PharmOmics: A Species- and Tissue-specific Drug Signature Database and Online Tool for Toxicity Prediction and Drug Repurposing

Yen-Wei Chen[1,2,3], Douglas Arneson[2,4], Graciel Diamente[2], Jennifer Garcia[2], Nima Zaghari[2], Paul Patel[2], Patrick Allard[1,3], and Xia Yang[1,2,4,5,6]

[1]Interdepartmental Program of Molecular Toxicology, University of California, Los Angeles, Los Angeles, CA

[2]Department of Integrative Biology and Physiology, University of California, Los Angeles, CA

[3]Institute for Society and Genetics, University of California, Los Angeles, Los Angeles, CA

[4] Interdepartmental Program of Bioinformatics, University of California, Los Angeles, Los Angeles, CA

[5]Institute for Quantitative and Computational Biosciences, University of California, Los Angeles, CA

[6]Molecular Biology Institute, University of California, Los Angeles, CA

Correspondence:

Xia Yang, Ph.D.

Dept. Integrative Biology and Physiology, UCLA

Email: xyang123@ucla.edu

Abstract

Drug development has been hampered by a high failure rate in clinical trials due to suboptimal efficacy or safety issues that are not predicted by preclinical studies in model systems. A key contributor to this hurdle is our incomplete understanding of how drug molecules function across organ systems and species. Therefore, elucidating species- and tissue-specific molecular actions of drugs can provide systems level insights into therapeutic efficacy, toxicity, adverse drug reactions (ADRs), and between-species differences that are necessary for more effective translational medicine. Here, we present a comprehensive drug knowledgebase and analytical tool, PharmOmics, comprised of genomic footprints of drugs in individual tissues from multiple species (human, mouse, and rat) based on meta-analysis of global transcriptome data from public repositories (GEO, ArrayExpress, TG-GATEs, drugMatrix). Using multi-species multi-tissue gene expression signatures as molecular indicators of drug functions, we demonstrate the potential of PharmOmics to predict drugs that can be used for new disease indications or that may induce ADRs. PharmOmics was able to accurately predict known drugs for multiple diseases and identify potential drugs targeting non-alcoholic fatty liver disease for which no approved drugs are currently available. In addition, we found that drugs causing similar ADRs tend to have gene signatures with close proximity in gene regulatory networks. Comparison of our PharmOmics platform with other commonly used drug signature databases, most of which are based on data derived from *in vitro* cell lines, demonstrates that our tissue- and species-specific drug signatures serve as a complementary resource to support drug characterization.

Key words: PharmOmics, drug gene signatures, tissue specificity, cross-species comparison, drug repositioning, toxicity, adverse drug reactions, gene regulatory networks

Background

Drug development has been challenging and costly over the past decades due to the continuously high failure rate in clinical trials (1). Most drugs with excellent efficacy and safety profiles in preclinical studies often encounter suboptimal efficacy or safety concerns in humans (2). This translational gap is likely attributable to our incomplete understanding of the systems level activities of drugs in individual tissues and organ systems (3) as well as the differences between humans and model systems (4).

Drug activities can be captured by gene expression patterns, commonly referred to as gene signatures. By measuring how a pharmacological agent affects the gene signature of a cell or a tissue type in a particular species, we can infer the cell- or tissue-specific biological pathways involved in therapeutic processes or toxicological responses. This concept has prompted drug repositioning studies and provided important predictions for repurposing approved drugs for new disease indications (5–10). Similarly, gene signatures can reveal mechanisms underlying adverse drug reactions (ADRs) and can therefore be leveraged to predict these ADRs as previously shown for liver and kidney toxicity (11–13).

A drug may affect different molecular processes between tissues, causing disease treatment effects in the desired target tissue(s) but toxicity or ADRs through off-target effects in other tissues. Therefore, tissue-specific drug treatment signatures will offer more systematic

understanding of drug actions *in vivo*. Additionally, rodent models have been commonly used in toxicology and preclinical studies, yet species-specific effects of drugs has been well described (14) and underlie the lack of efficacy or unexpected ADRs or toxicity responses of certain drugs when used in humans (15). Therefore, understanding the species-specific molecular effects of drugs is also of high biological importance. A detailed species- and tissue-specific drug genomic signature database will significantly improve our understanding of the molecular networks affected by drugs and facilitate network-based drug discovery and ADR prediction for translational medicine.

The tremendous potential of using gene signatures to facilitate target and toxicity identification has led to several major efforts in characterizing genomic signatures related to drug treatment (8,16–18). However, none of the existing platforms offer comprehensive cross-tissue and cross-species *in vivo* assessments of drug activities to allow predictions of the effects of a drug on individual tissues and to help assess the translational potential of a drug based on consistencies or discrepancies between species. For instance, the comparative toxicogenomics database (CTD), a literature-based resource curating chemical-to-gene/protein associations as well as chemical-to-disease and gene/protein-to-disease connections (16), lacks the cellular and tissue context of the curated interactions. More systematic, data-driven databases like CMAP (8) and LINC1000 (17) focus on characterizing and cataloging the genomic footprints of more than ten thousand chemicals using *in vitro* cell lines (primarily cancer cell lines) to offer global views of molecular responses to drugs in individual cellular systems. However, these *in vitro* cell-line based gene signatures may not capture *in vivo* tissue-specificity of drug activities. To move into the *in vivo* systems, large

toxicogenomics databases like TG-GATEs (19) and DrugMatrix from the National Toxicology Programs of the National Institute of Environmental Health Sciences (https://ntp.niehs.nih.gov/drugmatrix/index.html) have become available to provide unbiased transcriptome assessment for heart, muscle, liver, and kidney systems. However, information about other organ systems is limited. To broaden the tissue coverage, efforts to meta-analyze publicly deposited transcriptomic datasets in GEO (20) and ArrayExpress (21) have also been described (18), but the species- and tissue-specific effects of drugs have not been addressed.

Here, we establish a computational pipeline to meta-analyze 1,314 drug-related transcriptomic datasets from mouse, human, and rat studies curated in GEO, ArrayExpress, TG-GATEs and DrugMatrix, covering >20 tissue/organ systems, and summarize the results into tissue- and species-specific drug treatment signatures. In addition, we demonstrate the utility and performance of these signatures for gene network-based drug repositioning and ADR assessment. The tissue- and species-specific drug signatures are hosted on an interactive web server PharmOmics, to enable public access, download, and integrative analyses for drug repositioning (http://mergeomics.research.idre.ucla.edu/PharmOmics/).

Construction and content

Curation of tissue- and species-specific drug transcriptomic datasets

As illustrated in **Figure 1**, a list of clinically relevant target drugs were first compiled from Kyoto Encyclopedia of Genes and Genomes (KEGG) (16), followed by transcriptomic dataset search in GEO (20), ArrayExpress (21), TG-GATEs (19) and DrugMatrix (https://ntp.niehs.nih.gov/drugmatrix/index.html) based on matching generic names in either

the title or abstract fields of the dataset records. GEO and ArrayExpress datasets were identified as of January 2018. Duplicated datasets between repositories were removed. We developed a semi-automated pipeline combining automated search with manual checking to identify relevant datasets for drug treatment. The automated process first extracts datasets containing drug generic names or abbreviations and then inspects the potential datasets for availability of both drug treatment and control labels in the constituent samples. We also manually checked the recorded labels identified by the automated process to validate the labels to decrease potential false detections. Only datasets with n>=3/group in both drug treatment and control groups were included in our downstream analyses. We combined drug treatment samples regardless of dose and regimen for the same drug in order to obtain the consensus responses across conditions. The decision on not further stratifying the studies based on dose and regimen was to increase the generalizability of the drug signatures and to mitigate the reduced statistical power due to uneven coverage of dose and regimen information across data repositories and drugs.

Obtaining drug treatment signatures stratified by species and tissues

Based on the metadata of each dataset, we retrieved the species and tissue labels. Tissue names were standardized based on the Brenda Tissue Ontology (22). We implemented a search procedure to climb the ontology tree structure to determine the suitable tissue annotations. We first defined a tissue/organ priority list in the Brenda Tissue Ontology tree system based on the widely used organ systems in toxicological research. Tissue/organ priority order was set to "kidney", "liver", "pancreas", "breast", "ovary", "adipose tissue", "cardiovascular system", "nervous system", "respiratory system", "urogenital system",

"immune system", "hematopoietic system", "skeletal system", "integument" (endothelial and skin tissue), "connective tissue", "muscular system", "gland", "gastrointestinal system", and "viscus" (other non-classified tissue). Tissue terms relevant to each of these tissues or organs were curated from the ontology tree into a tissue/organ ontology table. Next, in our tissue/organ ontology table, we looked up the terms in the Cell/Organism/Tissue column of the metadata in each transcriptomic dataset. If not found, we searched the title and summary columns of the metadata as well to retrieve additional information. When searching returned multiple tissue terms, we used the one with the higher priority order as described above. Manual checking was conducted to confirm the tissue annotation for each dataset.

