Research Article

In Vitro to *In Vivo* Extrapolation for Drug-Induced Liver Injury Using a Pair Ranking Method

Zhichao Liu [1], Hong Fang [1], Jürgen Borlak [2], Ruth Roberts [3,4] and Weida Tong [1]

[1]National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, USA; [2]Centre for Pharmacology and Toxicology, Hannover Medical School, Hannover, Germany; [3]ApconiX, BioHub at Alderley Park, Alderley Edge, UK; [4]University of Birmingham, Edgbaston, Birmingham, UK

Summary

Preclinical animal toxicity studies may not accurately predict human toxicity. In light of this, *in vitro* systems have been developed that have the potential to supplement or even replace animal use. We examined *in vitro* to *in vivo* extrapolation (IVIVE) of gene expression data obtained from The Open Japanese Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-GATEs) for 131 compounds given to rats for 28 days, and to human or rat hepatocytes for 24 hours. Notably, a pair ranking (PRank) method was developed to assess IVIVE potential with a PRank score based on the preservation of the order of similarity rankings of compound pairs between the platforms using a receiver operating characteristic (ROC) curve analysis to measure area under the curve (AUC). A high IVIVE potential was noted for rat primary hepatocytes when compared to rat 28-day studies (PRank score = 0.71) whereas the IVIVE potential for human primary hepatocytes compared to rat 28-day studies was lower (PRank score = 0.58), indicating that species difference plays a critical role in IVIVE. When limiting the analysis to only those drugs causing drug-induced liver injury, the IVIVE potential was slightly improved both for rats (from 0.71 to 0.76) and for humans (from 0.58 to 0.62). Similarly, PRank scores were improved when the analysis focused on specific hepatotoxic endpoints such as hepatocellular injury, or cholestatic injury. In conclusion, toxicogenomic data generated *in vitro* yields a ranking of drugs regarding their potential to cause toxicity which is comparable to that generated by *in vivo* analyses.

Keywords: drug-induced liver injury (DILI), toxicogenomics, IVIVE

1 Introduction

Before a potential new medicine can progress to clinical trials in humans, it must be assessed for safety and tolerability in both rodent and non-rodent toxicology studies to limit and manage risk to human volunteers and patients. This current paradigm is based on law and also on historical data that show a concordance of the toxicity of pharmaceuticals in humans and animals (Olson et al., 2000). However, this concordance is challenged by many groups who argue that a systematic review of animal data demonstrate poor human clinical correlation and question the validity of these models (Knight, 2007; Bailey et al., 2013).

In addition to questions on the correlation between animal models and human toxicity, worldwide efforts are being made to reduce animal testing and to enhance safety assessment largely based on developing *in vitro* systems or *in silico* approaches tailored to toxicologically relevant mechanisms. An example of this is the REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) program in Europe (Abbott, 2005), which is driving for earlier identification of the intrinsic properties of chemical substances, aligned with encouraging methods to "reduce, refine and replace" (3Rs) animal testing. In the United States, there are several relevant programs such as "Advancing Regulatory Science" initiated by the Food and Drug Administration (FDA) (Hamburg, 2011), the Toxicology in the 21st Century (Tox21) program led by multiple governmental agencies (Tice et al., 2013) and the ToxCast program (Dix et al., 2007) developed by the U.S. Environmental Protection Agency

Disclaimer: The views presented in this article do not necessarily reflect current or future opinion or policy of the US Food and Drug Administration. Any mention of commercial products is for clarification and not intended as an endorsement.

Received October 20, 2016;
Accepted January 2, 2017;
Epub January 11, 2017;
doi:10.14573/altex.1610201

(EPA). The common driver for these programs is to encourage the development of animal-free approaches with the assistance of *in silico* methodologies for enhancing drug safety and risk assessment (DeJongh et al., 1999). A key indicator of success will be to define which methodological approaches conducted either *in vitro* or *in vivo* offer the best correlation with human data and therefore the best opportunity to predict human toxicity. Among these methodologies, toxicogenomics (TGx) has great potential as highlighted by a Health and Environmental Sciences Institute (HESI) survey (Pettit et al., 2010).

TGx has drawn wide attention as an approach to study the underlying molecular mechanisms of toxicity since it allows the generation of large and highly informative data sets that lend themselves to bioinformatic analyses (Nuwaysir et al., 1999; Aardema and MacGregor, 2002). Efforts have been made to investigate the potential of TGx data to supply better information for toxicity assessment than traditional animal toxicology studies (Liu et al., 2011, 2016; Ippolito et al., 2016; Sutherland et al., 2016). Furthermore, TGx has been used to investigate similarities between *in vitro* and *in vivo* assay systems and to develop translational biomarkers across the species. Hrach et al. (2011) developed an *in vitro* liver toxicity prediction model based on a rat primary hepatocyte sandwich culture, generating a 724-gene signature model that is capable of discriminating compounds according to their *in vivo* hepatotoxicity with a misclassification rate of only 7.5%. Furthermore, in order to assess the difference between *in vitro* systems and *in vivo* systems, several studies focused on comparing various *in vitro* liver models against liver tissue from *in vivo* exposure in terms of their gene expression profiles. Boess et al. (2003) demonstrated that *in vitro* TGx results, regardless of the system used, do not directly compare to the results obtained *in vivo*, at least not on a gene to gene comparison basis. Cheng et al. (2011) developed a novel genomic prediction technique based on rat *in vivo* TGx data, where a subset of 32 genes were subsequently used to predict hepatotoxicity in test sets of *in vitro* human liver and *in vivo* animal toxicity experiments. Deng et al. (2010) investigated 2,4,6-trinitrotoluene (TNT) effects on gene expression in the liver. It was suggested that gene regulatory networks obtained from an *in vitro* system can predict *in vivo* function and mechanisms. We recently described a text-mining methodology based on topic modeling that could assist in exploring the correlation between *in vitro* and *in vivo* assay systems (Lee et al., 2014, 2016; Chung et al., 2015).

