated using a similar experimental design (i.e., 15-point concentration-response data generated in three experimental runs). These 43 data sets correspond to 72 different readouts, where many of the agonist and antagonist assays monitored cytotoxicity as well as the response in the specified assay mode. A total of 7,229 chemicals were represented in all 72 readouts.

The barplot in **Figure 2** shows the fraction of these compounds that were classified as single clusters that are well-separated from the noise band (Conclusive Case 1), single clusters that extend outside of the noise band and points outside the noise threshold are not significantly different from the noise band (Inconclusive Case 1), non-responsive with all repeats located within the noise band (Case 2), multiple clusters where at least one cluster extends outside the noise band and points outside the noise threshold are not significantly different from the noise band (Inconclusive Case 3) or multiple clusters for which at least one cluster extends significantly beyond the noise band (Conclusive Case 3). Most chemicals do not exhibit any response in the tested assay conditions (Case 2). The fraction of single clusters among all 7,229 compounds with at least one detectable response in an assay ranges from 1.6% (*tox21-vdr-agonist-p1*) to 23.8% (*tox21-dt40-p1_100*) across the 72 readouts. As shown in the plots for selected compounds in **Figure 2**, this multiplicity in response is sometimes associated with one or more known experimental design factors such as supplier, library preparation site, compound purity, concentration spacing, or experimental day (**Figure 2**). For example, in the top panel of **Figure 2** supplier is confounded with site of library preparation so that one or both of these two experimental factors can potentially account for the separation of response patterns into two different clusters.

The Hill model (Hill, 1910) was used to estimate the concentration for half maximal activity (AC_{50}) for the 7,229 compounds common to all 43 data sets. Compounds with two or more clusters outside of the noise band and estimated AC_{50} values within about 10-fold of the typical concentration range in the assays (10^{-5} to 1,000 μM) were evaluated further in order to discover the variability in AC_{50} estimates within a multiple cluster compound. In **Figure 3A**, the percentage of multi-cluster compounds with AC_{50} estimates greater than 10-fold ranged from 16.7% for *the tox21-gh3-tre-agonist-p1* agonist assay to 65.6% for the *tox21-er-luc-bg1-4e2-antagonist-p1* viability assay. The percentage of multi-cluster compounds with AC_{50} estimate differences greater than 100-fold ranged between 10.7 and 43.8% for these two assays, respectively. The fraction of compounds with multiple cluster responses was not statistically different between agonist and antagonist/inhibitor assays. However, the distribution of multiple cluster compounds was greater in viability assays compared to the agonist and antagonist/inhibitor assays, when considering 10-fold (**Figure 3B**) or 100-fold (**Figure 3C**) potency differences ($p < 0.001$ using the two-sided Kolmogorov–Smirnov test). In **Figure 3B**, about 38% of the 7,729 tested compounds have at least a 10-fold spread in AC_{50} estimates in half of the agonist and antagonist/inhibitor assays, whereas about 54% of the tested compounds have at least a 10-fold spread in AC_{50} estimates in half of the viability assays. In **Figure 3C**, about 18% of the tested compounds have at least a 100-fold spread in AC_{50} estimates in half of the agonist and antagonist/inhibitor assays, while about 32% of the tested compounds have at least a 100-fold spread in AC_{50} estimates in half of the viability assays.

Simulation Studies to Evaluate the Performance of *CASANOVA*

Simulation error rates were determined by averaging error rates from each simulated data set of 2,000 compounds across 100 different simulated runs. Error rates were calculated for each run based on the proportion of compounds with a given error type. "Type A" error was assigned to a compound when the *CASANOVA* approach incorrectly separated any two repeats from a true cluster (i.e., when a true single cluster compound was classified as a Conclusive Case 3). Conversely, a "Type B" error was assigned to a compound when any two repeats from separate clusters were falsely combined (i.e., when a true multiple-cluster compound was classified as a Conclusive Case 1). In both cases, these error rates are less than 5% with $p < 0.05$ and $p < 0.10$ as