For each gene expression dataset, normalized data were retrieved, followed by assessment of quantile distribution of values and, when data was not normally distributed, log2-transformation using GEO2R (20). Due to the heterogeneity of the datasets in terms of study design, platforms, sample size, and normalization methods, we chose to use the characteristic direction method in the GeoDE package (23) to identify genes affected by drug treatment. GeoDE was designed to accommodate datasets from different settings and outputs differences between treatment and control groups using a “characteristic direction” measure analogous but superior to fold change in identifying biologically relevant genes and pathways. The normalized characteristic directions for all genes were then transformed into a non-parametric rank representation. Subsequently, gene ranks of a particular drug from the same tissue/organ system and the same organism were aggregated across datasets using the Robust Rank Aggregation method (24) to identify differentially expressed genes (DEGs) of the drug within each tissue for each species. Robust Rank Aggregation provides a non-parametric

meta-analysis across different ranked lists to obtain commonly shared genes across datasets, which avoids statistical issues associated with heterogeneous datasets. It computes a null distribution based on randomized gene ranks and then compares the null distribution with the empirical gene ranks to obtain a p-value for each gene. The robust rank aggregation process was done for the upregulated and downregulated genes separately to obtain DEGs for both directions under Bonferroni-adjusted p-value <0.01, a cutoff implemented in the Robust Rank Aggregation algorithm. To obtain species-level signatures for each drug, we further aggregated DEGs across different organs tested for each drug within each species. Pathway analysis of individual drug signatures was conducted using enrichr (25) by intersecting each signature with pathways or gene sets from KEGG (26), gene ontology biological pathways (GOBP) (27), Reactome pathways (28), Biocarta pathways (29) and Wikipathways (30). Pathways at false discovery rate (FDR) <0.05 were considered significant. For cross-tissue and species comparison described in this article we used all five pathway databases to make it more comprehensive.

As summarized in **Figure 2,** in total we identified 1,314 transcriptomic datasets across 551 drugs from GEO, ArrayExpress, DrugMatrix, and TG-GATEs. These datasets cover >20 tissue or organ systems in rats, human, and mouse and 289 KEGG drug classes. The toxicogenomics databases TG-GATEs and DrugMatrix mainly deposited rat liver and kidney datasets, while public data repositories GEO and ArrayExpress contained datasets with broader tissue and species coverage (**Figure 2A**). More liver and kidney datasets are from rats whereas more human and mouse datasets are available for the other tissues (**Figure 2B**). There is also a species bias between the data repositories, with GEO covering more mouse

and human datasets whereas DrugMatrix and TG-GATEs curate more rat datasets (**Figure 2C**) (http://mergeomics.research.idre.ucla.edu/PharmOmics/DrugSignatures/).

Comparison of PharmOmics with existing drug signature platforms revealed both shared and unique properties between platforms

We compared PharmOmics with the existing CREEDS (18) and L1000FWD (31) databases to assess the degree of agreement in drug signatures across different platforms **(Supplementary methods)**. As shown in **Supplementary Figure 1**, we found partial agreement among the three databases. Despite the fact that overlaps in gene signatures between databases were statistically significant for a majority of the drugs, the Jaccard scores for signature overlaps were low (mostly between 0.02 and 0.04). We were not able to systematically compare PharmOmics with CMAP for all drugs as done for CREEDS and L1000FWD, because CMAP is an online query system (32) only feasible to query select drugs. We searched atorvastatin in CMAP as an example and found that our PharmOmics signature of atorvastatin from primary hepatocytes matched well with CMAP signatures of the same drug from several human cancer cell lines including MCF7 (breast), HT29 (colon) and VCAP (prostate), but surprisingly, not with HepG2 (hepatocellular carcinoma) which is supposed to better match hepatocytes (**Supplementary Figure 2A**). The *in vivo* and *in vitro* atorvastatin rat liver signatures from PharmOmics also showed different matching patterns with CMAP signatures from the human cancer cell lines (**supplementary Figure 2B**). Taken together, our results highlight important differences between PharmOmics and major existing platforms. The differences between *in vivo* and *in vitro* drug signatures support the value of obtaining more physiologically relevant, species-specific and tissue-specific drug signatures.

Tissue-specificity of drug signatures

To examine tissue-specificity of drug signatures, we analyzed the overall overlaps in gene signatures for each drug across tissues and. Results were visualized with UpSetR package (33). shown in **Figure 3A**, the overlap rate in the DEGs of the same drug between tissues and organs is usually less than 5%, indicating a high variability in DEGs between tissues. However, shared genes across tissues at DEG and pathway levels also exist. As an example, we examined the HMGCR (β-Hydroxy β-methylglutaryl-CoA receptor) inhibitor atorvastatin since its mechanism is well known and atorvastatin has been broadly tested in different tissues under the human species label. We found that two DEGs (*TSC22D3*, *THBS1*) were shared across tissues (**Figure 3B**). These genes are involved in extracellular matrix and inflammation, suggesting these processes as common targets of atorvastatin across tissues. Among the pathways shared across tissues, immune related pathways were shared between blood cells and liver cells but not seen in prostate cells from the urogenital system (**Figure 3C, Supplementary Table 2)**. Despite these overlaps, we found that the majority of the DEGs were tissue-specific. Pathway analysis indicated that steroid synthesis and drug metabolism pathways were altered mostly in liver, as expected, since the known target of statin drugs is HMGCR, the rate limiting enzyme in cholesterol biosynthesis in liver. Blood monocyte DEGs indicate changes in inflammation related pathways, while GPCR ligand binding proteins were altered in prostate cancer cells. The tissue-specificity of drug signatures revealed through our analysis supports tissue-specific therapeutic response or side effects and the need for comprehensive maps of drug signatures across tissue systems as implemented in the PharmOmics framework.

Species-specificity of drug signatures

We also compared gene signatures for each drug between species and found evidence for species-specificity. Results were visualized with UpSetR package (33). As shown in **Figure 3D,** the pair-wise overlaps in DEGs between species for the same drug is generally lower than 5%. Here we chose PPAR gamma receptor agonist rosiglitazone as an example because this drug has datasets across human, rat, and mouse in PharmOmics and its mode of actions is well-studied. As shown in **Figure 3E** and **3F**, nine genes (*CPT1C, AKR1B1, VNN1, ACSM3, CD36, CPT1A, PDK4, ZNF669, ADH1C*) and several pathways including PPAR signaling and fatty acid, triacylglycerol, and ketone body metabolism are consistently identified from liver DEGs across species (**Supplementary Table 3**), supporting these as the major shared pharmacological effectors of rosiglitazone across species. Among the species-specific findings were rat-specific changes in bile acid and retinol metabolism, and adipocytokine pathways in human datasets. The differences revealed between species highlight the importance of understanding the physiological differences among model systems to facilitate drug design with better translational potential. Our cross-species comparative studies also highlight the urgent need to investigate drugs in multiple species, as only 21% of the drug-tissue pairs (237 out of 1140) have data from two or more species.

Implementation of a web server to enable public access to PharmOmics

Previously we have developed an Apache web server Mergeomics (34), which allows integrative analysis of multi-omics datasets to derive disease-associated pathways, networks, and network regulators (http://mergeomics.research.idre.ucla.edu/). We implemented the

interactive PharmOmics web server with Rshiny (35) and deployed it on the same server which we use to host Mergeomics, by enabling an instance of Apache to act as a front-end proxy to the PharmOmics Shiny Server (http://mergeomics.research.idre.ucla.edu/PharmOmics/). The PharmOmics web server hosts the following main functions: i) drug signature query, visualization and download (http://mergeomics.research.idre.ucla.edu/PharmOmics/DrugSignatures/), ii) gene network-based drug repositioning using our species- and tissue-stratified signatures (http://mergeomics.research.idre.ucla.edu/PharmOmics/Network_drug_repositioning/). iii) Jaccard score-based drug repositioning (http://mergeomics.research.idre.ucla.edu/PharmOmics/Jaccard_repositioning/) iv). Potential ADR related signature based on meta-analyzing drugs sharing same ADR (http://mergeomics.research.idre.ucla.edu/PharmOmics/ADRsignatures/). Due to higher computational cost in the network-based drug repositioning application, a downloadable shiny version is available at http://mergeomics.research.idre.ucla.edu/Download/Network_drug_repositioning/.