One limitation of many of the reported studies is that they are based on relatively small numbers of compounds, limiting the statistical measures that can be used for a comprehensive assessment, which in turn limits conclusions that can be drawn in general (Chen et al., 2012). Several large TGx databases derived from well-designed studies are available such as The Open Japanese Toxicogenomics Project (TG-GATEs) (Uehara et al., 2010; Igarashi et al., 2015), DrugMatrix (Ganter et al., 2005) and PredTox (Suter et al., 2011). These databases provide the opportunity to evaluate systematically different TGx assay systems for their consistency, predictivity and their ability to detect underlying toxicity mechanisms. Several studies have been reported to address different toxicological questions using these big TGx datasets (Otava et al., 2015; Hardt et al., 2016; Bell et al., 2016; Sutherland et al., 2016; Liu et al., 2016).

In this study, we examined *in vitro* to *in vivo* extrapolation (IVIVE) potential by assessing the similarity of gene activities between *in vitro* and *in vivo* TGx systems using Open TG-GATEs, a TGx database that stores gene expression profiles and traditional toxicological data derived from *in vivo* (rat) and *in vitro* studies (primary rat hepatocytes or primary human hepatocytes) on 131 compounds at multiple doses/concentrations and time points. A pair ranking (PRank) method was developed to assess the IVIVE potential, based on determining ranking preservation of drug-drug pairs according to their transcriptomic profiles between two assay systems. Furthermore, we examined IVIVE potential for specific human hepatotoxic endpoints such as drug-induced liver injury, hepatocellular injury, and cholestatic injury.

2 Materials and methods

Toxicogenomics datasets

The Open Japanese Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-GATEs, http://toxico.nibiohn.go.jp/english/) was employed for studying IVIVE potential (Uehara et al., 2010; Igarashi et al., 2015). The dataset was generated in two phases. Phase I, which involved 131 compounds (mainly drugs), was used in this study. The *in vivo* data was derived from standard preclinical studies with daily dosing using three doses (low, medium and high) and four treatment durations (3, 7, 14, and 28 days), sampled 24 h after the last dose. The *in vitro* data used 3 concentrations (low, medium and high) and 3 treatment durations (2, 4, 24 h). For the *in vitro* study, the highest concentration was defined as that yielding 80-90% survival; the low and middle concentrations were then derived using a 1:5:25 ratio for low: medium: high, all given as μM in DMSO. For the *in vivo* study, 0.5% methylcellulose or corn oil (oral dose, mg/kg), saline or 5% glucose solution (intravenous dose, mg/kg) were employed for repeat dosing with a ratio of the low, middle and high dose of 1:3:10. Blood samples for routine bio-chemical analyses were collected in heparinized tubes under ether anesthesia from the abdominal aorta at the time of sacrifice. In total, 12 time/dose combinations of each compound were profiled for the *in vivo* samples while 9 time/concentration combinations of each compound were profiled for the *in vitro* samples (Tab. S1[1]). There were 120 common compounds among the three assays.

Microarray data processing

Microarray data were processed with two methods, MAS5 (Hubbell et al., 2002) and FARMS (Factor Analysis for Robust Microarray Summarization) (Hochreiter et al., 2006). An

[1] doi:10.14573/altex.1610201s

extensive comparison of the results between the two methods was made and found to give similar results (see Discussion section), therefore we present FARMS processed data. The Open TG-GATEs data were downloaded from CAMDA 2013[2]. Briefly, for each compound, the gene expression profiles were generated for two (*in vitro*) or three (*in vivo*) replicate samples and two or three matched control replicate samples. First, the probe-level data of the microarrays were quantile normalized. Second, a compound batch correction was made by calculating the probe intensity ratios using the corresponding control measurement for the cell culture (vehicle-only without compound) as a reference. For the next preprocessing step, probe sets were defined corresponding to genes using alternative chip definition files (CDFs) (Version 15.1.0, ENTREZG) (Dai et al., 2005)

[2] http://dokuwiki.bioinf.jku.at/doku.php/contest_dataset

Tab. 1: IVIVE potential for hepatotoxic endpoints

Categories*	Number of compounds	InVitro_Rat-InVivo	InVitro_Human-InVivo	InVitro_Rat-InVitro_Human
All compounds	120	0.71	0.58	0.77
Drug induced liver injury (DILI)				
Most DILI concern	46	0.76	0.62	0.74
Xu's label	47	0.82	0.53	0.73
Sakatis's label	51	0.77	0.63	0.73
Hepatic failure	24	0.77	0.72	0.82
Other hepatotoxic endpoints				
Biochemical parameters				
AST increased	13	0.69	0.62	0.78
ALT increased	9	0.81	0.52	0.88
Hepatic enzyme increased	8	0.71	0.68	0.80
Blood bilirubin increased	6	0.81	–	–
Liver injury patterns				
Cholestasis	7	0.95	0.71	0.70
Hepatocellular injury	16	0.75	0.57	0.89
Hepatobiliary abnormality				
Cholelithiasis	6	0.86	–	–
Foetor hepaticus	7	0.81	0.62	0.96
Hepatomegaly	5	0.89	0.59	0.50
Jaundice	49	0.69	0.63	0.76
Jaundice cholestatic	21	0.40	0.61	0.74
Histologic findings				
Hyperbilirubinaemia	7	0.84	0.68	0.78
Hepatitis	43	0.70	0.61	0.73
Hepatic cirrhosis	7	0.80	0.44	0.89
Liver disorder	8	0.82	0.68	0.96
Hepatic function abnormal	33	0.77	0.62	0.71
Steatosis	13	0.67	0.70	0.85
Hepatic necrosis	13	0.64	0.63	0.68
Cytolytic hepatitis	10	–	0.61	0.87
Total (%) of IVIVE score increased		**78.3%**	**73.9%**	**43.5%**

* The different DILI endpoints are based on published datasets as described in the Materials and Methods section. The value in the table is the area under curve (AUC) value for comparison between assay systems based on the proposed PRank methodology.

from Brainarray[3] and applied FARMS (Hochreiter et al., 2006) for summarizing the intensity ratios at probe set level to obtain expression values per gene. Finally, the replicate samples were collapsed and the ratio data were calculated by using collapsed treated samples divided by the collapsed control samples.

