Supplementary Text

Table of contents

Supplementary Methods

1. A description of four expression datasets

We chose four gene expression datasets deriving from three different platforms and covering both cancer cell line and normal human tissues (i) Human U133A Gene Atlas (referred as U133A in this paper)[1]: a compendium of all transcript expression data in 84 human tissues running on the Affymetrix U133A microarray platform. This dataset was downloaded from BioGPS [2]. (ii) Human NCI60 Cell Lines (referred as NCI60 in this paper): a collection of all transcript expression data in 108 cancer cell lines running on the Affymetrix U133A microarray platform. This dataset was also downloaded from BioGPS[2]. We provided a mapping between the cell lines and the tissue origin of the cell lines (used in the column labels of Supplementary Fig. 4A) in Supplementary Table 12 (iii) HPM_PRT was released from the Human Proteome Map[3]. This project used Mass spectrometry to measure the levels of peptide sequences in 30 human tissues and mapped them to Human Refseq protein sequences to calculate protein level expression. (iv) GTEx Analysis V6 RNA-seq data (referred as GTEx in this paper)[4]: a collection of transcriptome data in 53 human normal tissues running on the Illumina TrueSeq RNA sequencing platform (we used the "Gene RPKM" file provided).

2. Calculation of over-expression p-value of target proteins

We mapped all the array or gene IDs in five datasets to Uniprot[5] protein ID and the average expression value was calculated if multiple arrays were mapped to a same Uniprot ID. Next, we converted all the expression value x to $\log_2(x+2)$ to adjust for 0 and extremely large values, so that all the converted values are no less than 1. We then normalized the expression of each protein by the median of all tissues, and took the log conversion again so that the final value will follow an approximate normal distribution across all the proteins. Then, we calculated an over-expression Z-score based on the normal distribution to represent the level of differential expression of a protein in a tissue. Z-score was converted to p-value with the "pnorm" function in R.

3. Calculation of tissue-specificity score of target proteins

The tissue-specificity score $S_{protein}$ of a target protein in a tissue is calculated as:

$$T_{protein} = \begin{cases} 0(p_{protein} \geq t_1) \\ 1(p_{protein} < t_1) \end{cases}$$

$$S_{protein} = \frac{T_{protein}}{\sum_{All\ Tissues} T_{protein}}$$

where $p_{protein}$ is the over-expression p-value of the target protein in that tissue. We used t_1 as a threshold $p_{protein}$ of to determine whether a target protein if is highly expressed in the tissue. We also used t_2 as a threshold of $S_{protein}$ to determine whether a target protein if is specifically expressed in the tissue. t_1 and t_2 were set as 0.05 and 0 in our analysis.

4. Calculation of pathway expression p-value

Since the normalized expression value after log conversion follows an approximate normal distribution across all the proteins, the average normalized value over a set of proteins in the pathway should also follow a normal distribution according to Central Limit Theorem, where the expected mean is the mean of all the proteins and the expected standard deviation is the standard deviation of all the proteins divided by the square root of the number of proteins in the pathway. Thus we calculate the expression Z-score of each pathway in a tissue based on the expected mean and standard deviation. Then Z-score was converted to p-value with the "pnorm" function in R. We used t_3 as a threshold of the pathway expression p-value determine whether a pathway if is highly expressed in the tissue. t_3 was set as 0.05 in our analysis.

5. Classification of target proteins

The classification of target proteins were mainly obtained from GtoPdb (http://www.guidetopharmacology.org/DATA/targets_and_families.csv)[6]. It classifies 2722 human proteins into 9 main classes as shown in Figure 1. The protein name of 1901 human enzymes were obtained from ENZYME[7]. In addition, we manually mapped 359 proteins to enzymes and 16 proteins to voltage-gated ion channels according to their functional annotation from Uniprot. If a protein from DrugBank is not included in any of the sources above, we classified it into "other_proteins". Altogether, we obtained the classification of 4609 unique proteins into 9 major classes.

6. Enrichment analysis of pathway category by drug class

Reactome pathways were classified into 25 major categories. Two different approaches were adopted to classify drugs: by ATC code and by class of target protein. Fisher's exact test was used to examine the significance of association between a drug class d_1 and a pathway category p_1. The 4 numbers in 2*2 contingency table were calculated as: # connections between drugs in d_1 and pathways in p_1, # connections between drugs in d_1 and pathways of other categories, # connections between drugs of other classes and pathways in p_1, # connections between drugs of other classes and pathways of other categories. Multiple testing correction was performed using FDR. FDR less than 0.01 was considered as significant.