Applications and Discussions

The value of PharmOmics to extract drug targets and target pathways

To explore the potential applications of our PharmOmics drug signatures, we evaluated whether the species- and tissue-specific DEGs have the potential to retrieve drug targets and target pathways. To this end, we first retrieved known drug targets for the drugs included in PharmOmics from the DrugBanks database (36). Three different methods were used to evaluate the potential of DEGs for drug target identification. The first method was based on

direct overlaps between DEGs and known drug targets, that is, if a known drug target was directly retrieved from the DEG list it would be considered a positive hit. The second method was based on overlaps in pathways, that is, whether the known pathway containing the drug target was retrieved by the pathway analysis of the drug DEGs. The last method was based on whether a known drug target is within the close neighborhood of drug DEGs in molecular networks, including the STRING network (37) and tissue-specific Bayesian networks (BNs) **(Supplementary methods)**.

As shown in **Table 1**, we found that all three drug target retrieval methods (gene overlap, pathway overlap, network overlap) support that drug DEGs are more likely to detect true drug targets than random genes. Compared to the gene overlap method which can retrieve known drug targets at a 22.1% rate, the pathway- and network-based methods reached higher target detection rates (62% for pathways, 40% for protein interaction network, 72% for liver BN with matching liver DEGs, and 40% for kidney BN with mismatching liver DEGs), although false positives also increased in the pathway and network-based approaches. Notably, matching the tissue between DEGs and networks improved the target detection rate. These results suggest that while known drug targets may not be directly detected among the DEGs, pathway relatedness and network proximity in tissue specific network can be informative of drug targets and can enhance the chance of target identification from drug DEGs.

Drug repositioning analyses support that PharmOmics drug signatures can retrieve known therapeutic drugs for various diseases

We hypothesized that if a drug is useful for treating a disease, the drug signatures and the disease signatures will likely target similar pathways and therefore will be highly overlapping or tightly connected in gene networks. To this end, we evaluated the ability of PharmOmics to identify drugs for diseases based on overlaps in gene signatures matched by tissue using the gene overlap-based measurement (Jaccard score) as well as a network proximity measurement between drug and diseases genes which was previously applied to protein interaction networks (5) **(Supplementary methods)**. Here, we used tissue-specific BN networks by testing the mean shortest distance between network hubs of the drug DEGs and disease genes. We further applied Mergeomics pipeline (38) which identifies hub genes on network (key drivers) in this method. Using key drivers of drug gene signatures can reduce noise and improve computing speed (**Supplementary Figure 3**) while performing similarly to drug DEGs (**Figure 4A** and **4B**).

We selected hyperlipidemia as a disease test case, as multiple known drugs are available as positive controls to evaluate the performance of drug repositioning. We have previously developed a method Mergeomics (38), which can accurately extract disease related genes, pathways, and networks (39,40). Because hyperlipidemia is most relevant to LDL and the liver tissue, we used Mergeomics to retrieve LDL causal genes and pathways in the liver tissue based on LDL GWAS and liver genetic regulation of gene expression (**Supplementary methods**).

The Jaccard score method revealed that our PharmOmics drug signatures were able to differentiate anti-hyperlipidemia drugs from other drugs with 79% AUC when using *in vivo*

rat liver signatures (Wilcoxon signed rank test p value=0.001), 56% AUC when using *in vitro* liver signatures (non-significant), and 74% AUC when using the aggregated rat liver signatures from both *in vivo* and *in vitro* settings (referred as combined, p=0.005) (**Figure 4C**). Among the top 5 predicted drugs are known LDL-lowering drug fluvastatin and other drugs that may affect LDL including loratadine, sertraline, gefitinib, and paroxetine.

The network-based approach performed similarly as the direct gene overlap method above (**Figure 4D**). The top 5 drugs predicted for hyperlipidemia based on the network approach included fluvastatin, loratadine, mestranol-lynestrenol mixture, sertraline and estriol. Between the network and Jaccard Score methods, 3 of the top 5 drugs from the two methods overlapped, which are fluvastatin, loratadine and sertraline. Of these, fluvastatin is a known HMG-CoA inhibitor for antihyperlipidemia therapy while oral contraceptives and sertraline were linked with hyperlipidemia ADR (41–43). Moreover, among the 53 significant predictions (out of 263 drugs) at $z < -2.52$ ($p < 0.01$) using our PharmOmics drug signatures (**Supplementary Table 4**), we retrieved 6 out of the 10 approved anti-hyperlipidemia drugs in the PharmOmics knowledgebase.

Using alternative hyperlipidemia signature from the Comparative Toxicogenomics database (CTD) (16) to match with our PharmOmics drug signatures also showed that anti-hyperlipidemia drugs had better Jaccard scores and better network repositioning z-scores compared to other non-hyperlipidemia drugs. In addition, *in vivo* drug signatures again showed better repositioning performance than the *in vitro* and combined drug signatures in the CTD-based hyperlipidemia analysis (**Figure 4E**, **4F**).

We further tested the performance of PharmOmics in retrieving known drugs for other diseases. Using CTD disease signatures for diabetes and hyperuricemia, we found PharmOmics signatures of known diabetes drugs had better repositioning scores than those of non-diabetes drugs (**Supplementary Figure 4**), and retrieved Benzbromarone as the top ranked drug for hyperuricemia (**Supplementary Table 5**). These results support the general utility of PharmOmics for drug repositioning.

Better or comparable drug repositioning performance of PharmOmics compared to existing platforms

We compared the performance of our PharmOmics drug signatures in drug repositioning with that of the well-established L1000 signatures, which were mainly based on *in vitro* cell models. We found that PharmOmics signatures achieved better repositioning results than L1000 signatures aggregated across dose and regimen for both Mergeomics and CTD hyperlipidemia signatures using both the Jaccard score-based and network-based repositioning methods (**Figure 4C-E**). Similarly, PharmOmics signatures also performed better than the raw L1000 signatures when using the best performing dose and regimen **(Supplementary Figure 5).** When compared with CMAP, PharmOmics also showed better performance regardless of whether the aggregate signature across CMAP cell lines (**Supplementary Figure 6A, Supplementary Table 4**) or the HEPG2-based signature was used (**Supplementary Figure 6B, Supplementary Table 4**). Notably, PharmOmics was able to retrieve 6 out of 10 (60%) of the known anti-hyperlipidemia drugs curated in the PharmOmics database, while CMAP retrieved only 4 out of the 18 anti-hyperlipidemia drugs

examined in CMAP (22%). These results indicate the benefits of using *in vivo* signatures in PharmOmics for drug repositioning to complement the existing cell-line based drug repositioning tools to broaden drug repositioning capacity.

We further compared our PharmOmics gene signatures with the CREEDS database (18) for drug repositioning. Using only the best scoring signature in CREEDS to represent each drug, we found comparable drug-repositioning ability between the two databases (**Supplementary Figure 7)**. PharmOmics was able to reach faster repositioning since consensus signatures across studies stratified by species and tissues were used instead of signatures from individual datasets, and contained signatures for more drugs.

Use of PharmOmics to predict drugs for NAFLD

After establishing the performance of PharmOmics in drug repositioning using hyperlipidemia, diabetes and hyperuricemia as test cases, we also applied PharmOmics to predict potential drugs for NAFLD, for which there is currently no approved drugs. Using NAFLD gene signatures from published studies (40,44) and the CTD database (16), we predicted PPAR alpha agonists (fenofibrate and clofibrate) and PPAR gamma agonist rosiglitazone to be strong matches to the NAFLD signatures (**Supplementary Table 6**). PPAR agonists have been supported as potential drugs for NAFLD (45). The beneficial effects of several PPAR alpha agonists on NAFLD have been demonstrated in *in vivo* models (46,47), but clinical evidence is inconclusive at this point (48–50). On the other hand, rosiglitazone and other PPAR gamma agonists showed beneficial effects in both rodent models (50–52) and clinical trials (53–56). These lines of evidence substantiate the drug

repositioning potential of our PharmOmics tool. Additional drugs predicted include aspirin which was also reported beneficial in mitigating NAFLD (57) and numerous other drugs (**Supplementary Table 6**) which warrant experimental testing in future studies.

PharmOmics drug signatures can predict ADRs in liver

We hypothesize that drugs with the same ADR likely share common gene signatures driving the specific ADR events. In this case, gene signature overlaps are expected between an ADR term and the drugs with the ADR. We further hypothesized that network-based modeling can better characterize hidden gene signatures to improve ADR prediction, which can be tested based on network distance from drug signature genes to ADR genes. We collected ADR information and ADR gene signatures were first aggregated across drugs with the same ADR regardless of their pharmacological classes (**Supplementary methods**). We then tested whether the ADR signatures can better predict drugs with the ADR than those with permuted ADR labels, either based on direct gene overlaps or based on network distances (design illustrated in **Supplementary Figure 8A)**.