Drug-induced liver injury endpoints

Three publicly available drug-induced liver injury (DILI) classification schemes were used: NCTR DILI annotation (Chen et al., 2011), Xu's dataset (Xu et al., 2008), and Sakatis's dataset (Sakatis et al., 2012). The drugs from the NCTR dataset (Chen et al., 2011) were divided into three categories based on their DILI potential: most-DILI concern; less-DILI concern; no-DILI concern. In this study, only those drugs falling into the "most-DILI-concern" category were considered from the NCTR dataset (Chen et al., 2011). The data in Xu's dataset (Xu et al., 2008) are annotated by cellular imaging of several toxicity endpoints in primary human hepatocyte cultures. The Sakatis dataset (Sakatis et al., 2012) employs *in vitro* bio-activation data to identify DILI risk of > 200 compounds. In this study, the DILI positive drugs in each of the annotations were mapped onto the Open TG-GATEs compound list for further analysis (Tab. S2[1]).

To address the complexity of DILI in humans, we also incorporated hepatotoxic related clinical manifestations in our analysis. Specifically, the human hepatic-related side effects were downloaded from the SIDER database[4] (Kuhn et al., 2010, 2016), which is based on Natural Language Processing (NLP) from drug labels and side effect terms standardized using the Medical Dictionary for Regulatory Activities (MedDRA v16.1) preferred terms (PTs). The hepatic related side effects were collected by attributing PTs to their primary System Organ Class (SOC) level of *hepatobiliary disorders* in MedDRA. Compounds that caused hepatic steatosis were curated from a literature survey and corrected by domain experts by observation of pathologic images (Liu et al., 2016; Sahini et al., 2014) (Tab. S2[1]). The collated hepatotoxic endpoints occurring across the 120 compounds and the three assays were further divided by domain experts into five different categories: liver transaminase elevations, hepatobiliary abnormality, histologic findings, liver injury patterns and severity of liver injury (Tab. 1).

Principal component analysis (PCA) and hierarchical clustering analysis (HCA)

The number of differentially expressed genes (DEGs) obtained from a treatment depends on the treatment condition. Use of different toxicants or varying doses/concentrations and treatment duration of the same toxicant will lead to a change in the number of DEGs identified. Thus, treatment state and effect is reflected on the number of DEGs identified. Specifically, a matrix (dose/concentration)/time *vs* compounds) was generated with each element denoting a number of DEGs for the corresponding treatment condition. Then, a principal component analysis (PCA) was implemented on the matrix. The Matlab function *princomp.m* under Statistics and Machine Learning Toolbox (Matlab R2014a) was employed to carry out the PCA calculation. Furthermore, we also carried out a hierarchical clustering analysis (HCA) to further investigate the treatment effect reflected in the TGx data within the same matrix.

Pair ranking (PRank) method

In order to study the IVIVE potential in TGx, we developed a pair ranking (PRank) method including the following steps:

1) We first ranked genes by fold change (treated *vs* control) for each compound. The top 200 (the highest up-regulated) and bottom 200 (the highest down-regulated) genes were selected as the signature for the compound. In each TGx assay system (e.g., *in vitro* or *in vivo*), the similarity between any two compounds was assessed by comparing their signatures using Dice's coefficient (Wang et al., 2014), as shown in the formula below,

$$D_{i,j} = \frac{2(N_{i,j})}{N_i + N_j} = \frac{2(N_{i,j})}{800} = \frac{N_{i,j}}{400}$$

where *Ni*, *j* denotes the number of overlapping genes between compound *i* and compound *j*, *Ni* and *Nj* represent the number of significant genes of compound *i* and compound *j*, respectively.

2) The compound-compound pairwise similarities were ranked from the most similar to least similar in each assay system.
3) The PRank score (a scale of 0~1) measured the IVIVE potential between two testing systems that was determined based on the area under curve (AUC) value from a receiver operating characteristic (ROC) curve analysis, which is used to measure the extent of rank preservation of the ranked similarity lists from two systems.

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis

The Database for Annotation, Visualization, and Integrated Discovery (DAVID)[5] (Huang et al., 2008) was used to conduct the KEGG pathway analysis. Specifically, the KEGG pathway was enriched by using DAVID with the top 200 and down 200 genes. A Benjamini-Hochberg adjusted *p* value less than 0.05 was used as a cut-off to identify the over-representative pathways.

Chemical structure similarity

The structural similarity among the compounds was also calculated to compare the compound pairwise similarity based on chemical structure information of compounds in the Open TG-GATEs database (see the chemical structure information in Table S3[1]). Specifically, the well-established functional class fingerprints (FCFPs) with a radius of FCFP 4 were used as chemical descriptors to calculate Tanimoto coefficients between two compounds (Hassan et al., 2006), which were implemented in Pipeline Pilot v8.0 (Accelrys, Biovia, and Dassault Systems).

3 http://brainarray.mbni.med.umich.edu/Brainarray/default.asp

4 http://sideeffects.embl.de/

5 http://david.abcc.ncifcrf.gov/

3 Results

3.1 The highest dose/concentration and longest exposure showed the most variance

TG-GATEs experiments were conducted using multiple time and dose/concentration levels for each compound. Figure 1 shows the PCA results based on the number of DEGs at each time and dose/concentration combination. It showed that the low and middle doses/concentrations even at different times clustered together, indicating a weak response to the assay system. However, the highest dose/concentration and longest duration in each assay system showed the most variance. We further used HCA to investigate the relationship among treatment/system/time/dose (or concentration), as shown in Figure S1[1]. It was observed that the high dose/concentration and long durations were clustered in each assay system. Therefore, only the highest dose/concentration and longest duration of treatment was taken forward for further analysis (28 day, highest dose; 24 h, highest concentration).