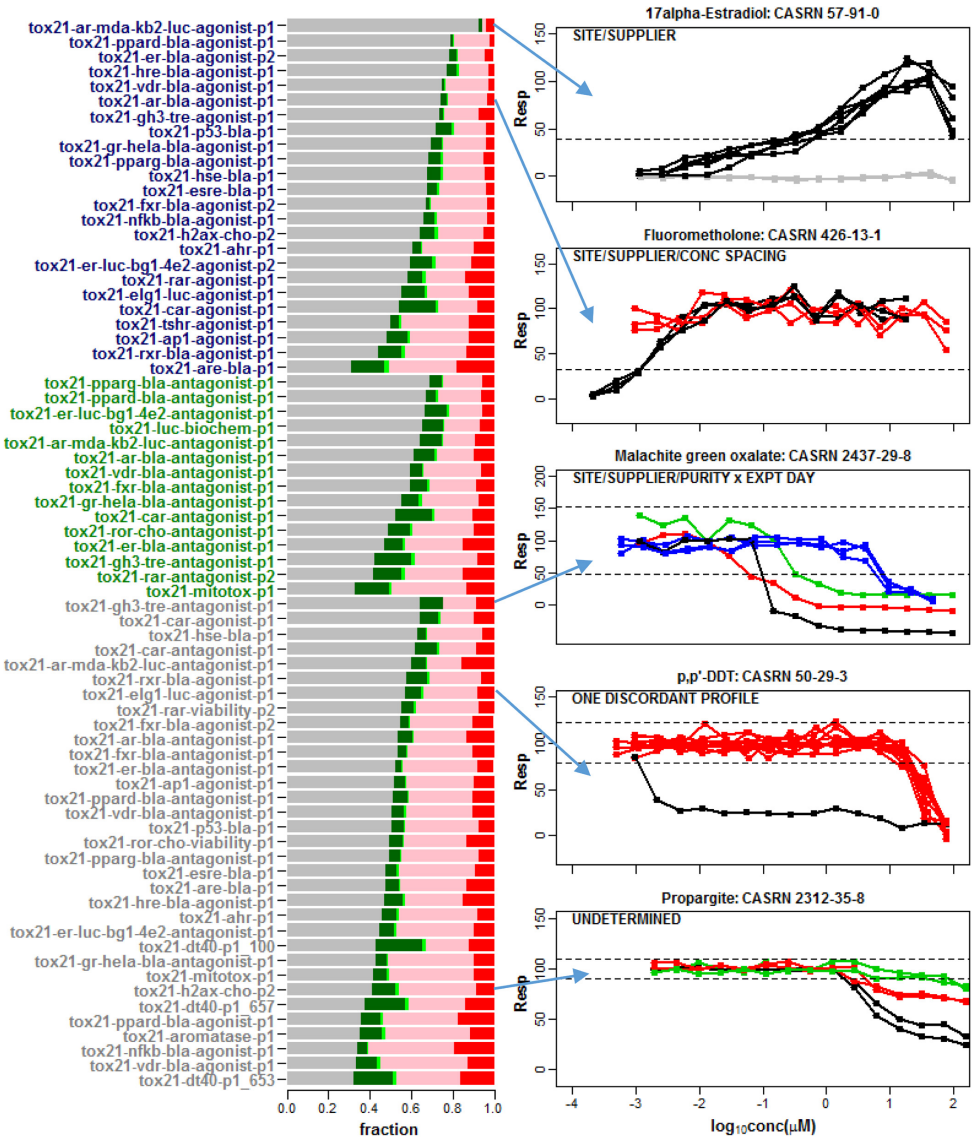


FIGURE 2 | A barplot was used to summarize the response patterns corresponding to 72 assay readouts from 43 different data sets. A total of 7,229 chemicals were common among all 43 data sets. In the barplot, the gray regions correspond to the fraction of chemicals clustered in the noise band (Case 2), the dark green regions refer to a single detectable cluster well-separated from the noise band (Conclusive Case 1), the light green regions represent a single cluster with response points not statistically separable from noise (Inconclusive Case 1), the pink regions correspond to multiple clusters with response points not statistically separable from the noise band (Inconclusive Case 3) and the red regions refer to multiple clusters well-separated from the noise band (Conclusive Case 3). Agonist assay labels are shown in dark blue, antagonist/inhibitor assay labels are shown in green and viability assay labels are shown in gray. Selected compound profiles from assays with multiple clusters (Conclusive Case 3) are shown to the right of the barplot. Known factors associated with different clusters are indicated in the upper left of each plot. These factors include supplier, library preparation site, concentration spacing, compound purity and experimental day. None of these factors explain the different patterns observed in the last two plots. Hence, adjusting or normalizing the concentration-response data for these known factors will not necessarily eliminate multiple cluster response patterns among repeats within a compound in qHTS data.

the selected criterion for identifying and separating clusters for either 10-fold AC_{50} spacing or 100-fold AC_{50} spacing (**Table 1**).