7. A description of MEDI dataset of drug-indication

68535 indications between 3112 drugs and 4396 UMLS CUI code were downloaded from MEDI (filename: MEDI_01212013_UMLS.csv.txt). 942 of 3112 drugs have target proteins annotated from Drugbank. These drugs were used for our analysis of indication similarity. Altogether they were annotated with 23990 indications involving 2351 unique UMLS disease concepts (counted by the indication description instead of the CUI code). On average, each drug is annotated with 25 indications with a standard deviation of 26. By median, each drug is annotated with 18 indications.

8. A description of the reference standard of 4 adverse events

We used a reference standard containing the positive and negative control drugs of four adverse events: gastrointestinal bleeding, acute kidney failure, acute liver failure and myocardial infarction[8]. The summary of 4 adverse events is as following: gastrointestinal bleeding containing 77 drugs with 24 positive controls and 53 negative controls; acute kidney failure containing 53 drugs with 19 positive controls and 34 negative controls; acute liver failure containing 95 drugs with 63 positive controls and 32 negative controls; myocardial infarction containing 79 drugs with 33 positive controls and 46 negative controls.

9. Expression dataset from L1000 experiments of drug treatment

We downloaded the expression dataset of 653394 compound treatment experiments (with perturbation type of "trt_cp") from lincscloud.org[9]. In each experiment, a compound was treated to a particular cancer cell line and the expression of all the genes were measured before, 6 and 24 hours after the treatment. 38908 experiments were performed with drugs from DrugBank (including 544 unique drugs and 64 cell lines). After filtering by our results from NCI60 datasets, 92 drugs and 16 cell lines were left, with 281 connections between them in our results. Four levels of data are provided by lincscloud.org: raw, unprocessed flow cytometry data (level 1), Gene expression values (level 2), normalized expression value (level 3) and signatures with differentially expressed genes computed by robust z-scores for each profile relative to population control (level 4). We used level 4 data, Z-score of each gene representing the level of expression change after the drug treatment. If replicates were performed under the same experimental condition, we averaged the Z-score of replicates for each gene. We then mapped array IDs to Uniprot IDs. The average Z-score was calculated if multiple array IDs were mapped to a same Uniprot ID.

10. Defining a reference standard for connections between drug, cell line and pathways

Since the absolute value of Z-score represents the level of gene expression change after the drug treatment, we quantified the expression change of a pathway (in 6 or 24 hours) by using Stouffer's method to combining all the Z-scores (absolute value) of genes in the pathway as follows:

$$Z_{pathway} = \frac{\sum_{i=1}^{N} |Z_i|}{\sqrt{N}}$$

N represents the number of genes in a pathway. A pathway with $Z_{pathway}$ greater than 4.08 (correspond to p-value < 2.2e-05, corrected for multiple hypothesis testing) in either 6 or 24 hours is considered to experience significant change after the drug treatment in a cell line. Such a pathway is defined in our standard as a "positive" and the rest, with no significant change, are defined as "negative". If DATE identifies a number of pathways as "positive" for a drug *d* in a cell line *c*, we obtain the reference standard of "positive" and "negative" from the experiment with *d* treated to *c*. Then we calculated precision (TP/TP+FP), recall (TP/TP+FN) and specificity (TN/TN+FP). The performance of GOTE was compared to the null distribution where we randomly assign pathways to each drug (as the same number of DATE) without considering any other information.

11. Selecting 22 compounds for further validation on coagulation activity

A list of 69 predicted drugs were not included our initial screen. We removed 7 known anti-coagulants from the list then ranked all the compounds by their price and availability. We then selected 22 compounds with the lowest price and available at Sigma-Aldrich. The order information of 22 compounds can be found in Supplementary Table 10.

References:

1 Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 6062-6067 (2004).

2 Wu, C., Jin, X., Tsueng, G., Afrasiabi, C. & Su, A. I. BioGPS: building your own mash-up of gene annotations and expression profiles. *Nucleic acids research* **44**, D313-D316 (2016).

3 Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509**, 575-581 (2014).

4 Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660-665 (2015).