3.2 The 200 genes from the top and down ranked gene list generated a stable list

Before calculating similarity between compounds in each assay system, the number of DEGs in the assay system was investigated (Fig. S2[1]). It was shown that the standard deviation of DEG distribution within the two *in vitro* assay systems was higher than in the *in vivo* system, which indicated that the compound response was the main influencing effect in the *in vitro* assay and the animal response had more influence in the *in vivo* assay system. Furthermore, the average number of DEG was quite different in the three TGx assay systems. Therefore, we used a fixed number of gene signatures of each compound and calculated pair-wise similarity (Iorio et al., 2010). To address the question of how many genes are sufficient to faithfully represent the compound response to the assay system, the rank order of the gene list for each compound was generated based on fold change from high to low. The top *N* genes from the up and down ranked list were picked to denote the DEG for each drug ($N = 50$: 50: 500). The pair-wise similarity was generated using Dice's coefficient between all the compound pairs. Ordered compound pair lists were generated based on pair-wise similarity (Spearman's correlation coefficient), which was used to investigate the stability of the ranked compound pair list (Fig. S3[1]). It could be seen that the ranked order of compound pairs tended to be stable with more than 200 genes from the top and down ranked gene list. Therefore, the top 200 and bottom 200 genes were selected for each compound as the signature to calculate pairs-wise similarity in the PRank process.

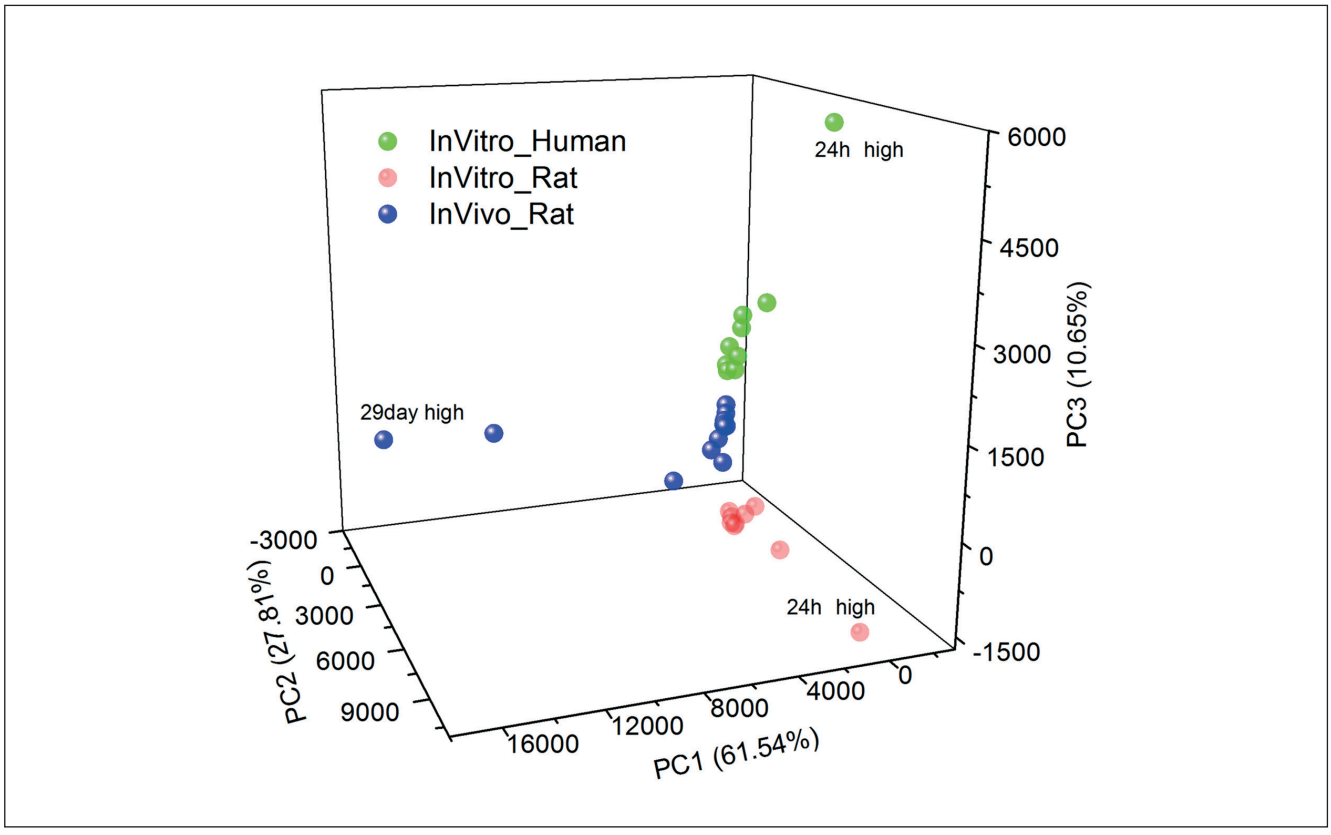

Fig. 1: PCA analysis of three toxicogenomics assays

In each compound/time/dose (concentration)/assay setting, the number of differentially expressed genes (DEGs) was calculated with the criteria fold change of ≥ 1.5 and p value ≤ 0.05. Then, the matrix about drug *vs* different time/dose/assay could be constructed using the number of DEGs. Finally, the PCA analysis was applied to the matrix and the first three PCs were drawn.

3.3 Rat *in vitro* showed high concordance with rat *in vivo*

Figure 2A shows the concordance among the three testing systems (referred to as InVivo_Rat, InVitro_Rat, and InVitro_Human hereafter) based on their PRank score. Concordance was highest between the two *in vitro* systems, i.e., InVitro_Rat and InVitro_Human (score = 0.77), indicating that species difference was less pronounced within the same testing system. There was also a high IVIVE potential (score = 0.70) between InVitro_Rat and InVivo_Rat, indicating the potential to substitute animal models with an animal-free *in vitro* assay. However, the inter-species concordance between InVitro_Human and InVivo_Rat was much lower (score = 0.58).