Simulation Studies to Evaluate AC_{50} Parameter Estimation

The bias and precision (1/variance) of AC_{50} estimation was evaluated in a separate simulation study. This simulation reflects the situation in which potency is estimated for single cluster compounds (Case 1). A total of 2,000 chemicals were simulated in activation mode, with increasing responses for increasing concentrations across a range of concentrations between 0.1 nM and 100 μM, where AC_{50} values were set to 0.001 μM (upper asymptote only), 0.1 μM (both asymptotes) or 10 μM (lower asymptote only). *RMAX* was considered at three values (25, 50, and 100% of positive

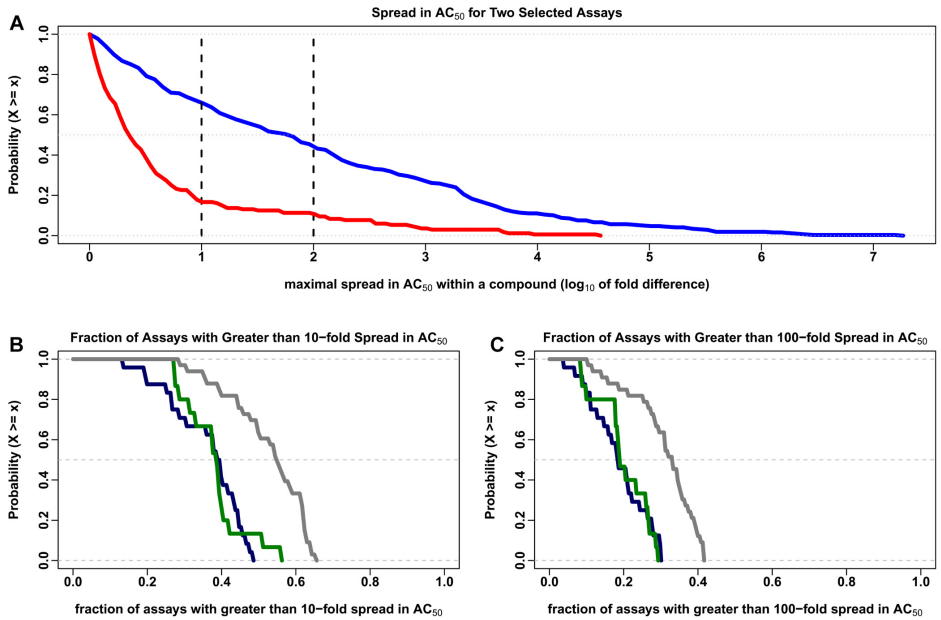


FIGURE 3 | Complementary empirical cumulative distribution (CCDF) describing the variability in AC_{50} values. The maximal range of AC_{50} values (on the $\log_{10}$ scale) was calculated for each compound in which two or more clusters were identified outside of the noise region for each of the 7,729 compounds investigated in the 43 data sets described in the text. The order of magnitude differences in intrachemical potency estimates shown here represent only those cases in which the calculated AC_{50} is between 10^{-5} and 1000 μM, which covers the typical testing concentration range of $\sim 10^{-4}$ to 100 μM evaluated in these assays. The number of compounds meeting this criterion ranged from 42 to 774 in the 72 assay types evaluated here, with a median of 255 compounds. **(A)** The CCDF (or 1-CDF) plots describing the proportion of compounds (y-axis) for a given spread in AC_{50} (x-axis) in the *tox21-er-luc-bg1-4e2-antagonist-p1* viability assay (blue) and the *tox21-gh3-tre-agonist-p1* agonist assay (red) are displayed. The vertical black lines indicate 10- and 100-fold differences in the calculated range of AC_{50} values. **(B)** The CCDF for the fraction of the 72 assays with greater than 10-fold range in AC_{50} values (y-axis) for a given spread in AC_{50} (x-axis) are shown for the agonist (dark blue), antagonist/inhibitor (dark green) and viability (dark gray) assays. **(C)** The CCDF for the fraction of the 72 assays with greater than 100-fold range in AC_{50} are shown for the same agonist, antagonist/inhibitor and viability assays presented in **(B)**.

control). The hill parameter was set to 1 for all curves in this simulation. Residual errors were modeled as $ERROR \sim N(0, \sigma^2)$ with σ = 5% or 10%.