5 Consortium, U. UniProt: a hub for protein information. *Nucleic acids research*, gku989 (2014).

6 Southan, C. *et al.* The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic acids research*, gkv1037 (2015).

7 Bairoch, A. The ENZYME database in 2000. *Nucleic acids research* **28**, 304-305 (2000).

8 Ryan, P., Madigan, D., Stang, P., Schuemie, M. & Hripcsak, G. Medication ‐ Wide Association Studies. *CPT: pharmacometrics & systems pharmacology* **2**, 1-12 (2013).

9 Duan, Q. *et al.* LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic acids research*, gku476 (2014).

Supplementary Figures

Figure S1: Justification of choosing 0.775 as threshold of significant anti-coagulation activity

Line chart shows the change of odds ratio and p-value when using 5 to 95 percentile of max ratio as cutoffs to divide all 337 screened drugs into a lower group and an upper group. Odds ratio of drugs with bleeding side effect was calculated for each cutoff value. The maximum odds ratio, 1.97 was achieved at a 17 percentile cutoff, with corresponding max ratio of 0.775. Therefore, 0.775 was defined as the threshold of significant anti-coagulation activity.

Figure S2: Tissue-specificity of distinct target classes in 4 datasets (related to Figure 1b-e)

(a-d): Boxplot showing the tissue-specificity of distinct target classes in 4 datasets. The tissue-specificity of a target protein is defined as the proportion of tissues in which the target is highly expressed, when compared to the 75 percentile of all the genes. To account for the variation in the absolute expression of different genes, each gene is normalized by the baseline level. Each box on the Y-axis represents one target class. The X-axis shows the tissue-specificity of proteins belonging to the target class. "all_drug_targets" represents the combination of all the target classes. "all_genes" represents all the genes in human genome.

Figure S3: Size of predicted pathways before and after filtering in 4 datasets

Boxplot showing the size of predicted pathways. Each point represents a single pathway and each box contains all the pathways predicted in the dataset. "_filter" represents the pathways after the filtering process introduced in the main text. "Reactome" represents all the 2223 Reactome pathways between size 5 to 500, which are used as background in this study.

Figure S4: Correlation between the size of a pathway and the number of drugs that it is connected to

(a-d): Scatter Plot showing the correlation between the size of a pathway and the number of drugs that it is connected to. The results are derived from 4 datasets: U133A (a), NCI60 (b), HPM_PRT (c), GTEx (d). The Pearson correlation coefficient (pcc) is shown on the topleft of each plot along with the p-value.

Figure S5: Correlation between the number of connected tissues and the number of target proteins in the pathway

(a): Line chart showing the correlation between the number of connected tissues (y axis) and the number of target proteins in the pathway (x axis). All the drug-pathway pairs were classified into 20 groups (x axis) based on the number of target proteins in the pathway (The 20th group includes all the pairs with more than 20 target proteins in the pathway). The mean and 95% confidence interval (shown as error bar) of each group was calculated. (b) Barplot showing the comparison between drug-pathway pairs with single target protein in the pathway and drug-pathway pairs with multiple target proteins (>=2) in the pathway. Error bar shows the 95% confidence interval calculated using bootstrap.

Figure S6: Tissue-specificity of distinct ATC drug classes (related to Figure 3e)

(a-c): Heatmap showing the tissue-specificity of distinct ATC drug classes in 3 datasets: NCI60 (a), HPM_PRT (b) and GTEx (c). Each column represents an ATC drug class while each row represents a tissue in the dataset. Each cell is colored in purple or white depending on whether drugs in the ATC class are connected to this tissue or not. The scale of purple is proportional to the tissue-specificity score.

Figure S7: Enrichment of different pathway categories in drug classes (related to Figure 3f-g)

(a-c): Heatmap showing the enrichment of different pathway categories by ATC drug class. (d-f) Heatmap showing the enrichment of different pathway categories by drug class defined by the class of target proteins. Results are derived from U133A (a,d), NCI60 (b,e), GTEx (c,f). Each column represents a drug class while each row represents a Reactome pathway category. Each cell is colored from white to purple, which is proportional to the percentage of drug-pathway connections (HPM_PRT) that belong to the corresponding drug class and pathway category. "*" in a cell indicates the pathway category is significantly enriched in the drug class by Fisher's Exact Test (FDR < 0.01).