Over-represented KEGG pathways for each compound were also determined in each testing platform. Taking all the compounds together, we examined the concordance between the TGx testing platforms with respect to the pathways perturbed by the compounds and ranked the pathway by its frequency of over-representation in each testing system. The concordance

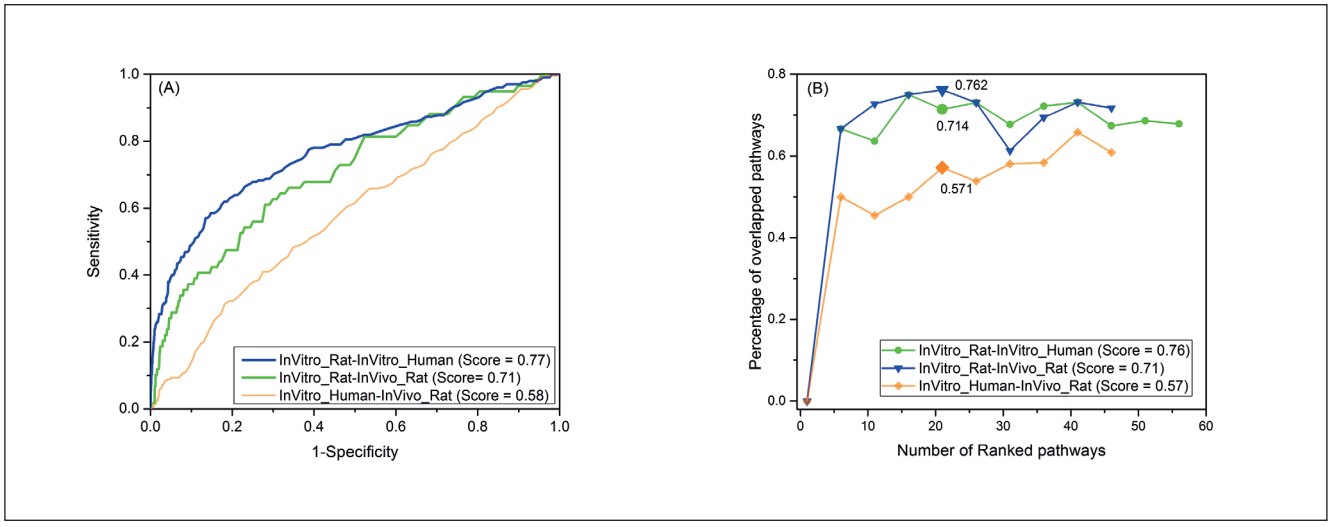

Fig. 2: The concordance among the three different toxicogenomic assay systems

(A) PRank methodology: the gene expression data is processed using FARMERS, (B) The percentage of overlapping adverse outcome pathways (AOPs) among the three toxicogenomic assay systems.

Tab. 2: Common pathways between InVitro_Rat and InVivo_rat

KEGG entry	Pathway names	Class
rno00830	Retinol metabolism	Lipid metabolism
rno00140	Steroid hormone biosynthesis	Lipid metabolism
rno01040	Biosynthesis of unsaturated fatty acids	Lipid metabolism
rno00071	Fatty acid metabolism	Lipid metabolism
rno00100	Steroid biosynthesis	Lipid metabolism
rno00280	Valine, leucine and isoleucine degradation	Amino acid metabolism
rno00330	Arginine and proline metabolism	Amino acid metabolism
rno00380	Tryptophan metabolism	Amino acid metabolism
rno00480	Glutathione metabolism	Metabolism of other amino acids
rno00982	Drug metabolism	Xenobiotics biodegradation and metabolism
rno00980	Metabolism of xenobiotics by cytochrome P450	Xenobiotics biodegradation and metabolism
rno00650	Butanoate metabolism	Carbohydrate metabolism
rno03320	PPAR signaling pathway	Endocrine system

between the two testing platforms was defined as the percentage of overlapped pathways (POP) between them. As depicted in Figure 2B, the highest concordance in the POP plot (0.76) was between the two *in vitro* systems, followed by InVitro_Rat-InVivo_Rat (0.71) and InVitro_Human-InVivo_Rat (0.57). Thus, the pathway-level analysis was consistent with the findings from the proposed PRank methodology at the gene level. Thirteen pathways were common between InVitro_Rat and InVivo_Rat (Tab. 2), most of them involved in different metabolic processes such as lipid metabolism.

Since TGx data are generated using microarray technology, the influence of the different preprocessing strategies was also investigated. Here, the TGx data of three different assay systems were processed using a MAS5.0 expression summary (Pepper et al., 2007). The same conclusion could be drawn from the analysis (see Fig. S4[1]).

3.4 IVIVE potential for drug-induced liver injury (DILI)

We further examined whether IVIVE potential could be improved when the proposed PRank methodology was applied to different DILI-related endpoints. To this end, we analyzed IVIVE for four groups of compounds (under DILI classification in Tab. 1), two of which are associated with severe DILI ("most DILI concern" and "hepatic failure") and two for general DILI (Xu and Sakatis labels). The IVIVE potential (PRank score) of InVitro_Rat-InVivo_Rat was marginally increased by 7% for all four groups ($Score_{Most\text{-}DILI\ concern} = 0.76$, $Score_{hepaticFailure} = 0.77$, $Score_{Xu} = 0.82$, and $Score_{Sakatis} = 0.77$) in comparison to all the compounds studied (Score = 0.71). In addition, the PRank scores of InVitro_Human-InVivo_Rat were also improved for three out of four DILI groups. In contrast, a slight decrease in the score was observed between the two *in vitro* systems for three of the four DILI groups (from 0.77 to 0.73-0.74, last column of Tab. 1).

To ensure that the above observations were not due to chance, an equal number of compounds from all the study compounds in each of four DILI groups was selected and analyzed by PRank individually. This process was repeated $N = 100{,}000$ to remove the potential bias in the compound selection process. Results in Table 1 were statistically significantly different from those of the random test with an adjusted p value less than 1×10^{-6}.

Lastly, compounds were sorted according to their DILI manifestations and the PRank analysis was conducted for each group. There was a total of 20 different DILI endpoints grouped into 4 categories (see Tab. 1). The IVIVE potential was significantly improved for most of the categories when compared to all the compounds used.