Outliers were removed (Wang et al., 2010), separate curves were fit to each response curve and the $log_{10}AC_{50}$ parameter value was calculated for each profile (Shockley, 2012). We evaluated n profiles per compound for n = 3, 6, 9, or 12. For each compound, profile-specific estimates were summarized using the average, median or weighted average of the estimates, or a single model fit (Shockley, 2015). As described above, the weighted average approach uses $(1/SE)^2$ for weights, where SE is the standard error of the parameter estimate. The bias was less than 0.01 (1.02-fold) and the variance was less than 0.04 (1.1-fold) when both plateaus/asymptotes were present in the simulated sigmoidal curve for σ = 5% (**Table 2**). These errors were larger for σ = 10% (**Supplementary Table 6**). The weighted average approach produced the most repeatable results, where both bias and variance of the estimated $log_{10}AC_{50}$ for a compound were typically within one order of magnitude (10-fold).

DISCUSSION

Millions of dollars are being invested in developing qHTS assays and there are far reaching economic and public health implications for these large-scale studies. We believe that there is a pressing need for a rigorous, yet simple, Q/C process such as the one we offer in this work. Chemical genomics efforts inevitably involve multiple sources of variation imposed by limited resources and the technological constraints of robotic plate handling (Attene-Ramos et al., 2013). On the one hand, it can be advantageous to have compound activity data generated across multiple design factors in order to increase the chances that an observed response is related to the biological assay of interest rather than technical error (Ding et al., 2017). However, differences in chemical supplier, compound purity, laboratory protocol, or the day of the experiment may produce systematic errors that vary from chemical to chemical. Assay interference arising from autofluorescence and compound-induced cytotoxicity can also cause misleading signals (Tice et al., 2013; Hsieh et al., 2015). Other influential factors may be unknown or difficult to take into account (Malo et al., 2006). The proximity of wells in microtiter test plates may yield misleading signals due to signal flare or inadvertent contamination. Well-composition could also change over time due to evaporation, alterations in dissolvability, volatility, or chemical reaction. Artifacts can have an unpredictable effect on the biological response (Hsieh et al., 2015). Unfortunately, these design restrictions may lead to discordant intrachemical response patterns even after data normalization.

TABLE 1 | *CASANOVA* classification errors for the given *p*-value threshold.

	***p*-Value threshold**							
	0.00	**0.001**	**0.01**	**0.05**	**0.10**	**0.20**	**0.50**	**1.00**
10-fold AC_{50} spacing								
Dataset 1[a]								
errorA	0.019	0.019	0.019	0.023	0.031	0.051	0.151	0.585
errorB	0.418	0.102	0.057	0.032	0.022	0.014	0.001	0.000
Dataset 2[b]								
errorA	0.022	0.022	0.023	0.027	0.036	0.058	0.163	0.565
errorB	0.392	0.104	0.059	0.033	0.022	0.014	0.006	0.000
Dataset 3[c]								
errorA	0.017	0.017	0.018	0.022	0.028	0.047	0.141	0.588
errorB	0.423	0.098	0.055	0.030	0.021	0.013	0.005	0.000
Dataset 4[d]								
errorA	0.031	0.031	0.032	0.037	0.047	0.072	0.182	0.416
errorB	0.312	0.109	0.064	0.035	0.024	0.014	0.004	0.000
100-fold AC_{50} spacing								
Dataset 1[a]								
errorA	0.014	0.014	0.015	0.020	0.028	0.050	0.166	0.606
errorB	0.348	0.037	0.014	0.005	0.003	0.001	0.000	0.000
Dataset 2[b]								
errorA	0.017	0.017	0.018	0.024	0.033	0.058	0.178	0.583
errorB	0.331	0.039	0.015	0.005	0.003	0.002	0.001	0.000
Dataset 3[c]								
errorA	0.013	0.013	0.014	0.018	0.025	0.046	0.154	0.609
errorB	0.351	0.034	0.012	0.004	0.002	0.001	0.000	0.000
Dataset 4[d]								
errorA	0.025	0.025	0.026	0.034	0.045	0.074	0.202	0.521
errorB	0.273	0.048	0.019	0.006	0.003	0.001	0.000	0.000

[a]*tox21-elg1-luc-agonist,* [b]*tox21-are-bla-p1,* [c]*tox21-er-luc-bg1-4e2-agonist-p2,* [d]*tox21-fxr-bla-agonist-p2.*

In this article we present a simple methodology to group intrachemical repeats in an automated manner. In theory, if a compound is active, then we expect the responses to be active at the lowest tested concentration (i.e., exceeding the noise limits), monotonic, or partially ordered (e.g., up-turn or down-turn responses) with concentrations. Our data driven approach to cluster compound-specific response patterns, termed *CASANOVA*, finds clusters in which repeats group together across the entire concentration-response domain as well as clusters which distinguish repeats in concentration subgroups.