We focused on rat liver drug signatures as more than 200 drugs were available for this tissue. The other tissues had more limited coverage of drugs, preventing a comprehensive investigation. A total of 617 ADR terms each with more than six drugs were tested. For each ADR, aggregate gene signatures were obtained from all constituent drugs with the ADR. The gene overlap matrix between ADR signature and individual drug signatures, or the distance matrix between ADR signatures and the key drivers of drug signatures in liver Bayesian network were obtained. A random forest model was built with 10-fold-5-repeat cross

validation combined with smote algorithm (58) to accommodate unbalanced ADR samples (i.e., fewer drugs with an ADR than those without the ADR, **Supplementary Methods**). For the 617 ADR terms investigated, the distribution of the raw permuted p-values from random forest models based on the gene overlapping matrix was similar to that based on the network distance matrix (**Supplementary Figure 8B**, **8C**), both shifting towards the low p value end (Kolomogorov-Smirnov test p-value <2.2e-16 against uniform distribution)**.** However, the network-based method achieved better ADR predictions with lower FDR and higher accuracy, with 21 ADR terms passing FDR < 0.05 (compared to 0 ADR term based on the gene overlap matrix at the same threshold) and cross-validation accuracy > 0.99 (**Table 2**). Therefore, PharmOmics signatures can also be used to derive ADR signatures which can predict drugs with ADRs using a network-based approach. We were also able to retrieve top predictive genes for many ADR terms based on the random forest models and predicted additional drugs that may have these ADRs (**Table 2**).

Discussion

Here we present a publicly available tissue- and species-stratified drug signature database along with an online web server, PharmOmics. PharmOmics fully utilizes published drug-related transcriptomic datasets and provides unique tissue- and species-stratified gene signatures that are more reflective of *in vivo* activities of drugs. We also developed a framework for fast drug repositioning based on tissue-specific gene network models. We demonstrate that PharmOmics can be used for molecular characterization of drugs, comparisons of molecular activities between tissues and between species, and various

applications including target and target pathway identification, drug repurposing, and ADR prediction.

Compared to the well-established CMAP and LINC1000 platforms, PharmOmics focuses more on *in vivo* settings and likely captures more physiologically relevant drug signatures to improve drug repositioning performance. Compared to a previous crowdsourcing effort, CREEDS, which also utilizes publicly available drug datasets (18), our PharmOmics platform included more curated drugs (343 in PharmOmics vs. 271 drugs in CREEDS) and involved systematic tissue and species stratification and signature aggregation across different dosing and treatment regimens to facilitate drug repositioning.

The use of tissue annotation with Brenda Tissue Ontology in PharmOmics helps normalize organ labels and makes datasets more comparable. Although the tree structure in Brenda Tissue Ontology may assign a cell or tissue to multiple systems (for example, breast cancer cell line may be categorized as both epithelial and breast organ), we prioritized the tissue terms based on the relevance to toxicology to make tissue assignments unique for each dataset to reduce ambiguity. The unique tissue- and species-specific analyses implemented in PharmOmics allows for comprehensive molecular insight into the actions of drug molecules in individual tissues and species. Our results support that different species indeed have unique drug responses in addition to shared features, thereby implicating that drug responses and toxicity measures obtained in animal models may require caution before directly translating the insights into humans. This notion agrees with the long-observed high failure

rate of drug development that has primarily relied on preclinical animal models and argues for improved appreciation and understanding of inter-species differences in drug actions.

We demonstrate that PharmOmics can be used to retrieve drug targets and target pathways. In general, gene signatures of drugs reflect cascades of downstream events after drug administration. The initial drug target(s) may or may not be captured by drug DEGs due to the lack of dynamic information in the DEGs. For example, changes in the direct drug targets may have occurred transiently early upon treatment, but the change may have weakened or disappeared at the time of transcriptome profiling. Nevertheless, our analyses support that drug signatures are more likely to contain information about drug targets as well as target pathways and networks. These findings support the value of using drug signatures to reflect, at least partially, the molecular mechanisms of drugs.

Previous drug repositioning studies support that protein network-based approach is powerful for drug repositioning. Here we show that combining the drug transcriptomic signatures in PharmOmics, particularly the *in vivo* signatures, with tissue-specific gene regulatory Bayesian networks and gene signatures of diseases can retrieve known therapeutic drugs and predict potential new therapeutic avenues. Compared to other platforms, the use of tissue- and species drug signatures along with network biology is a unique strength of PharmOmics, which enables better performance and faster computation in repositioning.

We also demonstrate the potential of PharmOmics in ADR prediction. Although this effort is limited by the small numbers of existing ADR datasets at each tissue level, we were able to

show the value of incorporating drug signatures with gene network models to perform ADR predictions. The ADR signatures derived from our study and the predictive models from the machine learning-based methods can be informative of mechanisms involved in ADRs, although we acknowledge that further investigations are necessary to validate this potential.

There are several limitations in this study. First, our existing pipeline may not be able to identify all of the drug datasets from GEO and ArrayExpress database. Along the same line, the tissue coverage across the species is unbalanced, preventing a thorough cross-tissue and cross-species comparison of drug signatures. We will keep refining the pipeline and update our PharmOmics database periodically to include more datasets as they become available to increase the coverage of datasets and drug signatures across tissues and species. In addition, aggregating across different dosage and treatment durations may lose dose or time-specific drug responses. However, we believe our approach can capture the most consistent drug signatures regardless of treatment conditions, which are more representative of the general properties of the drugs. The comparable performance of our platform with that of CREEDS, which details dose and treatment regimens, in drug repurposing supports this notion. Lastly, more comprehensive applications and experimental evaluation of PharmOmics in drug repurposing for different disease areas and in ADR prediction are necessary in future studies.

Conclusion

We have established a new drug treatment signature database, PharmOmics, across different species and tissues, which comprehensively captures the systems level *in vivo* activities of drug molecules. In addition, we demonstrate the possible means to integrate these signatures

to address drug repositioning needs for disease treatment and to predict and characterize ADRs. PharmOmics can complement other available drug signature databases to accelerate drug development and toxicology research. The pipeline and our PharmOmics database will be updated periodically to include newly available datasets to increase the coverage of these datasets and drug signatures across tissues and species.

List of abbreviations

ADR	adverse drug reactions
CTD	comparative toxicogenomics database
KEGG	Kyoto Encyclopedia of Genes and Genomes
DEG	differential expressed genes
FDR	false discovery rate
wKDA	weighted key driver analysis
NAFLD	non-alcoholic fatty liver disease
LDL	low-density lipoprotein cholesterol
GWAS	genome-wide association study
BN	Bayesian gene regulatory network
ROC	Receiver operating characteristic
HMGCR	β-Hydroxy β-methylglutaryl-CoA receptor
PPAR	Peroxisome proliferator-activated receptor
GPCR	G-protein coupled receptor

Declaration

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Indexed dataset catalog, pre-computed gene signatures and pre-computed pathway enrichments for individual drugs are deposited to and accessible through the PharmOmics web server (http://mergeomics.research.idre.ucla.edu/PharmOmics/DrugSignatures/). We also implemented functions for same-tissue between-species comparison and same-species between-tissue comparison. Direct downloading the whole database or interested parts of the database is also enabled. In addition, network-based drug repositioning analysis using human and rat liver signatures is available at http://mergeomics.research.idre.ucla.edu/PharmOmics/Network_drug_repositioning/. Jaccard score-based drug repositioning analysis for all the tissues is available at http://mergeomics.research.idre.ucla.edu/PharmOmics/Jaccard_repositioning/. Potential ADR signatures aggregated from drugs having ADR is available at http://mergeomics.research.idre.ucla.edu/PharmOmics/ADRsignatures/

Competing interests

The authors declare that they have no competing interests.

Funding

YC is supported by UCLA Eureka fellowship and Burroughs Wellcome Fund Inter-school Training Program in Chronic Diseases. DA is supported by NIH-NCI National Cancer Institute (T32CA201160), UCLA dissertation year fellowship and UCLA Hyde fellowship.

PA is supported by R01 NIH/NIEHS ES02748701, R21 NIH/NIAAA AA024889, and the Burroughs Wellcome Foundation. XY is funded by NIH-DK104363, DK117850 and NIH-NS103088.

Authors' contributions

YC curated and analyzed data, constructed database, designed and conducted application studies and wrote the manuscript. DA provided supports in curation and analysis. GA, JG, NZ, PP assisted with data curation. PA contributed to manuscript writing. XY designed and supervised the research. All authors contributed to manuscript editing.