4 Discussion

Owing to the poor correlation between animal models and human toxicology, tremendous efforts have been made to investigate whether *in vitro* systems or *in silico* approaches could provide better representation of toxicologically relevant mechanisms. Among these approaches, toxicogenomics (TGx) shows great potential to facilitate method developments for predicting long-term toxic effects and to provide mechanistic elucidation at the molecular level (Pettit et al., 2010). However, if TGx data generated *in vitro* are to supplement or replace animal use in predicting human safety, it is key to understand how the data are similar to those from rat *in vivo* systems. We therefore explored *in vitro* to *in vivo* extrapolation using gene expression data retrieved from TG-GATEs to determine the utility of IVIVE.

Overall, the data show a good potential for InVitro_Rat to InVivo_Rat extrapolation suggesting that a 1-day *in vitro* TGx system could yield meaningful data. When the analysis was limited to DILI related drugs, the IVIVE potential improved in both systems. An analysis of POPs as an indicator of overlapped pathways shows large numbers of overlapping pathways between rat hepatocytes and other assay systems, but the POPs between rat *in vivo* and human *in vitro* was low, indicating that when both the species and assay system differ, extrapolation can be challenging.

Various *in vitro* and *in vivo* testing strategies have been developed to assess drug sensitivity and toxicity, and within a given system a wide range of endpoints can be studied. For example, *in vitro* testing strategies may rely on immortalized cell lines, primary cell cultures, or more physiologically relevant cell environments (Tice et al., 2013; Michelini et al., 2010). Additionally, cell lines can be grown in three-dimensional (3D) models and cells may be derived from susceptible individuals to mimic idiosyncratic responses. The diversity of choice creates a challenge as to whether the different assay results yield comparable assessments, especially on how they rank drugs with respect to efficacy or toxicity (Haibe-Kains et al., 2013; Garnett et al., 2012; Barretina et al., 2012).

In the pair ranking analysis (Prank) we refer to "similar" compounds that imply similar toxicity profiles based on toxicogenomic data and the phrase "a highly similar pair of compounds" must be seen in the context of a comparison against other pairs. If the pairwise similarity of two compounds is consistently ranked at the top by various assays among all other pairs of compounds compared, these two compounds are highly likely similar. The same concept can be used to assess the similarity of any two assays where if two assays produce the same ranking resolution, we consider them interchangeable. Therefore, the PRank method compares any two assay systems by their preservation of the ranking of transcriptomic profiles perturbed by compounds.

The read-across concept assumes that compounds with similar chemical structure have similar biological activity or toxicity (Zhu et al., 2016). However, this approach is not always robust as illustrated by ibuprofen and ibufenac. These two drugs are NSAIDs with very similar chemical structures (only one methyl group difference) but ibufenac was withdrawn from the market due to severe DILI whereas ibuprofen continues as one of the most popular over-the-counter pain relief medications. Three TGx assay systems were used to carry out a comparison between the chemical similarity and outcome (Fig. 3). The PRank

score between chemical space and TGx space was less than 0.6 for all three TGx testing systems, which implied that there was no concordance between chemical space and TGx space.

We further investigated the ability of the TGx systems to detect compounds with high chemical similarity (Tanimoto correlation coefficient cut-off = 0.4, which is close to similarity value of the top one percentile of total pairs, see Fig. S5[1]) (Baldi and Benz, 2008). There were 48 drug pairs with chemical similarity greater than 0.4. We further compared the similarity observed to the similarity in different TGx space, as listed in Tab. S3[1]. There are 20.8%, 12.5% and 8.3% compound pairs with similarity more than 0.4 in the InVitro_Rat, InVitro_Human, and InVivo_Rat, respectively. In addition, 33.3% and 39.6% compound pairs with chemical similarity more than 0.4 have the same therapeutic uses and DILI concern, respectively. It was suggested that the combination of chemical structure and TGx may be a better way to conduct a safety assessment (Low et al., 2011). For example, chlorpheniramine and ticlopidine have a chemical similarity of 0.45 but the therapeutic uses and DILI concern are quite different for the two compounds. The TGx systems show that the two compounds are quite different with similarity less than 0.2, demonstrating that the difference between the compounds could be distinguished in the TGx assay systems.

The IVIVE between the assay systems was endpoint dependent and we found lipid metabolism to be consistently perturbed among the three testing strategies (rat or human *in vitro*, rat *in vivo*). For instance, a high similarity was observed among the PPARα agonists (fenofibrate, clofibrate and WY-14643), which influence lipid metabolism (Tab. S4[1]). In our previous study, we also demonstrated that these three compounds clustered together using topic modeling with network approaches (Lee et al., 2014).

There are several caveats to the current study. Firstly, pharmacokinetics and pharmacodynamics (PK/PD) are not considered within the proposed IVIVE assessment of toxicogenomic data. The PK/PD properties of a molecule influence the concentration that ultimately reaches the cell, and are essential in interpreting *in vitro* data in the context of drug potency (Groothuis et al., 2015; Kramer et al., 2015). In the current PRank methodology, we evaluated IVIVE potential based on the gene activity among different assay systems, which does not take into account the actual concentration reaching the cell. In future studies, PK/PD properties will be factored in to enhance interpretation of the data and further improve the assessment. Secondly, there are no negative controls in the Open TG-GATEs database for the different DILI related endpoints; this limits the assessment of the methodology for discrimination of IVIVE potential. Furthermore, in future, genetic elements such as miRNA and long non-coding RNAs implicated in regulating specific toxicological or biological processes could be integrated into our proposed PRank methodology.

In the current drug discovery paradigm, animal testing systems are still considered the standard way to assess drug safety and detect potential safety concerns. As emerging techniques develop, it is the key to understand how to apply these techniques in drug discovery through an improved understanding of the IVIVE potential of different techniques such as toxicogenomics, high throughput screening assays and other testing systems. The data presented suggest that PRank methodology offers a promising approach to assess transferability between the testing systems.

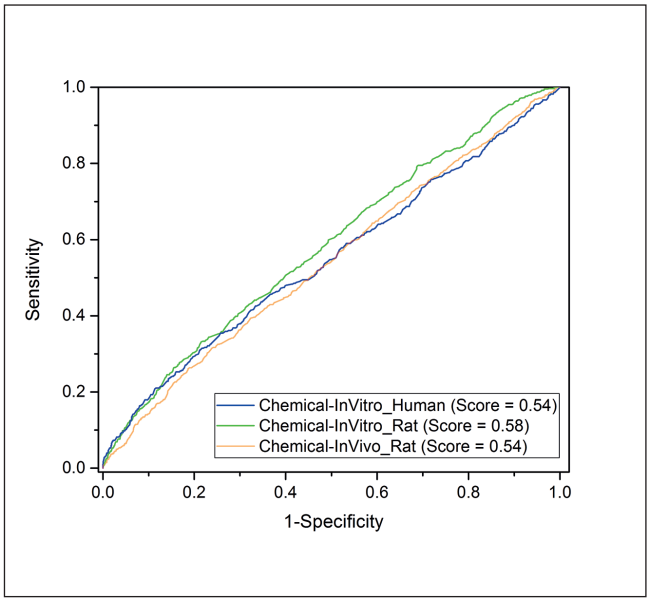

Fig. 3: The concordance between three different toxico-genomic assay systems and chemical space using the PRank method

References

Aardema, M. J. and MacGregor, J. T. (2002). Toxicology and genetic toxicology in the new era of "toxicogenomics": Impact of "-omics" technologies. *Mutat Res 499*, 13-25. doi:10.1016/S0027-5107(01)00292-5

Abbott, A. (2005). More than a cosmetic change. *Nature 438*, 144-146. doi:10.1038/438144a

Bailey, J., Thew, M. and Balls, M. (2013). An analysis of the use of dogs in predicting human toxicology and drug safety. *Altern Lab Anim 41*, 335-350.

Baldi, P. and Benz, R. W. (2008). BLASTing small molecules – statistics and extreme statistics of chemical similarity scores. *Bioinformatics 24*, i357-i365. doi:10.1093/bioinformatics/btn187

Barretina, J., Caponigro, G., Stransky, N. et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature 483*, 603-307. doi:10.1038/nature11003

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

Boess, F., Kamber, M., Romer, S. et al. (2003). Gene expression

in two hepatic cell lines, cultured primary hepatocytes, and liver slices compared to the in vivo liver gene expression in rats: Possible implications for toxicogenomics use of in vitro systems. *Toxicol Sci 73*, 386-402. doi:10.1093/toxsci/kfg064

Chen, M., Vijay, V., Shi, Q. et al. (2011). FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov Today 16*, 697-703. doi:10.1016/j.drudis.2011.05.007

Chen, M., Zhang, M., Borlak, J. and Tong, W. et al. (2012). A decade of toxicogenomic research and its contribution to toxicological science. *Toxicol Sci 130*, 217-228. doi:10.1093/toxsci/kfs223

Cheng, F., Theodorescu, D., Schulman, I. G. et al. (2011). In vitro transcriptomic prediction of hepatotoxicity for early drug discovery. *J Theor Biol 290*, 27-36. doi:10.1016/j.jtbi.2011.08.009

Chung, M. H., Wang, Y. P., Tang, H. L. et al. (2015). Asymmetric author-topic model for knowledge discovering of big data in toxicogenomics. *Front Pharmacol 6*, 81. doi:10.3389/fphar.2015.00081

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

DeJongh, J., Nordin-Andersson, M., Ploeger, B. A. and Forsby, A. (1999). Estimation of systemic toxicity of acrylamide by integration of in vitro toxicity data with kinetic simulations. *Toxicol Appl Pharmacol 158*, 261-268. doi:10.1006/taap.1999.8670

Deng, Y. P., Johnson, D. R., Guan, X. et al. (2010). In vitro gene regulatory networks predict in vivo function of liver. *BMC Syst Biol 4*, 18. doi:10.1186/1752-0509-4-153

Dix, D. J., Houck, K. A., Martin, M. T. et al. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci 95*, 5-12. doi:10.1093/toxsci/kfl103

Ganter, B., Tugendreich, S., Pearson, C. I. 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

Garnett, M. J., Edelman, E. J., Heidorn, S. J. et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature 483*, 570-575. doi:10.1038/nature11005

Groothuis, F. A., Heringa, M. B., Nicol, B. et al. (2015). Dose metric considerations in in vitro assays to improve quantitative in vitro-in vivo dose extrapolations. *Toxicology 332*, 30-40. doi:10.1016/j.tox.2013.08.012

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J. et al. (2013). Inconsistency in large pharmacogenomic studies. *Nature 504*, 389-393. doi:10.1038/nature12831

Hamburg, M. A. (2011). Advancing regulatory science. *Science 331*, 987-987. doi:10.1126/science.1204432

Hardt, C., Beber, M. E., Rasche, A. et al. (2016). ToxDB: Pathway-level interpretation of drug-treatment data. *Database 2016*, baw052. doi:10.1093/database/baw052

Hassan, M., Brown, R. D., Varma-O'Brien, S. and Rogers, D. (2006). Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers 10*, 283-299. doi:10.1007/s11030-006-9041-5

Hochreiter, S., Clevert, D.-A. and Obermayer, K. (2006). A new summarization method for affymetrix probe level data. *Bioinformatics 22,* 943-949. doi:10.1093/bioinformatics/btl033

Hrach, J., Mueller, S. O. and Hewitt, P. (2011). Development of an in vitro liver toxicity prediction model based on longer term primary rat hepatocyte culture. *Toxicol Lett 206*, 189-196. doi:10.1016/j.toxlet.2011.07.012

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protocols 4,* 44-57. doi:10.1038/nprot.2008.211

Hubbell, E., Liu, W.-M. and Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics 18*, 1585-1592. doi:10.1093/bioinformatics/18.12.1585

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

Iorio, F., Bosotti, R., Scacheri, E. et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci 107*, 14621-14626. doi:10.1073/pnas.1000138107

Ippolito, D. L., AbdulHameed, M. D. M., Tawa, G. J. et al. (2016). Gene expression patterns associated with histopathology in toxic liver fibrosis. *Toxicol Sci 149*, 67-88. doi:10.1093/toxsci/kfv214

Knight, A. (2007). Systematic reviews of animal experiments demonstrate poor human clinical and toxicological utility. *Altern Lab Anim 35*, 641-659.