We assessed the consistency of intra-chemical response patterns within and between Tox21 Phase II assays interrogating nuclear receptor activity and stress response. While most chemicals do not exhibit any response in the tested assay conditions, a fraction of compounds (i.e., 1.6 to 23.8% across the tested assays) with at least one profile extending outside of the noise band represent single cluster response patterns (**Figure 2**). Multiplicity in response can often be attributed to one or more known experimental design factors. Still, it may not be possible to account for all confounding factors associated with an observed disparity of responses (e.g., **Figure 1C**). The wide range of AC_{50} estimates obtained for the same compound in experimental data sets (**Figure 3**) underscores the importance of a clustering algorithm such as *CASANOVA* to identify compounds with single cluster patterns of response. Otherwise, compound potency estimates may not be reliable.

Simulation studies were used to evaluate the ability of *CASANOVA* to cluster compound profiles into reliable subgroups and provide suitable AC_{50} potency estimates. The overall error rates for *CASANOVA* to correctly cluster similar patterns ("Type A" errors) and separate disjoint patterns ("Type B" errors) was found to be less than 5% across a range of simulation studies based on Tox21 Phase II qHTS data using 10- or 100-fold AC_{50} spacing. We employed a *p*-value threshold of 0.05 to describe patterns in the Tox21 Phase II data. However, the results from our simulation studies reveal that selecting a less stringent *p*-value threshold (e.g., $p < 0.10$) can be used to increase the "Type A" error and decrease the "Type B" error according to different research motivations. Assuming that all the profiles belong to a single cluster, simple averaging of individual AC_{50} estimates leads to the greatest bias and least precise estimates. However, the weighted average approach produces the most repeatable results, where both bias and variance are generally within one order of magnitude.

TABLE 2 | Bias and variance of $log_{10}AC_{50}$ parameter for Hill model curves (5% error).

True	True		Bias (and variance) of log_{10} AC_{50}			
AC_{50}	*RMAX*	*n*	Avg	Median	WT Avg	One model
1.00e-03 Upper plateau only	25	3	1.26(4.07)	0.42(2.63)	0.03(0.61)	0.52(4.07)
	25	6	1.22(2.13)	0.21(0.61)	0.10(0.07)	0.44(3.16)
	25	9	1.19(1.42)	0.20(0.08)	0.11(0.04)	0.46(3.06)
	25	12	1.24(1.10)	0.08(0.10)	0.10(0.03)	0.41(2.84)
	50	3	0.28(0.52)	0.07(0.19)	0.05(0.10)	0.05(0.11)
	50	6	0.27(0.24)	0.03(0.02)	0.08(0.02)	0.04(0.05)
	50	9	0.26(0.14)	0.02(0.01)	0.09(0.01)	0.04(0.06)
	50	12	0.26(0.11)	0.02(0.01)	0.09(0.01)	0.04(0.06)
	100	3	0.02(0.01)	0.01(0.01)	0.03(0.01)	0.01(0.01)
	100	6	0.03(*)	0.01(*)	0.03(*)	0.01(0.01)
	100	9	0.03(*)	0.01(*)	0.04(*)	0.01(0.01)
	100	12	0.03(*)	*(*)	0.04(*)	0.01(0.01)
0.1 Upper and lower plateaus	25	3	0.13(1.52)	0.01(0.05)	*(0.04)	0.01(0.03)
	25	6	0.08(0.63)	0.01(0.02)	*(0.02)	*(0.03)
	25	9	0.09(0.45)	*(0.01)	*(0.01)	*(0.03)
	25	12	0.07(0.32)	0.01(0.01)	*(0.01)	0.01(0.03)
	50	3	*(0.01)	*(0.01)	*(0.01)	*(0.01)
	50	6	*(*)	*(*)	*(*)	*(0.01)
	50	9	*(*)	*(*)	*(*)	*(0.01)
	50	12	*(*)	*(*)	*(*)	*(0.01)
	100	3	*(*)	*(*)	*(*)	*(*)
	100	6	*(*)	*(*)	*(*)	*(*)
	100	9	*(*)	*(*)	*(*)	*(*)
	100	12	*(*)	*(*)	*(*)	*(*)
10 Lower plateau only	25	3	1.86(4.78)	1.04(5.08)	0.13(1.65)	0.72(4.72)
	25	6	1.91(2.42)	0.72(1.82)	0.06(0.43)	0.73(5.10)
	25	9	1.90(1.70)	0.48(1.20)	0.11(0.11)	0.78(5.27)
	25	12	1.90(1.21)	0.37(0.56)	0.09(0.10)	0.75(4.93)
	50	3	0.74(1.08)	0.30(0.78)	0.03(0.34)	0.09(0.15)
	50	6	0.74(0.51)	0.19(0.16)	0.03(0.13)	0.12(0.30)
	50	9	0.77(0.36)	0.14(0.06)	0.04(0.07)	0.13(0.27)
	50	12	0.75(0.27)	0.12(0.03)	0.07(0.03)	0.12(0.24)
	100	3	0.14(0.10)	0.06(0.03)	0.01(0.04)	0.02(0.01)
	100	6	0.14(0.06)	0.04(0.01)	0.02(0.02)	0.02(0.01)
	100	9	0.14(0.04)	0.03(0.01)	0.03(0.01)	0.02(0.01)
	100	12	0.14(0.02)	0.03(*)	0.03(*)	0.02(0.01)