Acknowledgements

We thank Avi Ma'ayan (Icahn School of Medicine at Mount Sinai) for providing L1000FWD and CREEDS data for comparison. We also thank Luoting Zhang for her contribution in data curation for our research.

References:

1. Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: a systematic review. Health Policy. 2011 Apr;100(1):4–17.

2. Yu HWH. Bridging the translational gap: collaborative drug development and dispelling the stigma of commercialization. Drug Discov Today [Internet]. 2016;21(2):299–305. Available from: http://www.sciencedirect.com/science/article/pii/S1359644615003906

3. Lin Z, Will Y. Evaluation of Drugs With Specific Organ Toxicities in Organ-Specific Cell Lines. Toxicol Sci [Internet]. 2011 Dec 13;126(1):114–27. Available from: https://doi.org/10.1093/toxsci/kfr339

4. Denayer T, Stöhr T, Van Roy M. Animal models in translational medicine: Validation and prediction. New Horizons Transl Med [Internet]. 2014;2(1):5–11. Available from: http://www.sciencedirect.com/science/article/pii/S2307502314000022

5. Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási A-L, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun [Internet]. 2018;9:2691. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6043492/

6. Guney E, Menche J, Vidal M, Barábasi A-L. Network-based in silico drug efficacy screening. Nat Commun [Internet]. 2016;7:10331. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4740350/

7. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2010/12/18. 2011;12(1):56–68.

8. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell.

2017/12/02. 2017;171(6):1437-1452.e17.

9. Hall CJ, Wicker SM, Chien A-T, Tromp A, Lawrence LM, Sun X, et al. Repositioning drugs for inflammatory disease – fishing for new anti-inflammatory agents. Dis Model & Mech [Internet]. 2014;7(9):1069. Available from: http://dmm.biologists.org/content/7/9/1069.abstract

10. Corbett A, Pickett J, Burns A, Corcoran J, Dunnett SB, Edison P, et al. Drug repositioning for Alzheimer's disease. Nat Rev Drug Discov [Internet]. 2012;11:833. Available from: http://dx.doi.org/10.1038/nrd3869

11. Godoy P, Bolt HM. Toxicogenomic-based approaches predicting liver toxicity in vitro. Arch Toxicol [Internet]. 2012;86(8):1163–4. Available from: https://doi.org/10.1007/s00204-012-0892-5

12. Low Y, Uehara T, Minowa Y, Yamada H, Ohno Y, Urushidani T, et al. Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. Chem Res Toxicol [Internet]. 2011;24(8):1251–62. Available from: https://doi.org/10.1021/tx200148a

13. AbdulHameed MD, Ippolito DL, Stallings JD, Wallqvist A. Mining kidney toxicogenomic data by using gene co-expression modules. BMC Genomics. 2016/10/12. 2016;17(1):790.

14. Toutain P-L, Ferran A, Bousquet-Melou A. Species differences in pharmacokinetics and pharmacodynamics. Handb Exp Pharmacol. 2010;(199):19–48.

15. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. 2018 Jan 31; Available from: https://dx.doi.org/10.1093/biostatistics/kxx069

16. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, et al. The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Res [Internet].

2017;45(Database issue):D972–8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210612/

17. Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. Bioinformatics [Internet]. 2016;32(15):2338–45. Available from: http://dx.doi.org/10.1093/bioinformatics/btw168

18. Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. Nat Commun [Internet]. 2016;7:12846. Available from: http://dx.doi.org/10.1038/ncomms12846

19. Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, et al. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res. 2014/10/15. 2015;43(Database issue):D921-7.

20. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res [Internet]. 2013;41(D1):D991–5. Available from: http://dx.doi.org/10.1093/nar/gks1193

21. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res [Internet]. 2015;43(D1):D1113–6. Available from: http://dx.doi.org/10.1093/nar/gku1057

22. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res [Internet]. 2011;39(Database issue):D507–13. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013802/

23. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, et al. The characteristic

direction: a geometrical approach to identify differentially expressed genes. BMC Bioinformatics [Internet]. 2014;15:79. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24650281

24. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics [Internet]. 2012;28(4):573–80. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22247279

25. Kuleshov M V, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res [Internet]. 2016;44(Web Server issue):W90–7. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987924/

26. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res [Internet]. 2017;45(Database issue):D353–61. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210567/

27. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2016/12/03. 2017;45(D1):D331-d338.

28. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2017/11/18. 2018;46(D1):D649-d655.

29. Nishimura D. BioCarta. Biotech Softw Internet Rep [Internet]. 2001;2(3):117–20. Available from: https://doi.org/10.1089/152791601750294344

30. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics

research. Nucleic Acids Res [Internet]. 2018;46(Database issue):D661–7. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753270/

31. Wang Z, Lachmann A, Keenan AB, Ma'ayan A. L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. Bioinformatics [Internet]. 2018 Jun 15;34(12):2150–2. Available from: http://dx.doi.org/10.1093/bioinformatics/bty060

32. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science (80-). 2006/09/30. 2006;313(5795):1929–35.

33. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics [Internet]. 2017;33(18):2938–40. Available from: http://dx.doi.org/10.1093/bioinformatics/btx364

34. Arneson D, Bhattacharya A, Shu L, Mäkinen V-P, Yang X. Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration. BMC Genomics [Internet]. 2016;17(1):722. Available from: https://doi.org/10.1186/s12864-016-3057-8

35. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. shiny: Web Application Framework for R [Internet]. 2017. Available from: https://cran.r-project.org/package=shiny

36. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017/11/11. 2018;46(D1):D1074-d1082.

37. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res [Internet]. 2017;45(Database issue):D362–8. Available

from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210637/

38. Shu L, Zhao Y, Kurt Z, Byars SG, Tukiainen T, Kettunen J, et al. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. BMC Genomics [Internet]. 2016;17(1):874. Available from: https://doi.org/10.1186/s12864-016-3198-9

39. Shu L, Chan KHK, Zhang G, Huan T, Kurt Z, Zhao Y, et al. Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States. PLoS Genet. 2017/09/29. 2017;13(9):e1007040.

40. Chella Krishnan K, Kurt Z, Barrere-Cain R, Sabir S, Das A, Floyd R, et al. Integration of Multi-omics Data from Mouse Diversity Panel Highlights Mitochondrial Dysfunction in Non-alcoholic Fatty Liver Disease. Cell Syst. 2018/01/24. 2018;6(1):103-115.e7.

41. Kesim M, Tiryaki A, Kadioglu M, Muci E, Kalyoncu NI, Yaris E. The effects of sertraline on blood lipids, glucose, insulin and HBA1C levels: A prospective clinical trial on depressive patients. J Res Med Sci [Internet]. 2011 Dec;16(12):1525–31. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22973359

42. Naz F, Jyoti S, Akhtar N, Afzal M, Siddique YH. Lipid profile of women using oral contraceptive pills. Pakistan J Biol Sci PJBS. 2012 Oct;15(19):947–50.

43. Connelly PW, Stachenko S, MacLean DR, Petrasovits A, Little JA. The prevalence of hyperlipidemia in women and its association with use of oral contraceptives, sex hormone replacement therapy and nonlipid coronary artery disease risk factors. Canadian Heart Health Surveys Research Group. Can J Cardiol. 1999 Apr;15(4):419–27.

44. Kurt Z, Barrere-Cain R, LaGuardia J, Mehrabian M, Pan C, Hui ST, et al. Tissue-

specific pathways and networks underlying sexual dimorphism in non-alcoholic fatty liver disease. Biol Sex Differ. 2018 Oct;9(1):46.

45. Liss KHH, Finck BN. PPARs and nonalcoholic fatty liver disease. Biochimie. 2017 May;136:65–74.

46. Abd El-Haleim EA, Bahgat AK, Saleh S. Resveratrol and fenofibrate ameliorate fructose-induced nonalcoholic steatohepatitis by modulation of genes expression. World J Gastroenterol. 2016 Mar;22(10):2931–48.

47. Zhang N, Lu Y, Shen X, Bao Y, Cheng J, Chen L, et al. Fenofibrate treatment attenuated chronic endoplasmic reticulum stress in the liver of nonalcoholic fatty liver disease mice. Pharmacology. 2015;95(3–4):173–80.

48. Laurin J, Lindor KD, Crippin JS, Gossard A, Gores GJ, Ludwig J, et al. Ursodeoxycholic acid or clofibrate in the treatment of non-alcohol-induced steatohepatitis: a pilot study. Hepatology. 1996 Jun;23(6):1464–7.

49. Fernandez-Miranda C, Perez-Carreras M, Colina F, Lopez-Alonso G, Vargas C, Solis-Herruzo JA. A pilot trial of fenofibrate for the treatment of non-alcoholic fatty liver disease. Dig Liver Dis. 2008 Mar;40(3):200–5.

50. Hertz R, Bar-Tana J. Peroxisome proliferator-activated receptor (PPAR) alpha activation and its consequences in humans. Toxicol Lett. 1998 Dec;102–103:85–90.

51. Kawaguchi K, Sakaida I, Tsuchiya M, Omori K, Takami T, Okita K. Pioglitazone prevents hepatic steatosis, fibrosis, and enzyme-altered lesions in rat liver cirrhosis induced by a choline-deficient L-amino acid-defined diet. Biochem Biophys Res Commun. 2004 Feb;315(1):187–95.

52. Nan Y-M, Fu N, Wu W-J, Liang B-L, Wang R-Q, Zhao S-X, et al. Rosiglitazone

prevents nutritional fibrosis and steatohepatitis in mice. Scand J Gastroenterol. 2009;44(3):358–65.

53. Neuschwander-Tetri BA, Brunt EM, Wehmeier KR, Oliver D, Bacon BR. Improved nonalcoholic steatohepatitis after 48 weeks of treatment with the PPAR-gamma ligand rosiglitazone. Hepatology. 2003 Oct;38(4):1008–17.

54. Torres DM, Jones FJ, Shaw JC, Williams CD, Ward JA, Harrison SA. Rosiglitazone versus rosiglitazone and metformin versus rosiglitazone and losartan in the treatment of nonalcoholic steatohepatitis in humans: a 12-month randomized, prospective, open- label trial. Hepatology. 2011 Nov;54(5):1631–9.

55. Ratziu V, Charlotte F, Bernhardt C, Giral P, Halbron M, Lenaour G, et al. Long-term efficacy of rosiglitazone in nonalcoholic steatohepatitis: results of the fatty liver improvement by rosiglitazone therapy (FLIRT 2) extension trial. Hepatology. 2010 Feb;51(2):445–53.

56. Ratziu V, Giral P, Jacqueminet S, Charlotte F, Hartemann-Heurtier A, Serfaty L, et al. Rosiglitazone for nonalcoholic steatohepatitis: one-year results of the randomized placebo-controlled Fatty Liver Improvement with Rosiglitazone Therapy (FLIRT) Trial. Gastroenterology. 2008 Jul;135(1):100–10.

57. Simon TG, Henson J, Osganian S, Masia R, Chan AT, Chung RT, et al. Daily Aspirin Use Associated With Reduced Risk For Fibrosis Progression In Patients With Nonalcoholic Fatty Liver Disease. Clin Gastroenterol Hepatol. 2019 May;

58. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2016;16:321–57.

Tables

Table 1 The capacity of PharmOmics drug signatures to retrieve known drug targets based on various methods.

Method	Median gene number count	Target detection rate	Target detection rate from randomized gene symbols
Jaccard gene overlap	427	22.1%	10.5% ($p < 0.001$)
Pathway (GOBP and KEGG)	1402	62.4%	26.2% ($p < 0.001$)
Network (one-edge neighbors of DEGs in STRING network)	1143	41.8%	17.4% ($p < 0.001$)
Network (one-edge neighbors of liver DEGs in liver BN)	4258	71.8%	44.7% ($p < 0.001$)
Network (one-edge neighbors of liver DEGs in mismatched kidney BN)	861	40.5%	19.2% ($p < 0.001$)

Table 2 Predictive ADR signatures in rat liver, top predictive genes and new drugs predicted to have the corresponding ADRs.

ADR	FDR	Top predictive genes	New drugs predicted to have ADR
Abdominal cramps	1.8e-2	*Atg10, Fgf1, Plekhb1, Gne, Fabp5*	
Acute generalised exanthematous pustulosis	1.8e-2	*Ptpn4, Arhgdib, Slc26a8, Acly, Itgal*	Cyproterone
Bursitis	1.8e-2	*Esm1, Lgals2, Zfp395, Fnip2, Gpx3*	Chloroquine, Estriol
Calculus of kidney	1.5e-2	*Cenpf, Mvd, Esco2, Top2a, Clec4a1*	Anastrozole, Ethylestrenol
Cardiovascular disorder	4.9e-2	*Chd1l, Gfra3, Rcan1, Esm1, Dnase2b*	
Folliculitis	1.8e-2	*Ahr, Comt, Cds1, Nucb2, Ccnb2*	
Generalised oedema	1.8e-2	*Slc15a1, Fads2, Fosl1, Clec4a3, Cxcl1*	Cyclophosphamide, Tolbutamide
Hostility	4.6e-2	*Otc, Col4a2, Gbp2, Zfp395, Sema3e*	Chloroquine, Estriol
Hyperreflexia	1e-2	*S100a6, Il9r, Ddr1, Mlc1, Dcn*	
Hyperuricaemia	1.8e-2	*Psat1, Fnip2, Syngr1, Gpnmb, Slc22a8*	Benzethonium chloride, Tolbutamide
Lymphocytosis	1.8e-2	*Sema3c, Larp4, S100a6, Tmem132a, Pdlim2*	
Meningitis	1.8e-2	*Tradd, Chka, Cldn1, Usp2, Nampt*	Pramoxine hydrochloride
Nail disorder	8e-3	*Bach2, Lcp1, Coro1a, Rhpn2, Hao2*	Dexamethasone, Diazepam
Nephritis	2e-3	*Dusp6, Tgfb3, Cdkn1a, Jup, Nr1d1*	Cimetidine, Flutamide
Nephrotoxicity	4.6e-3	*Cidea, Prlr, Msl3l2, Dmbt1, Slc25a30*	Ranitidine, Tacrine
Pelvic pain	1.6e-2	*Paqr9, Top2a, Kif20a, Gpx2, Gpx3*	Chloroquine, Estriol
Pruritus	1.8e-2	*Ncapd2, Cited2, Dusp9, Mpzl1, Klf15*	
Psoriasis	1.8e-2	*Hao2, B3galt1, Gdf15, Tekt1, Acsm5*	Nefazodone, Tolbutamide
Slurred speech	1.8e-2	*Ppp1r3b, Adora2b, Paqr9, Pttg1, Slc6a13*	
Tenosynovitis	2e-3	*Dpys, Bub1b, Gcat, Ldhd, Aldoc*	
Thrombosis	1.8e-2	*Fnip2, Rhbdd2, Atf3, Fgf1, Pdlim2*	Acarbose, Tolbutamide

Figure legends

Figure 1 PharmOmics data processing pipeline. A list of clinically relevant drugs was selected from KEGG database, followed by searching across GEO, ArrayExpress, TG-GATEs and drugMatrix data repositories for datasets where drug treatment and control samples were available. Individual transcriptome datasets were analyzed with the GeoDE method to obtain ranked gene list for each treatment experiment. Datasets were annotated with tissue and species information and meta-analyzed within tissue, species, and drug to obtain genes that are altered consistently across datasets for each drug as "drug signatures". These signatures were used to conduct drug repositioning analysis and ADR prediction analysis based on direct gene overlaps or a gene network-based approach.

Figure 2. Summary of available datasets based on data sources, tissues, and species. Y-axis indicates unique dataset counts and X-axis indicates (A) tissue and data resources, (B) tissue and species, and (C) data resources and species.

Figure 3. Cross-tissue and cross-species comparisons of drug signatures in PharmOmics. (A) Distribution of drug signature overlap percentages between tissue pairs in matching species from PharmOmics. Arrow points to the pairs of tissues for drugs with high overlap in gene signatures. (B) Upset plot of cross-tissue comparison for atorvastatin signatures genes. Y-axis indicates number of genes. (C) Upset plot of cross-tissue comparison for pathways enriched in atorvastatin signatures. Y-axis indicates number of pathways. (D) Distribution of drug signature overlap percentages between pairs of species for matching tissues from PharmOmics. Arrow points to the species pair with high gene signature overlap for a

matching drug. (E) Upset plot of cross-species comparison for rosiglitazone liver gene signatures. (F) Upset plot of cross-species comparison for pathways enriched in rosiglitazone liver signatures. Pairs of tissues with shared drug signature genes or pathways are connected with black vertical lines in the bottom portion of the Upset plots.

Figure 4. AUC plot for drug repositioning for hyperlipidemia, comparing anti-hyperlipidemia drugs against other drugs under different testing conditions. Hyperlipidemia signatures were derived from either the Mergeomics tool or the CTD database. We first compared key drivers (KD) of DEGs vs DEGs on network-based drug repositioning using disease signatures from (A) MergeOmics and (B) CTD. The drug repositioning performance of PharmOmics was then compared with L1000 (C-F). Only matching drugs between L1000 and PharmOmics were used to make results comparable. (C) AUC plot based on the Jaccard score method and the hyperlipidemia signature from MergeOmics. (D) AUC plot based on the gene network method and the hyperlipidemia signature from MergeOmics. (E) AUC plot based on the Jaccard score method and the hyperlipidemia signature from CTD. (F) AUC plot based on the gene network repositioning method and the hyperlipidemia signature from CTD. Wilcoxon signed rank test was used to calculate significance between Jaccard/Network z-scores between groups. *, **, *** inducates $p < 0.05$, $p < 0.01$ and $p < 0.001$ repectively.

Figures

Figure 1.

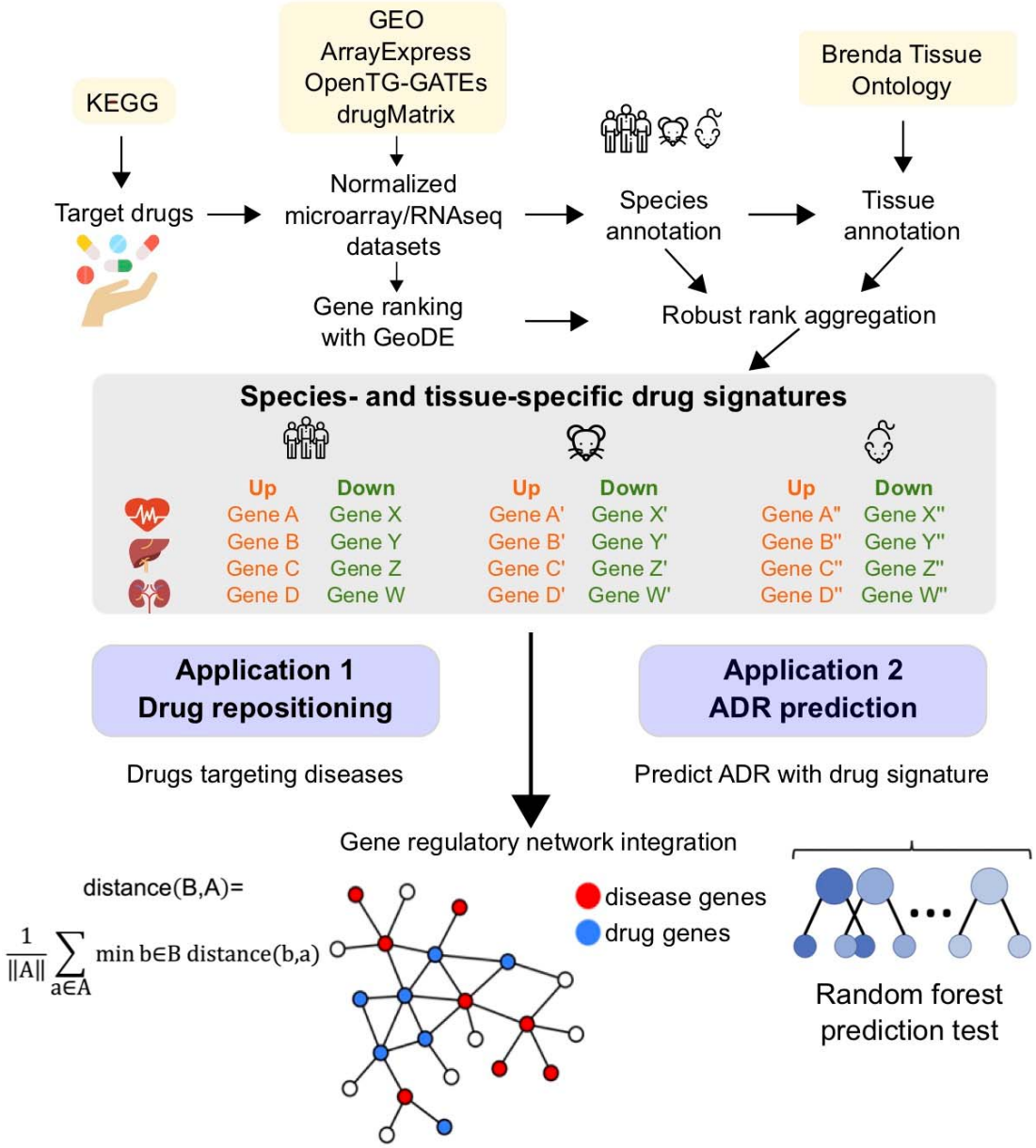


Figure 2.

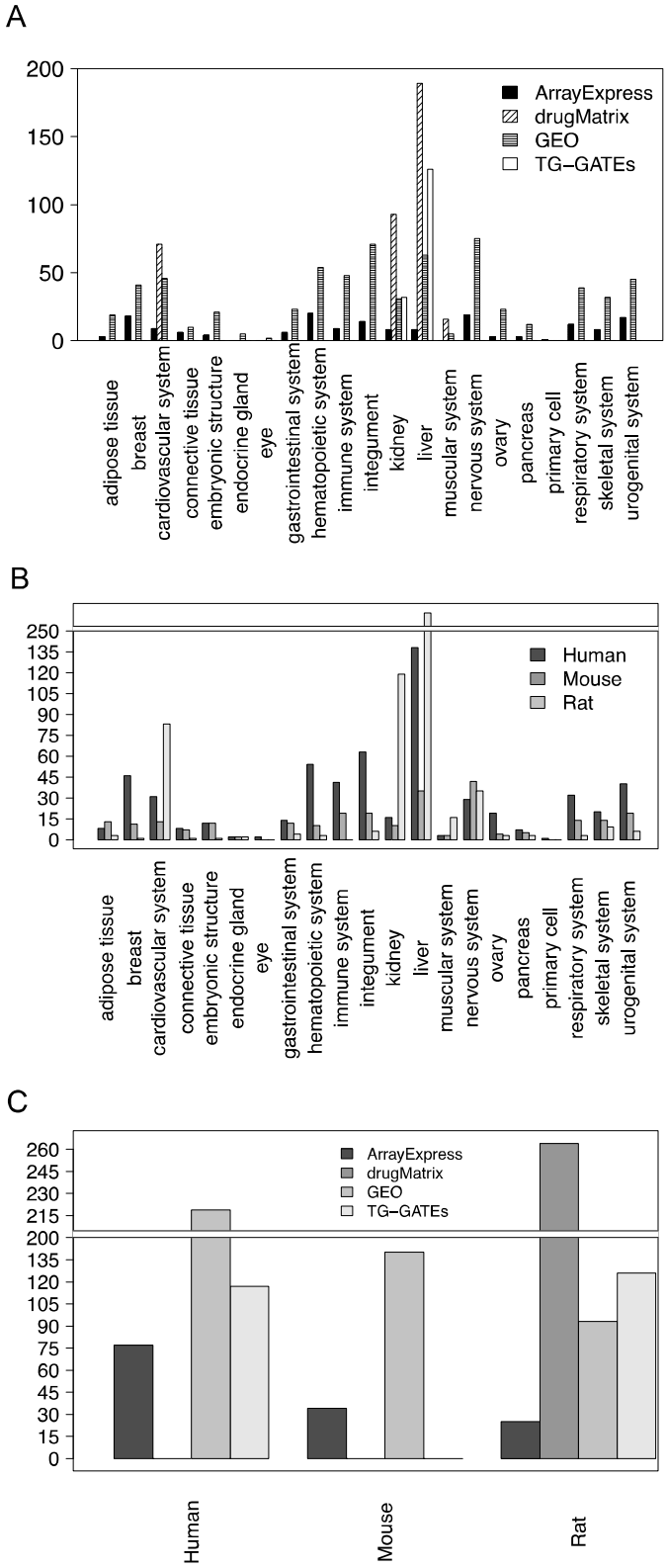

A
ArrayExpress
drugMatrix
GEO
TG–GATEs
200
150
100
50
0
adipose tissue
breast
cardiovascular system
connective tissue
embryonic structure
endocrine gland
eye
gastrointestinal system
hematopoietic system
immune system
integument
kidney
liver
muscular system
nervous system
ovary
pancreas
primary cell
respiratory system
skeletal system
urogenital system
B
Human
Mouse
Rat
250
135
120
105
90
75
60
45
30
15
0
C
ArrayExpress
drugMatrix
GEO
TG–GATEs
260
245
230
215
200
Human
Mouse
Rat

Figure 3.

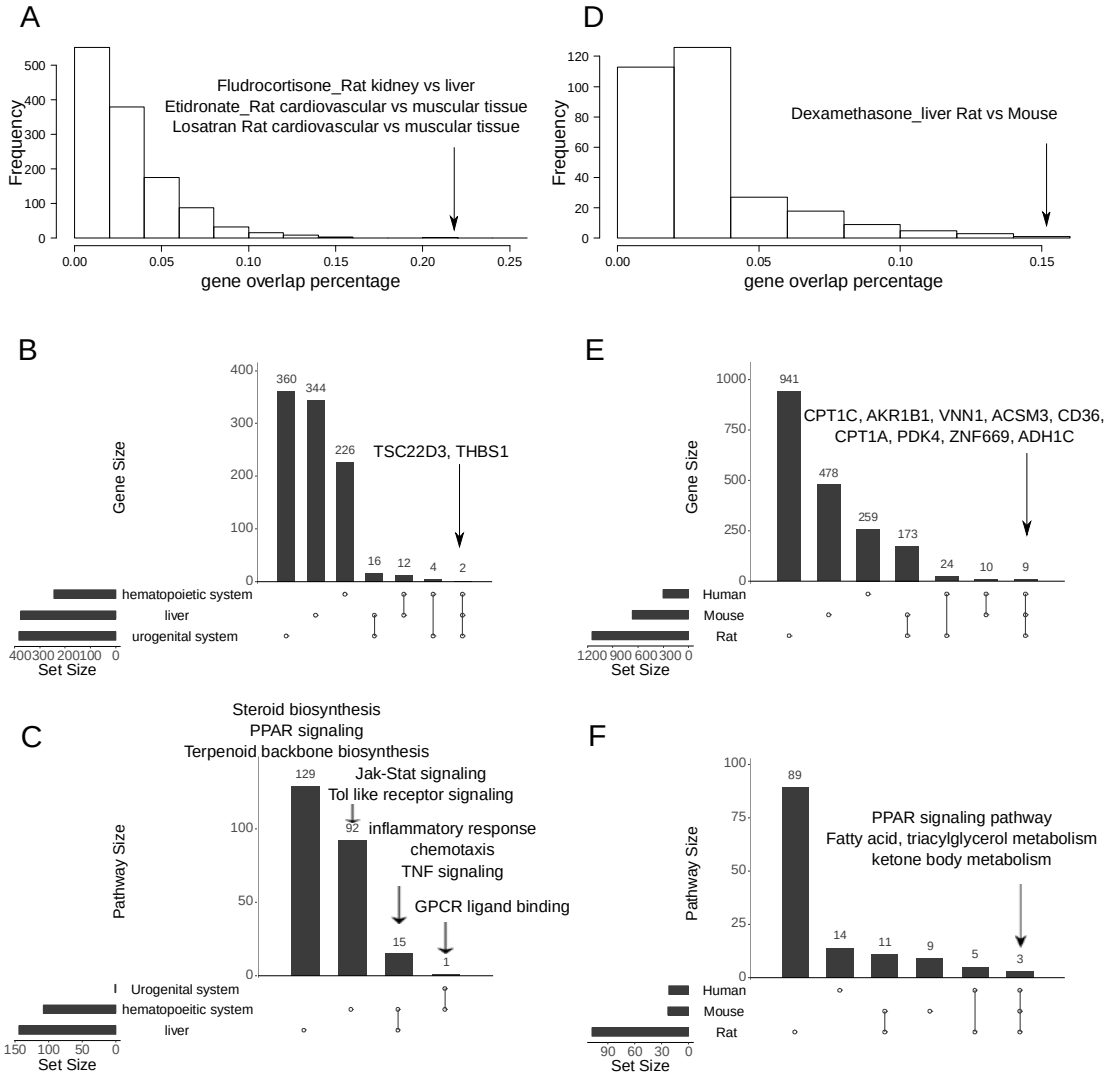
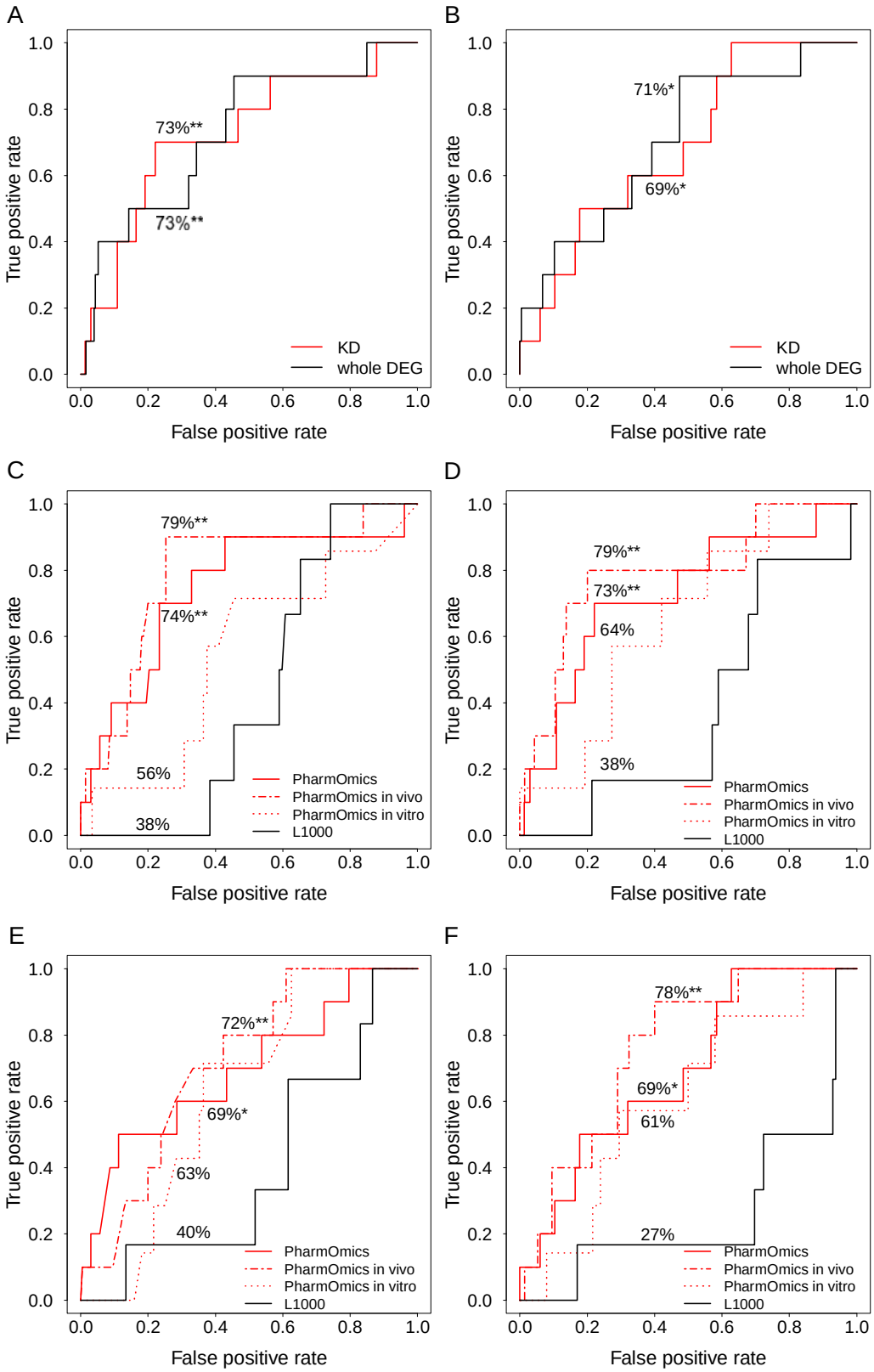

A
Frequency
Fludrocortisone_Rat kidney vs liver
Etidronate_Rat cardiovascular vs muscular tissue
Losatran Rat cardiovascular vs muscular tissue
gene overlap percentage
D
Frequency
Dexamethasone_liver Rat vs Mouse
gene overlap percentage
B
Gene Size
TSC22D3, THBS1
hematopoietic system
liver
urogenital system
Set Size
E
Gene Size
CPT1C, AKR1B1, VNN1, ACSM3, CD36,
CPT1A, PDK4, ZNF669, ADH1C
Human
Mouse
Rat
Set Size
C
Steroid biosynthesis
PPAR signaling
Terpenoid backbone biosynthesis
Jak-Stat signaling
Tol like receptor signaling
inflammatory response
chemotaxis
TNF signaling
GPCR ligand binding
Pathway Size
Urogenital system
hematopoeitic system
liver
Set Size
F
Pathway Size
PPAR signaling pathway
Fatty acid, triacylglycerol metabolism
ketone body metabolism
Human
Mouse
Rat
Set Size

Figure 4.


A
B
C
D
E
F
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
73%**
73%**
KD
whole DEG
71%*
69%*
79%**
74%**
56%
38%
PharmOmics
PharmOmics in vivo
PharmOmics in vitro
L1000
79%**
73%**
64%
38%
72%**
69%*
63%
40%
78%**
69%*
61%
27%