Kramer, N. I., Di Consiglio, E., Blaauboer, B. J. et al. (2015). Biokinetics in repeated-dosing in vitro drug toxicity studies. *Toxicol In Vitro 30*, 217-224. doi:doi:10.1016/j.tiv.2015.09.005

Kuhn, M., Campillos, M., Letunic, I. et al. (2010). A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol 6*, 343. doi:10.1038/msb.2009.98

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

Lee, M., Liu, Z. C., Kelly, R. and Tong, W. (2014). Of text and gene – using text mining methods to uncover hidden knowledge in toxicogenomics. *BMC Syst Biol 8*, 93. doi:10.1186/s12918-014-0093-3

Lee, M., Huang, R. and Tong, W. (2016). Discovery of transcriptional targets regulated by nuclear receptors using a probabilistic graphical model. *Toxicol Sci 150*, 64-73. doi:10.1093/toxsci/kfv261

Liu, Z., Kelly, R., Fang, H. et al. (2011). Comparative analysis of predictive models for nongenotoxic hepatocarcinogenicity using both toxicogenomics and quantitative structure-activity relationships. *Chem Res Toxicol 24*, 1062-1070. doi:10.1021/tx2000637

Liu, Z., Wang, Y., Borlak, J. and Tong, W. (2016). Mechanis-

tically linked serum miRNAs distinguish between drug induced and fatty liver disease of different grades. *Sci Rep 6*, 23709. doi:10.1038/srep23709

Low, Y., Uehara, T., Minowa, Y. et al. (2011). Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chem Res Toxicol 24*, 1251-1262. doi:10.1021/tx200148a

Michelini, E., Cevenini, L., Mezzanotte, L. et al. (2010). Cell-based assays: Fuelling drug discovery. *Anal Bioanal Chem 398*, 227-238. doi:10.1007/s00216-010-3933-z

Nuwaysir, E. F., Bittner, M., Trent, J. et al. (1999). Microarrays and toxicology: The advent of toxicogenomics. *Mol Carcinog 24*, 153-159. doi:10.1002/(sici)1098-2744(199903)24:3<153::aid-mc1>3.0.co;2-p

Olson, H., Betton, G., Robinson, D. et al. (2000). Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol 32*, 56-67. doi:10.1006/rtph.2000.1399

Otava, M., Shkedy, Z., Talloen, W. et al. (2015). Identification of in vitro and in vivo disconnects using transcriptomic data. *BMC Genomics 16*, 1-10. doi:10.1186/s12864-015-1726-7

Pepper, S., Saunders, E., Edwards, L. et al. (2007). The utility of MAS5 expression summary and detection call algorithms. BMC *Bioinformatics 8*, 273. doi:10.1186/1471-2105-8-273

Pettit, S., des Etages, S. A., Mylecraine, L. et al. (2010). Current and future applications of toxicogenomics: Results summary of a survey from the HESI genomics state of science subcommittee. *Environ Health Perspect 118*, 992-997. doi:10.1289/ehp.0901501

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

Sakatis, M. Z., Reese, M. J., Harrell, A. W. et al. (2012). Preclinical strategy to reduce clinical hepatotoxicity using in vitro bioactivation data for >200 compounds. *Chem Res Toxicol 25*, 2067-2082. doi:10.1021/tx300075j

Suter, L., Schroeder, S., Meyer, K. et al. (2011). EU Framework 6 Project: Predictive toxicology (PredTox)-overview and outcome. *Toxicol Appl Pharmacol 252*, 73-84. doi:10.1016/j.taap.2010.10.008

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

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 Perspect 121*, 756-765. doi:10.1289/ehp.1205784

Uehara, T., Ono, A., Maruyama, T. et al. (2010). The Japanese toxicogenomics project: Application of toxicogenomics. *Mol Nutr Food Res 54*, 218-227. doi:10.1002/mnfr.200900169

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

Xu, J. J., Henstock, P. V., Dunn, M. C. et al. (2008). Cellular imaging predictions of clinical drug-induced liver injury. *Toxicol Sci 105*, 97-105. doi:10.1093/toxsci/kfn109

Zhu, H., Bouhifd, M., Kleinstreuer, N. et al. (2016). t^4 report: Supporting read-across using biological data. *ALTEX 33*, 167. doi:10.14573/altex.1601252

Conflict of interest

Ruth Roberts is co-founder and co-director of Apconix, an integrated toxicology and ion channel company that provides expert advice on nonclinical aspects of drug discovery and drug development to academia, industry and not-for-profit organisations.

Acknowledgements

We specially thank Dr John Senior for categorizing DILI patterns and also thank Dr Takeki Uehara, Dr Ikuo Kato and the TGP group for helpful discussions and for their expert opinion on histopathology of the livers after treatment of rats with steatosis/phospholipidosis causing drugs.

Correspondence to

Zhichao Liu, PhD
National Center for Toxicological Research (NCTR)
U.S. Food and Drug Administration
3900 NCTR Road, HFT-020
Jefferson, AR 72079, USA
Phone: +1 870 543 7909
Fax: +1 870 543 7662
e-mail: zhichao.liu@fda.hhs.gov

Weida Tong, PhD
National Center for Toxicological Research (NCTR)
U.S. Food and Drug Administration
3900 NCTR Road, HFT-020
Jefferson, AR 72079, USA
Phone: +1 870 543 7142
Fax: +1 870 543 7662
e-mail: weida.tong@fda.hhs.gov