Values of bias or variance less than 0.01 are indicated by "". *(*) indicates that both the bias and the variance are less than 0.01.*

The *CASANOVA* approach provides an unsupervised method to agnostically separate multiple cluster response compounds from compounds with reasonably concordant concentration-response repeats. Our approach therefore avoids a complicated modeling effort to account for all potentially influential variables in the data, many of which may not be explicit or identifiable in any given study. Compound potency estimates in qHTS experiments can vary substantially (well over 100-fold in some cases) in large scale *in vitro* bioassay data due to multiple cluster intrachemical responses. Lim et al. (2013) discussed possible strategies to derive optimal experimental designs for qHTS experiments to improve the precision of potency estimates and statistical inference on these parameters. Nevertheless, *CASANOVA* can improve the detection of single cluster intrachemical repeats and potency estimation for candidate hits irrespective of the underlying study design. Multiple cluster compounds identified using *CASANOVA* can be studied further to understand the source of the variation which may arise from technological disturbances such as compound carryover, interference between signal channels, autofluorescence, or potential fluctuations in the laboratory environment. However, by focusing research efforts on compounds with single cluster response patterns, potency estimation is expected to be more accurate and precise. We anticipate that *CASANOVA* can be applied to other types of sequential data types involving non-linear responses, including dose-response and longitudinal genomics studies, where divergent responses in subregions of the data are important. The R code for CASANOVA is

available upon request or can be downloaded online from www.niehs.nih.gov/research/atniehs/labs/bb/staff/shockley/index.cfm.

AUTHOR CONTRIBUTIONS

KS and SP designed the study, analyzed the data, and wrote the manuscript. KS, SP, SH, SG, and SL edited the manuscript. SP, SG, and SL conceived the application of two-way ANOVA algorithm for qHTS. SH performed automation of *CASANOVA*.

FUNDING

This work was supported (in part) by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (ZIA ES102865 and Contract HHSN273201600011C).

ACKNOWLEDGMENTS

We thank Dr. Grace Kissling (NIEHS), Dr. Marjo Smith (Social and Scientific Systems), and Dr. Raymond Tice (NIEHS) for providing helpful suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00387/full#supplementary-material

REFERENCES

Abdo, N., Xia, M., Brown, C. C., Kosyk, O., Huang, R., Sakamuru, S., et al. (2015). Population-based in vitro hazard and concentration-response assessment of chemicals: the 1000 genomes high-throughput screening study. *Environ. Health Pers.* 123, 458–466. doi: 10.1289/ehp.1408775

Anthony Tony Cox, L., Popken, D. A., Kaplan, A. M., Plunkett, L. M., and Becker, R. A. (2016). How well can in vitro data predict in vivo effects of chemicals? Rodent carcinogenicity as a case study. *Regul. Toxicol. Pharmacol.* 77, 54–64. doi: 10.1016/j.yrtph.2016.02.005

Attene-Ramos, M. S., Miller, N., Huang, R., Michael, S., Itkin, M., Kavlock, R. J., et al. (2013). The Tox21 robotic platform for the assessment of environmental chemicals–from vision to reality. *Drug Disc. Today* 18, 716–723. doi: 10.1016/j.drudis.2013.05.015

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Bouhaddou, M., DiStefano, M. S., Riesel, E. A., Carrasco, E., Holzapfel, H. Y., Jones, D. C., et al. (2016). Drug response consistency in CCLE and CGP. *Nature* 540, E9–E10.

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

Chen, I., Mathews-Greiner, L., Li, D., Abisoye-Ogunniyan, A., Ray, S., Bian, Y., et al. (2017). Transcriptomic profiling and quantitative high-throughput (qHTS) drug screening of CDH1 deficient hereditary diffuse gastric cancer (HDGC) cells identify treatment leads for familial gastric cancer. *J. Trans. Med.* 15:92. doi: 10.1186/s12967-017-1197-5

Collins, F. S., Gray, G. M., and Bucher, J. R. (2008). Toxicology. transforming environmental health protection. *Science* 319, 906–907. doi: 10.1126/science.1154619

Dahlin, J. L., Inglese, J., and Walters, M. A. (2015). Mitigating risk in academic preclinical drug discovery. *Nat. Rev. Drug Dis.* 14, 279–294. doi: 10.1038/nrd4578

Ding, K. F., Finlay, D., Yin, H., Hendricks, W. P. D., Sereduk, C., Kiefer, J., et al. (2017). Analysis of variability in high throughput screening data: applications to melanoma cell lines and drug responses. *Oncotarget* 8, 27786–27799. doi: 10.18632/oncotarget.15347

Eduati, F., Mangravite, L. M., Wang, T., Tang, H., Bare, J. C., Huang, R., et al. (2015). Prediction of human population responses to toxic compounds by a collaborative competition. *Nat. Biotechnol.* 33, 933–940. doi: 10.1038/nbt.3299

Filer, D. L., Kothiya, P., Setzer, R. W., Judson, R. S., and Martin, M. T. (2017). Tcpl: the ToxCast pipeline for high-throughput screening data. *Bioinformatics* 33, 618–620. doi: 10.1093/bioinformatics/btw680

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483, 570–575. doi: 10.1038/nature11005

Geeleher, P., Gamazon, E. R., Seoighe, C., Cox, N. J., and Huang, R. S. (2016). Consistency in large pharmacogenomic studies. *Nature* 540, E1–E2.

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J., et al. (2013). Inconsistency in large pharmacogenomic studies. *Nature* 504, 389–393. doi: 10.1038/nature12831

Haverty, P. M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., et al. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* 533, 333–337. doi: 10.1038/nature17987

Hill, A. V. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J. Physiol.* 40, 4–7. doi: 10.1371/journal.pone.0041098

Hsieh, J. H., Sedykh, A., Huang, R., Xia, M., and Tice, R. R. (2015). A data analysis pipeline accounting for artifacts in Tox21 quantitative high-throughput screening assays. *J. Biomol. Screen.* 20, 887–897. doi: 10.1177/1087057115581317

Huang, R., Xia, M., Cho, M. H., Sakamuru, S., Shinn, P., Houck, K. A., et al. (2011). Chemical genomics profiling of environmental chemical modulation of human nuclear receptors. *Environ. Health Pers.* 119, 1142–1148. doi: 10.1289/ehp.1002952

Inglese, J., Auld, D. S., Jadhav, A., Johnson, R. L., Simeonov, A., Yasgar, A., et al. (2006). Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11473–11478. doi: 10.1073/pnas.0604348103

Jamal, S., Arora, S., and Scaria, V. (2016). Computational analysis and predictive cheminformatics modeling of small molecule inhibitors of epigenetic modifiers. *PloS One* 11:e0083032. doi: 10.1371/journal.pone.0083032

Judson, R., Houck, K., Martin, M., Richard, A. M., Knudsen, T. B., Shah, I., et al. (2016). Editor's highlight: analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space. *Toxicol. Sci.* 152, 323–339. doi: 10.1093/toxsci/kfw092

Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., et al. (2010). In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ. Health Pers.* 118, 485–492. doi: 10.1289/ehp.0901392

Karmaus, A. L., Filer, D. L., Martin, M. T., and Houck, K. A. (2016). Evaluation of food-relevant chemicals in the ToxCast high-throughput screening program. *Food Chem. Toxicol.* 92, 188–196. doi: 10.1016/j.fct.2016.04.012

Kleinstreuer, N. C., Ceger, P., Watt, E. D., Martin, M., Houck, K., Browne, P., et al. (2017). Development and validation of a computational model for androgen receptor activity. *Chem. Res. Toxicol.* 30, 946–964. doi: 10.1021/acs.chemrestox.6b00347

Kleinstreuer, N. C., Dix, D. J., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., et al. (2013). In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis. *Toxicol. Sci.* 131, 40–55. doi: 10.1093/toxsci/kfs285

Kleinstreuer, N. C., Yang, J., Berg, E. L., Knudsen, T. B., Richard, A. M., Martin, M. T., et al. (2014). Phenotypic screening of the ToxCast chemical library

to classify toxic and therapeutic mechanisms. *Nat. Biotechnol.* 32, 583–591. doi: 10.1038/nbt.2914

Lim, C., Sen, P. K., and Peddada, S. D. (2013). Robust analysis of high throughput screening (HTS) assay data. *Technometrics* 55, 150–160. doi: 10.1080/00401706.2012.749166

Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J., and Nadon, R. (2006). Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* 24, 167–175. doi: 10.1038/nbt1186

Martin, M. T., Knudsen, T. B., Reif, D. M., Houck, K. A., Judson, R. S., Kavlock, R. J., et al. (2011). Predictive model of rat reproductive toxicity from toxCast high throughput screening. *Biol. Reprod.* 85, 327–339. doi: 10.1095/biolreprod.111.090977

Martinez, N. J., Rai, G., Yasgar, A., Lea, W. A., Sun, H., Wang, Y., et al. (2016). A high-throughput screen identifies 2,9-diazaspiro[5.5]undecanes as inducers of the endoplasmic reticulum stress response with cytotoxic activity in 3d glioma cell models. *PloS One* 11:e0161486. doi: 10.1371/journal.pone.0161486

Mpindi, J. P., Yadav, B., Ostling, P., Gautam, P., Malani, D., Murumagi, A., et al. (2016). Consistency in drug response profiling. *Nature* 540, E5–E6.

Rotroff, D. M., Wetmore, B. A., Dix, D. J., Ferguson, S. S., Clewell, H. J., Houck, K. A., et al. (2010). Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. *Toxicol. Sci.* 117, 348–358. doi: 10.1093/toxsci/kfq220

Roy, A., McDonald, P. R., Sittampalam, S., and Chaguturu, R. (2010). Open access high throughput drug discovery in the public domain: a mount everest in the making. *Curr. Pharm. Biotechnol.* 11, 764–778. doi: 10.2174/138920110792927757

Safikhani, Z., El-Hachem, N., Smirnov, P., Freeman, M., Goldenberg, A., Birkbak, N. J., et al. (2016a). Safikhani et al. reply. *Nature* 540, E2–E4.

Safikhani, Z., El-Hachem, N., Smirnov, P., Freeman, M., Goldenberg, A., Birkbak, N. J., et al. (2016b). Safikhani et al. reply. *Nature* 540, E6–E8.

Safikhani, Z., El-Hachem, N., Smirnov, P., Freeman, M., Goldenberg, A., Birkbak, N. J., et al. (2016c). Safikhani et al. reply. *Nature* 540, E11–E12.

Safikhani, Z., El-Hachem, N., Quevedo, R., Smirnov, P., Goldenberg, A., Juul Birkbak, N., et al. (2016d). Assessment of pharmacogenomic agreement. *F1000Res.* 5:825. doi: 10.12688/f1000research.8705.1

Shockley, K. R. (2012). A three-stage algorithm to make toxicologically relevant activity calls from quantitative high throughput screening data. *Environ. Health Persp.* 120, 1107–1115. doi: 10.1289/ehp.1104688

Shockley, K. R. (2015). Quantitative high-throughput screening data analysis: challenges and recent advances. *Drug Disc. Today* 20, 296–300. doi: 10.1016/j.drudis.2014.10.005

Shockley, K. R. (2016). Estimating potency in high-throughput screening experiments by maximizing the rate of change in weighted shannon entropy. *Sci. Rep.* 6:27897. doi: 10.1038/srep27897

Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: a Tox21 update. *Environ. Health Pers.* 121, 756–765. doi: 10.1289/ehp.1205784

U.S. National Toxicology Program [U.S. NTP] (2017). *About the NTP*. Available at: https://ntp.niehs.nih.gov/about/ (accessed December 12, 2017).

Wang, Y., Jadhav, A., Southal, N., Huang, R., and Nguyen, D. T. (2010). A grid algorithm for high throughput fitting of dose-response curve data. *Curr. Chem. Genomics* 4, 57–66. doi: 10.2174/1875397301004010057

Weinstein, J. N., and Lorenzi, P. L. (2013). Cancer: Discrepancies in drug sensitivity. *Nature* 504, 381–383. doi: 10.1038/nature12839

Wetmore, B. A., Wambaugh, J. F., Ferguson, S. S., Sochaski, M. A., Rotroff, D. M., Freeman, K., et al. (2012). Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. *Toxicol. Sci.* 125, 157–174. doi: 10.1093/toxsci/kfr254

Xu, M., Lee, E. M., Wen, Z., Cheng, Y., Huang, W. K., Qian, X., et al. (2016). Identification of small-molecule inhibitors of *Zika virus* infection and induced neural cell death via a drug repurposing screen. *Nat. Med.* 22, 1101–1107. doi: 10.1038/nm.4184

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership