Bioinformatics, 34(9), 2018, 1498–1505
doi: 10.1093/bioinformatics/btx800
Advance Access Publication Date: 11 December 2017
Original Paper

Gene expression

gene2drug: a computational tool for pathway-based rational drug repositioning

Francesco Napolitano[1], Diego Carrella[1], Barbara Mandriani[1], Sandra Pisonero-Vaquero[1], Francesco Sirci[1,2], Diego L. Medina[1], Nicola Brunetti-Pierri[1,3] and Diego di Bernardo[1,4,*]

[1]Systems and Synthetic Biology Lab, Telethon Institute of Genetics and Medicine (TIGEM), Pozzuoli (NA) 80078, Italy, [2]Institute for Research in Biomedicine (IRB Barcelona), C/ Baldiri Reixac 10, 08028 Barcelona, Spain, [3]Department of Translational Medicine, Federico II University, 80131 Naples, Italy and [4]Department of Chemical, Materials and Industrial Production Engineering, University of Naples Federico II, 80125 Naples, Italy

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on June 7, 2017; revised on November 9, 2017; editorial decision on December 6, 2017; accepted on December 10, 2017

Abstract

Motivation: Drug repositioning has been proposed as an effective shortcut to drug discovery. The availability of large collections of transcriptional responses to drugs enables computational approaches to drug repositioning directly based on measured molecular effects.

Results: We introduce a novel computational methodology for rational drug repositioning, which exploits the transcriptional responses following treatment with small molecule. Specifically, given a therapeutic target gene, a prioritization of potential effective drugs is obtained by assessing their impact on the transcription of genes in the pathway(s) including the target. We performed in silico validation and comparison with a state-of-art technique based on similar principles. We next performed experimental validation in two different real-case drug repositioning scenarios: (i) upregulation of the glutamate-pyruvate transaminase (GPT), which has been shown to induce reduction of oxalate levels in a mouse model of primary hyperoxaluria, and (ii) activation of the transcription factor TFEB, a master regulator of lysosomal biogenesis and autophagy, whose modulation may be beneficial in neurodegenerative disorders.

Availability and implementation: A web tool for Gene2drug is freely available at http://gene2drug.tigem.it. An R package is under development and can be obtained from https://github.com/franapoli/gep2pep.

Contact: dibernardo@tigem.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The study of approved drugs for new therapeutic applications, i.e. drug repositioning, is a potential shortcut in the drug discovery process (Ashburn and Thor, 2004). Computational analysis of transcriptional responses of cells to chemical and genetic perturbations or in disease has been successfully applied for preclinical investigations of new drug indications (Chen and Butte, 2016; Iorio *et al.*, 2010; Kunkel *et al.*, 2011; Malcomson *et al.*, 2016; Pesce *et al.*, 2016; Ramsey *et al.*, 2013).

Transcriptomic approaches in industrial and academic settings leverage large scale collections of gene expression profiles such as the Connectivity Map (CMap) (Lamb *et al.*, 2006), which includes a total of 7056 genome-wide expression profiles obtained upon treatment of 5 different cell lines with different concentrations of 1309 small molecules. The CMap project is currently being scaled up by 3 orders of magnitudes by including 1.4 M profiles, derived from treatment of 15 cell lines with 15 000 small molecules and 5000 genetic perturbagens, although only 1000 genes are measured in this

case (LINCS) (Musa *et al.*, 2017; Vidović *et al.*, 2014). Recent applications of CMap include identifying drugs inducing a transcriptional response opposite to the one induced by a disease, which, therefore, may exert therapeutic effects independently of their molecular targets Chen and Butte (2016). We recently introduced Drug Set Enrichment Analysis (DSEA) (Napolitano *et al.*, 2016), in order to identify the shared mechanism of action among a set of pharmacologically diverse small molecules inducing the same phenotype, by analyzing their corresponding gene expression profiles in CMap.

An advantage of transcriptomics approaches is that they can be applied in a completely data-driven fashion, without prior knowledge about disease or therapeutic mechanisms. Although fully data-driven approaches do not require any prior information, they cannot take advantage of it when available. From this perspective, a complementary well-established route to drug repositioning is provided by rational approaches. In rational drug repositioning, a specific therapeutic target gene is known in advance and drugs modulating its activity are investigated. Many drug-target prioritization methods have been proposed in which chemo-structural considerations guide the selection of small molecules for their binding affinity with the target (target-based), or for their similarity to existing small molecules known to bind the target (ligand based) (Ma *et al.*, 2013). Nonetheless, 80% of newly discovered drugs tend to bind targets that are interactors of previously known therapeutic targets (Csermely *et al.*, 2013). Available information about protein–protein interactions can thus be exploited to improve target-based drug discovery methods. Indeed, a number of computational drug repositioning approaches exploit the known protein interactors of the therapeutic target to predict the small molecules with highest probability of modulating the target (Bultinck *et al.*, 2012; Hase *et al.*, 2009; Zhu *et al.*, 2009).

In the context of rational approaches, transcriptional responses to drug treatment can also provide important information about drug mode of action, i.e. its molecular target (Iorio *et al.*, 2010; Kibble *et al.*, 2015; Napolitano *et al.*, 2016). However, drug-induced differential expression of the molecular target, if present, can be masked by the much larger differential expression of off-target genes (Isik *et al.*, 2015). Nevertheless, off-targets may be functionally related to the intended target, so that their differential expression level can be exploited as an indirect marker of the therapeutic target activity. For this reason, computational drug repositioning methods exploiting both drug-induced transcriptional responses and protein–protein interaction networks have been recently developed (Emig *et al.*, 2013; Isik *et al.*, 2015; Laenen *et al.*, 2013).

A recently published comparison of 13 different computational approaches to drug repositioning (Isik *et al.*, 2015), including methods based on the expression of the target alone, on protein–protein interactions alone, or on a combinations of gene expression and protein–protein interactions, found the best performance for a method of the latter type, namely 'Local Radiality' (LR). LR takes into account the protein interactions among the significantly differentially expressed genes in the drug-induced transcriptional response and the therapeutic target in order to predict which drugs may modulate the target and thus are candidates for repositioning.

Here we developed a novel approach to rational drug repositioning combining drug-induced transcriptional responses with annotated pathways as an alternative to protein interaction networks. Specifically, our method relies on the identification of drugs inducing significant transcriptional modulation of pathways that involve the target gene, as opposed to its protein interactors in a protein–protein interaction network. While this approach may prioritize drugs directly acting on the therapeutic target, any drug modulating the expression of the target-related pathways, even not directly, will be selected as a potential candidate for repositioning.

We implemented the method as on line tool named 'Gene2Drug', which takes advantage of publicly available pathway annotations from different sources, and exploits the CMap data preparation pipeline previously developed for DSEA (Napolitano *et al.*, 2016). We computationally assessed the performance obtained by Gene2Drug using 10 different pathway databases and compared its performance both to the LR method and a naive method based on the target gene expression alone.

To investigate the efficiency of the method on real case scenarios, we tested Gene2Drug experimentally in two different settings: (i) to find drugs able to induce the expression of the Glutamic-Pyruvate Transaminase (GPT, aka Alanine Aminotransferase) whose overexpression was reported to reduce oxalate levels in mouse models of Primary Hyperoxaluria Type I, an inborn error of liver metabolism (Pagliarini *et al.*, 2016); (ii) to find drugs activating the transcription factor TFEB, a master regulator of lysosomal biogenesis and autophagy, whose modulation maybe beneficial in the treatment of neurodegenerative disorders (Settembre *et al.*, 2011).

2 Materials and methods

2.1 Approach

Gene2Drug uses gene expression data obtained from the Connectivity Map (CMap) (Lamb *et al.*, 2006), including genome-wide transcriptional response to treatments with 1309 different small molecules. The CMap is currently the largest single collection of drug-induced gene expression profiles in which the expression of most genes is measured [LINCS data include the expression of just 1000 genes, while the expression of the other ≈11 000 genes is computationally inferred (Vidović *et al.*, 2014)]. Gene2Drug relies on a pathway-wise version of the CMap dataset that we previously derived for the DSEA tool (Napolitano *et al.*, 2016). In the pathway-wise CMap, all of the pathways in a database are ranked according to how much the expression of genes annotated to each pathway changes after drug treatment, as shown in Figure 1. Ten different pathway databases are supported (Table 1).

The first three steps of the CMap data preparation process in (Fig. 1) are the same ones we implemented for the Drug Set Enrichment Analysis (DSEA) (Napolitano *et al.*, 2016). The two methods, however, differ in the last preprocessing step, where the ranking phase is performed column-wise for Gene2Drug and row-wise in the DSEA. The two methods have indeed two very different applications: DSEA is meant to predict a common mechanism of action (in terms of pathways) that is shared by a set of drugs given as input. Gene2drug is meant to predict drugs that are able to target a set of pathways given as input.

Given a subset of pathways including the therapeutic target gene, Gene2Drug computes for each drug an Enrichment Score (ES) and its *P*-value according to how much they tend to be up- or down-regulated by that drug, as shown in Figure 2. This is done by applying Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005) but for a set of pathways rather than a set of genes. Gene2Drug then outputs the list of 1309 drugs ranked according to the computed *P*-values.

To support gene-drug prioritization directly, Gene2Drug takes a single gene as input and it generates automatically the subset of pathways including the input gene. While a user may want to manually select a specific subset of pathways of interest and possibly

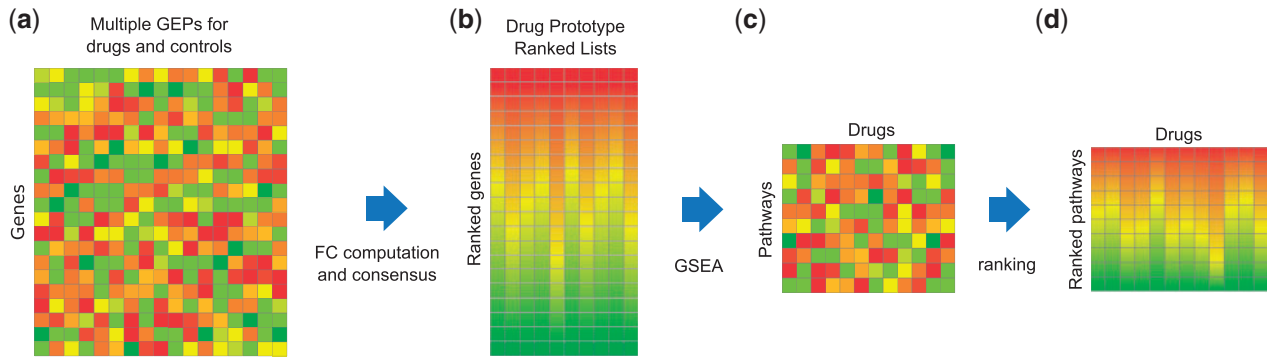

Fig. 1. Bioinformatics pipeline to compute the pathway-based version of the Connectivity Map. (**a**) Raw genome wide expression profiles were collected from the Connectivity Map and preprocessed. (**b**) Control-treatment fold change values were computed and converted to ranks. Profiles referring to the same small molecule in different experimental conditions were merged together. (**c**) Enrichment Scores and *P*-values are computed for each Drug-pathway pair. (**d**) ESs are converted to column-wise ranks according to their *P*-values (most significantly upregulated on top, most significantly downregulated at the bottom) (Color version of this figure is available at *Bioinformatics* online.)

Table 1. Pathway databases currently supported by Gene2Drug

Source	Name	Description	#
BioMart	GO BP	Gene Ontology—Biological Processes	3262
BioMart	GO MF	Gene Ontology—Molecular Function	939
BioMart	GO CC	Gene Ontology—Cellular Component	556
MSigDB	CP	Expert-defined Canonical Pathways	243
MSigDB	KEGG	Kyoto Encyclopedia of Genes and Genomes	186
MSigDB	Biocarta	Biocarta community-fed molecular relationships	217
MSigDB	Reactome	Open-source, open access, manually curated and peer-reviewed pathway database	669
MSigDB	CGP	Genetic and Chemical Perturbations	3262
MSigDB	TFT	Transcription Factor Targets	615
MIPS	CORUM	Comprehensive Resource of Mammalian protein complexes	300

Note: Pathways were obtained from 10 publicly available collections.

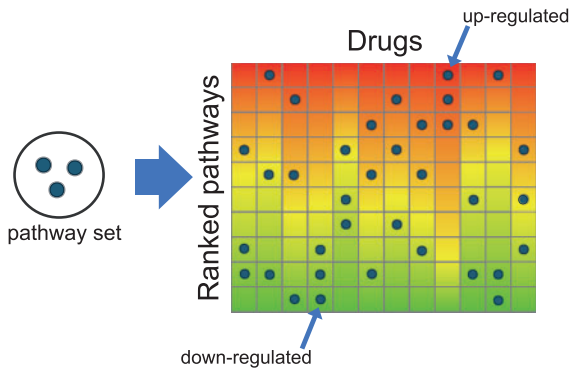

Fig. 2. Schematics of the Gene2Drug approach. Given a set of pathways containing the therapeutic target, the Enrichment Score is computed for each drug to identify those able to significantly upregulate (or downregulate) the pathways in the set. In this example, the drug in the 4th column is predicted to inhibit the target, whereas the drug in the 9th column is predicted to activate the target (Color version of this figure is available at *Bioinformatics* online.)

obtain better results in this way, the performance of the tool was assessed using this automatic selection.

2.2 Data preparation

The 6100 differential gene expression profiles from the CMap were first reduced to 1309 ranked lists of genes (one per drug), as shown in Figure 1, by merging together those obtained with the same small molecule (Iorio *et al.*, 2010). Each ranked list of genes was subsequently converted to a ranked list of pathways by means of Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005). GSEA uses a generalization of the Kolmogorov-Smirnov statistic as Enrichment Score (ES) to assess how much the genes in a pathway are distributed towards the top or bottom of the ranked list of genes. Additional details on the steps above can be found in the Supplementary Material and in Napolitano *et al.* (2016).

Given a database of pathways, we thus obtained a matrix of signed *P*-values (a negative *P*-value is the *P*-value of a pathway with a negative ES), with pathways along the rows and drugs along the columns, as shown in Figure 1. Each element (i, j) of the matrix thus contains a *P*-value assessing how significant the modulation of pathway i following treatment with drug j is. Finally, we ranked each of the columns according to the signed *P*-values, so that significantly up- (down-) regulated pathways appear at the top (bottom) of each column (Fig. 1). We applied this procedure to 10 different pathway databases, listed in Table 1, thus obtaining 10 pathway-drug matrices.

2.3 Identification of drugs modulating a target gene of interest

Given a subset of pathways containing the target gene, Gene2Drug assesses how much these pathways tend to appear at the top or bottom of each ranked list of pathways (one for each drug), as exemplified in Figure 2. To this end, GSEA is used to compute an Enrichment Score and a *P*-value for each drug. The final output is a list of drugs ranked by the corresponding *P*-values.

2.4 Experimental validation methods

Besides computational validation, Gene2Drug was tested in two different experimental settings. The corresponding experimental procedures follow.

2.4.1 GPT: Luciferase assays

Both the human Huh-7 hepatic cells and the mouse Hepa1-6 hepatoma cells were cultured at 37°C with 5% CO_2 in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. Cells were plated in 24-well plates and transfected with a reporter construct carrying the human GPT promoter driving the expression of the luciferase reporter gene (Switchgear Genomics) using Lipofectamine 2000 Transfection reagent according to the manufacturer's instructions (Life Technologies). 24 h post-transfection cells were treated with different concentrations of small molecule drugs: Fulvestrant (Sigma-Aldrich) (20, 50, 100, 125, 250, 500 μM), Tomatidine (Extrasynthese) (5, 10, 20 μM) and Nifuroxazide (Sigma-Aldrich) (5, 10, 20 μM). Prior to the treatment, the concentrations of each small molecule were tested in cells and were not found to result in cell mortality by microscopic observation. DMSO was used as drug-vehicle. After 24 h treatment, cells were washed with 1X phosphate buffered saline (PBS), lysed and assayed for Renilla luciferase activity using the Dual-GLO Luciferase Assay System (Promega) by Glomax 96 microplate luminometer. All assays were performed at least in triplicate and the data are presented as means ± standard deviation.

2.4.2 TFEB: nuclear translocation assay

Following a previously described protocol (Medina *et al.*, 2015), HeLa cells stably expressing TFEB-GFP construct were seeded in a 384-well plate, incubated for 24 h and treated with the different compounds at 0, 0.1, 1, 10, 20 and 30 μM for additional 24 h. After that cells were fixed with 4% paraformaldehyde or ice-cold methanol and permeabilized/blocked with 0.05% (w/v) saponin, 0.5% (w/v) BSA and 50 μM NH4Cl in PBS (blocking buffer). Images were acquired using the high content Opera system (Perkin Elmer) and analyzed using Harmony software and a dedicated script. Non-linear regression curves were determined by using Prism software.

3 Results

3.1 Gene2Drug

The aim of Gene2Drug is to identify one or more drugs able to modulate a therapeutic target of interest and thus prioritize these drugs for repositioning.

As shown in Figure 1, Gene2Drug makes use of publicly available data: the CMap, a collection of transcriptional responses to drug treatments, and a collection of annotated pathway databases, as listed in Table 1. Gene2Drug uses these data to build a ranked list of pathways for each drug in CMap representing the cellular response to drug treatment (Fig. 1 and Methods). Pathways at the top (bottom) of the list include those genes which tend to be transcriptionally upregulated (downregulated) following drug treatment. This resource constitutes a higher level description of the well-known ranked lists of genes and it is available for download from the Gene2Drug website. We derived 10 different versions of the pathway-wise CMap, one for each of 10 different pathway databases (Table 1) including signaling pathways, cellular components, biological processes, transcription factor targets, co-expressed and co-localized genes.

As shown in Figure 2, and described in details in the Methods, Gene2Drug exploits the well-established GSEA statistics to rank the 1309 drugs according to their ability to modulate the therapeutic target. To this end, Gene2Drug first identifies the set of pathways in the database that includes the therapeutic target. It then quantifies the drug-induced transcriptional modulation of these pathways by applying the GSEA method, where the gene-set is replaced by the pathway-set and the ranked list of genes by the ranked list of pathways. In this way, Gene2Drug assigns to each drug an Enrichment Score and a *P*-value.

Finally, drugs are ranked according to their signed *P*-value (a negative *P*-value is the *P*-value of a pathway with a negative Enrichment Score). Drugs at the top of the list are those predicted to most activate the therapeutic target, whereas the drugs at the bottom of the list, are the ones most inhibiting the therapeutic target.

Gene2Drug is implemented as a user-friendly web site publicly available at http://gene2drug.tigem.it. The web site supports both the manual input of a pathway set of interest and the automatic generation of the set starting from a target gene.

3.2 Validation

We first validated the method *in silico* in order to have a general assessment of its performances as compared to existing state-of-the-art tools. We then performed two experimental validations on different two molecular targets: the liver-specific enzyme GPT (aka ALT) and the transcription factor TFEB.

3.2.1 In silico validation

To assess Gene2Drug performance in comparison to a state-of-the-art method, we implemented the LR method, which was reported to be the best performing one across 13 different approaches (Isik *et al.*, 2015). LR makes use of a protein–protein interaction (PPI) network as obtained from the STRING database (Szklarczyk *et al.*, 2015). Given the set of significantly differentially expressed genes (DEGs) following a drug treatment, the shortest paths across the PPI network from each DEG to the therapeutic target gene of interest is computed. The average length of such paths is used to score the drug-target pair (the shorter the better). Note that there is a weak correlation between LR and Gene2Drug as genes in the same pathway tend to be closer in the PPI network (Supplementary Fig. S1).

We also implemented a naive single-gene based method as a baseline to compare with. This method simply ranks drugs according to the differential expression of the therapeutic target of interest in the CMap dataset. We refer to this method as Single Gene Expression (SGE).

To build a gold standard, we followed the approach described by Isik *et al.* (2015) based on the STITCH protein-chemical database (Szklarczyk *et al.*, 2016). Of the 1309 compounds in CMap, 607 small molecules are present in the STITCH database, corresponding to total of 133 146 drug-target pairs. This number includes 4 different evidence types ('experimental', 'prediction', 'database' and 'text-mining'). Each pair has a score for at least one of the four evidence types. A 'combined score' is also provided, which is computed over the *non-missing* scores, but it is heavily biased by the 'text-mining' evidence (61% of the pairs have a 'text-mining' evidence, against 14, 2 and 33% respectively for the 'experimental', 'prediction' and 'database' evidences, see Supplementary Fig. S2). For this reason we used all the marginal scores separately. In order to retain the most reliable predictions for the gold standard, we selected only the drug-target pairs in the top quartile of the score. To limit the computational burden, when more than 5000 drug-target pairs are present in

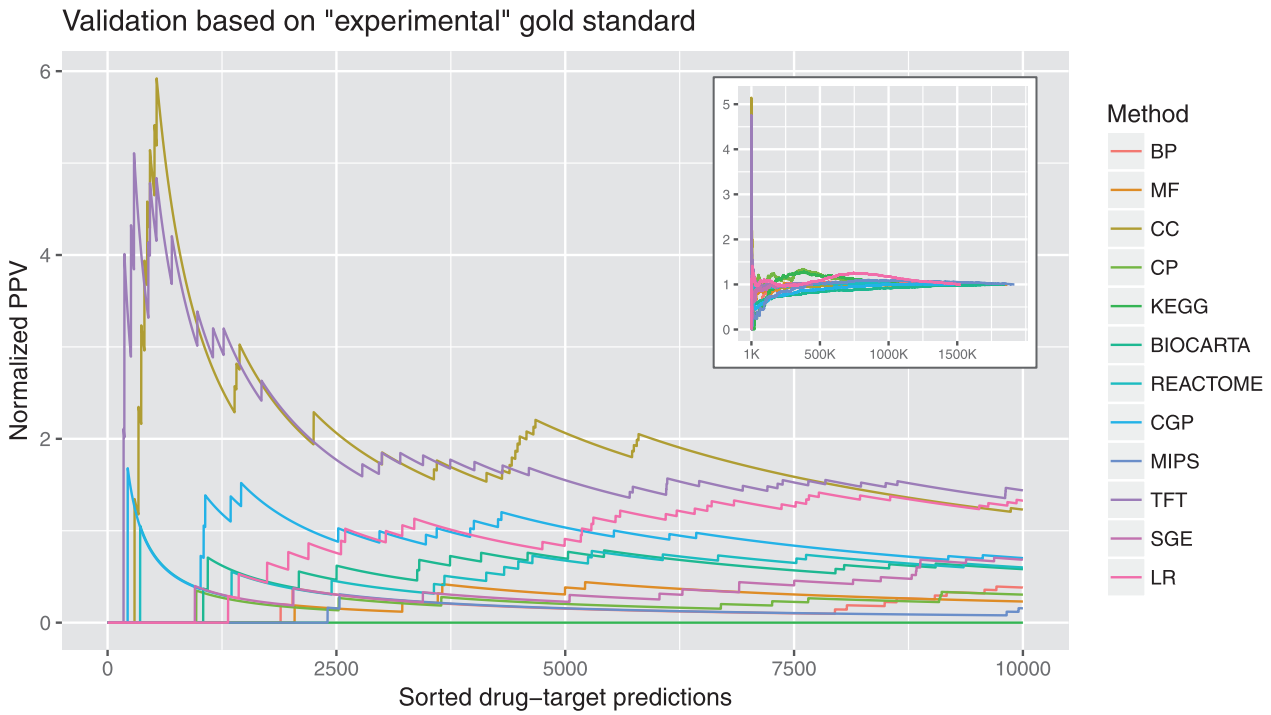

Fig. 3. Validation of Gene2Drug against the STITCH experimental gold-standard. Only drug-target pairs present in STITCH and supported by experimental evidence were included in this gold-standard (refer to Tab. 2 for additional gold-standards). For each one of the 12 methods, predicted drug-target pairs are sorted along the *x* axis and the corresponding normalized PPVs (see Materials and methods section) are reported on the *y* axis. Only the top 10 000 drug-target pairs are shown. **Inset:** PPV versus drug-targets but this time showing the PPV for all of the possible drug-target pairs, which vary according to the chosen database (BP: 1 853 544; MF: 1 864 016; CC: 1 233 078; CP: 1 117 886; KEGG: 1 127 049; BIOCARTA: 1 848 308; REACTOME: 1 424 192; CGP: 1 479 170; MIPS: 1 916 376; TFT: 442 442; SGE: 335 104; LR: 1 524 985) (Color version of this figure is available at *Bioinformatics* online.)

the top quartile, only the first 5000 were used (this condition happened only for the 'database' and 'text-mining' evidences).

Note that assessing Gene2Drug and other computational methods against such a gold standard is likely to provide significantly pessimistic outcomes, as they necessarily include a large number of false negatives. For example, it was noted that while the estimated number of possible drug targets lies between 6000 and 8000 (Overington *et al.*, 2006; Yuan *et al.*, 2016), only around 1300 known drugs are in DrugBank (Law *et al.*, 2014). Nonetheless, the gold standard represents a fair common ground to assess the relative performances of different methods.

Given one of the 4 gold standards, we used Gene2Drug with each of the 10 pathway databases and compared the resulting performances with those obtained using the naive SGE and the state-of-the-art LR methods. For the sake of clarity, we will refer to all of them as a set of 12 different methods.

Given one of the 12 methods, for each known target, we obtained a full weighed ranking of the 1309 CMap small molecules (where the weights are given by the *P*-values for Gene2Drug, the gene ranks for the SGE and the Local Radiality scores for the LR method). We then generated a single ranked list of drug-target pairs by merging together the lists obtained for the different targets according to their weights. Finally, we assessed the performance of each method by analyzing how much the top ranked drug-target pairs were enriched for true positives according to the 4 STITCH gold standards.

Figure 3 reports a summary of such analysis for the STITCH 'experimental' gold standard which includes only drug-target pairs supported by experimental evidence. In Figure 3, the *precision* value (or Positive Predicted Values, PPV) is normalized against the expected PPV for a random ordering of drug-target pairs and it is plotted as a function of drug-target pairs ordered according to one of the 12 methods (see Supplementary Material). For this gold standard, the Gene2Drug approach using the CC and TFT databases performed significantly better than the others, scoring up to 5 and 6 fold better then random respectively.

To further summarize the results, we computed also the median PPV of each method and for each gold standard considering only the top ranked 1% drug-target pairs. Table 2 reports the ranking of the methods according to this score for each gold standard. The Gene2Drug method with databases TFT and CC generally outperformed the LR method, which appeared particularly powerful when matched against the 'text-mining' gold standards. The SGE method is almost always outperformed by the others, confirming again the hypothesis that the expression of the target gene alone is not a good predictor of drug mode of action.

3.2.2 Experimental validation: upregulation of GPT expression

The glutamate-pyruvate transaminase (GPT) plays a key role in the intermediary metabolism of glucose and amino acids. GPT overexpression reduces oxalate in mouse models of Primary Hyperoxaluria Type I (Pagliarini *et al.*, 2016), a rare genetic disorder caused by loss-of-function mutations of the Alanine-Glyoxylate Aminotransferase gene (AGXT). We applied Gene2Drug to find drugs effective at increasing the expression of GPT to reduce the hyperoxaluria.

The GPT gene was annotated within seven of our pathway databases. However, Reactome, a metabolism-centric database, included the pathways most relevant to the known GPT function (*metabolism of amino acids and derivatives* end *amino acid synthesis and interconversion transamination*). Note that this is an example of

Table 2. Rankings of the methods across the 4 STITCH gold standards and corresponding medians

Method	Experimental	Database	Prediction	Text-mining	Median
CC	2	2	2	8	2.0
TFT	1	8	1	5	3.0
LR	3	4	4	1	3.5
MIPS	11	1	7	6	6.5
MF	7	7	10	7	7.0
BIOCARTA	6	10	8	3	7.0
REACTOME	5	3	9	12	7.0
CP	10	12	6	4	8.0
CGP	4	6	11	10	8.0
BP	8	9	3	9	8.5
KEGG	12	5	12	2	8.5
SGE	9	11	5	11	10.0

Note: The Gene Ontology—Cellular Component database (CC) showed the most consistent performance across all the gold standards except *Text-mining*. The Transcription Targets database (TFT) was the best performing method for the *Experimental* and *Prediction* gold standards. Local Radiality (LR) was the best for *Text-mining*. MIPS ranked top for *Database*, however performed poorly for *Experimental*, which is likely the most reliable gold standard. The Single-gene method (SGE) performed consistently worse than the others.

Table 3. Summary of luciferase assays for the three top ranked drugs predicted to upregulate GPT by Gene2Drug

Rank	Compound	ES	*P*-value	Concentration 5–20μM	Concentration 125–500μM
1	Fulvestrant	0.98	0.002	No	Yes
2	Tomatidine	0.97	0.002	Yes	N.T.
3	Nifuroxazide	0.97	0.002	No	toxic

Note: Results refer to Huh-7 cells and were confirmed in Hepa1-6 (Supplementary Figs S3, S5, S7). "Yes" indicates increased luciferase expression compared to vehicle treated controls. "No'" indicates lack of increased luciferase expression. Higher concentrations of Nifuroxazide were not tested because they resulted in cell toxicity.

user-directed choice driven by prior knowledge on the therapeutic target function that is not readily supported by methods other than Gene2Drug.

We thus run Gene2Drug with the Reactome database using GPT as input. Table 3 shows the first 3 compounds (out of 1309 compounds in CMap) ranked by Gene2Drug as those ones most upregulating the pathways involving GPT (a list of the top 30 is reported in Supplementary Table S2). We experimentally tested the efficacy of these 3 compounds to upregulate GPT in two different cell lines: Huh-7 (human hepatocyte derived cellular carcinoma cell line) and Hepa1-6 (mouse hepatoma cells). Cells were transfected with the GPT promoter driving the expression of the luciferase reporter gene (Materials and methods section). Table 3 summarizes the results for the Huh-7 cells, additional details and results for the Hepa1-6 cells are reported in the Supplementary Material. Fulvestrant resulted in significant upregulation of luciferase only at concentrations above 125 μM (left panel of Fig. 4). Tomatidine resulted in dose-dendent increase of luciferase expression at low concentrations (right panel of Fig. 4). Nifuroxazide did not show any effect at concentration less than 20 μM and was toxic in cells at higher concetrations (data not shown). Similar results for all the three compounds were confirmed also in the Hepa1-6 cell line (Supplementary Figs S3–S8).

Because of the significant effect induced by fulvestrant and tomatidine, we wondered whether using the naive single gene expression (SGE) approach, i.e. simply ranking compounds according to the differential expression of GPT, would yield similar results. Surprisingly, SGE would rank fulvestrant 695th out of 1309 among drugs overexpressing GPT, while GPT would be 6126th out of 12 012 among genes overexpressed by fulvestrant (data not shown). Similarly, tomatidine would rank 247th for GPT, while GPT would rank 2391th for tomatidine (Supplementary Table S3). Finally, we also run the LR method to identify compound modulating GPT expression, however nor fulvestrant neither tomatidine had significant LR scores (Supplementary Fig. S9 and Supplementary Table S4). Obviously, we cannot exclude that drugs predicted by LR to upregulate GPT are indeed positive hits, however we can conclude that LR and gene2drug output distinct sets of drugs.

3.2.3 Experimental validation: induction of TFEB nuclear translocation

We then asked whether Gene2Drug can help identifying compounds able to modulate the activity of a transcription factor (TF)—a particularly difficult task because TFs are usually considered undruggable targets (Bakheet and Doig, 2009). To this end, we chose TFEB, a master regulator of lysosomal biogenesis and autophagy whose modulation has potential for the treatment of neurodegenerative disorders (Settembre *et al.*, 2011). Also in this case, we chose to run Gene2Drug using a specific database based on the functional relevance of the pathways in which TFEB is annotated.

Being a transcription factor, the obvious choice would have been to select the TFT database of transcription factor targets. Unfortunately, a set of TFEB targets is not present in the current release of the MsigDB collection. We therefore chose the *Gene Ontology—Biological Processes* (GO-BP) database which included terms such as *lysosome organization* and *positive regulation of autophagy*.

Table 4 shows the 10 compounds (out of 1309 compounds in CMap) ranked by Gene2Drug as those most upregulating the GO-BP terms containing TFEB.

Among the top 10 drugs ranked by Gene2Drug, 9 were available to us. For each of the 9 drugs, we performed a High Content Screening assay for the TFEB nuclear translocation (TFEB-NT) (Medina *et al.*, 2015) at 3 h following drug administration at concentrations between 0.1 μM and 30 μM (Materials and methods section). Out of these 9 drugs, 4 were able to induce TFEB nuclear translocation, 3 of which at concentrations below 10 μM, as reported in Table 4. A representative experimental result for deptropine (one of the 4 positives drugs) is shown in Figure 5. As in the case of GPT, we verified that neither the naive SGE approach nor the LR method assigned significant ranks to the drugs identified by gene2drug (Supplementary Tables S7–S8).

4 Discussion and conclusions

We introduced a computational approach for rational drug repositioning integrating transcriptional responses to small molecules with prior knowledge in the form of annotated pathway databases.

Gene2Drug implements a complementary approach to other state-of-the-art computational methods which exploit prior knowledge in the form of gene and protein interaction networks (Isik *et al.*, 2015).

Gene2Drug is designed to be a semi-automated pipeline where the user chooses the pathways that best describe the function of the

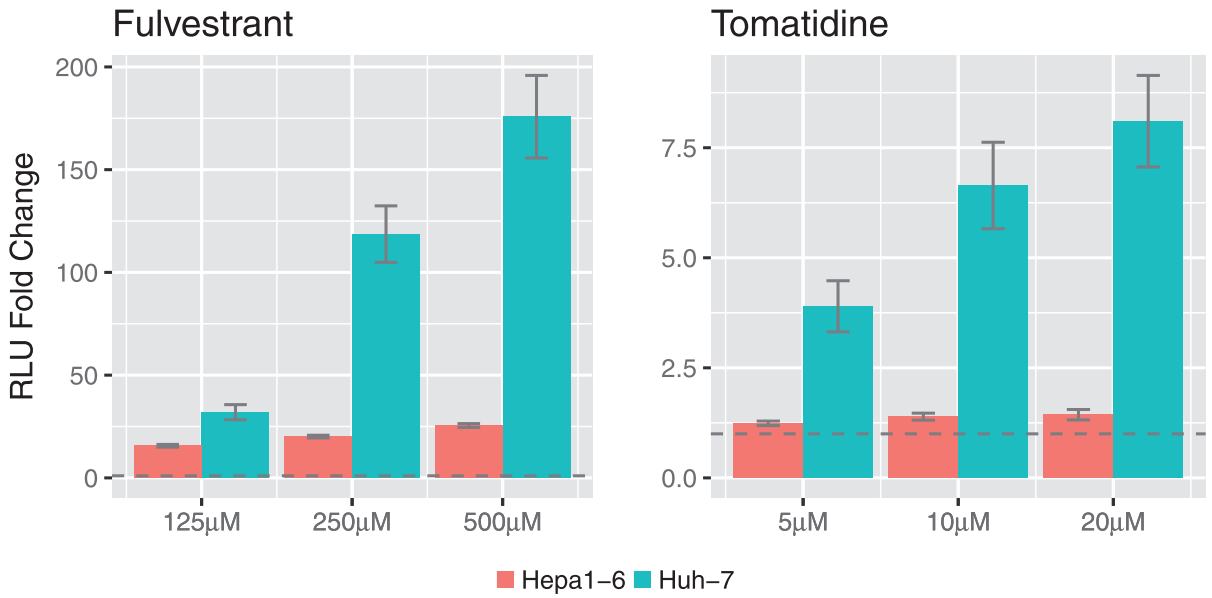

Fig. 4. Experimental validation of predicted GPT modulators. Dose-dependent increase in relative luminescence units (RLU) in Hepa1-6 and Huh-7 cells transfected with a plasmid expressing the luciferase gene under the control of the GPT promoter and incubated with various concentrations of fulvestrant (left) or tomatidine (right). The dashed line indicates RLU fold change = 1 (no effect). The two compounds were ranked 1st and 2nd respectively among the small molecules predicted to upregulate pathways including GPT by Gene2Drug (Color version of this figure is available at *Bioinformatics* online.)

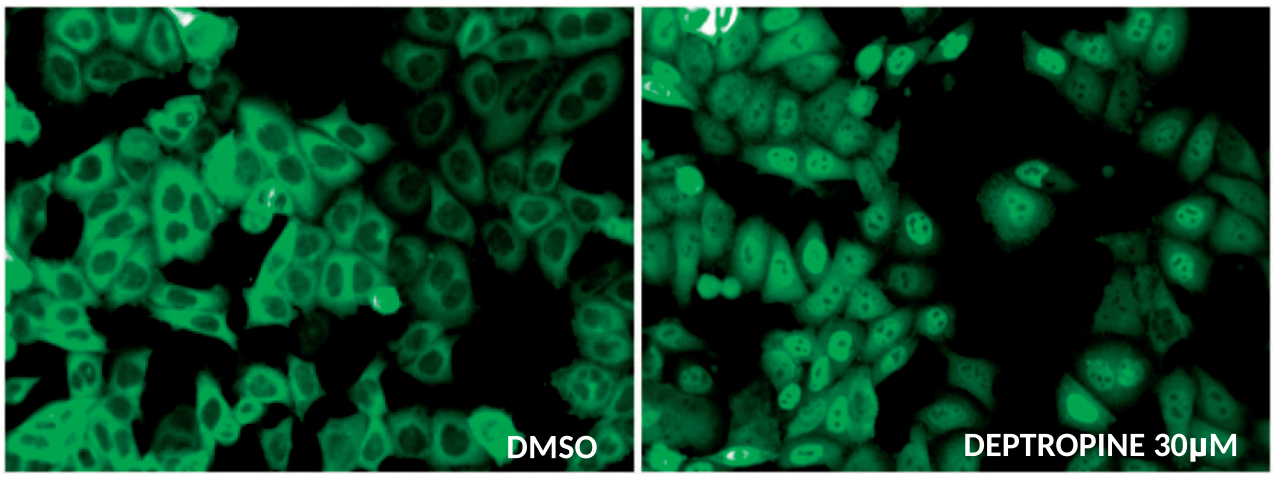

Fig. 5. Deptropine induces TFEB nuclear translocation. HeLa cells stably expressing TFEB-GFP were seeded, incubated for 24 h and treated with deptropine 30 μM or DMSO 0.2% (control) for 6 h (Color version of this figure is available at *Bioinformatics* online.)

Table 4. Drugs tested for TFEB nuclear translocation

Rank	Compound	ES	*P*-value	EC50 (μM)
1	Pimozide	0.79	0.001	>30
2	**Deptropine**	**0.78**	**0.001**	**6.636**
3	**Maprotiline**	**0.77**	**0.002**	**11.579**
4	Nifurtimox	0.75	0.002	>30
5	Benzethonium	0.75	0.003	>30
6	Alprenolol	0.75	0.003	>30
7	0297417-0002B	0.74	0.003	N.A.
8	**Miconazole**	**0.74**	**0.003**	**7.223**
9	**Loperamide**	**0.73**	**0.003**	**6.585**
10	Etoposide	0.73	0.003	>30

Note: 9 of the top 10 ranked by Gene2Drug were available to us. Four of them induced TFEB translocation at concentrations lower than 30 μM (highlighted in bold).

target gene, whose pharmacological modulation is deemed to be therapeutic. The tool ranks small molecules in the CMap database according to their ability to induce transcriptional changes in the pathways involving the therapeutic target.

Using the STITCH database as a gold standard for drug–chemical interactions, we demonstrated that Gene2Drug consistently outperforms the naive single-gene method, where drugs are ranked according to the differential expression of the therapeutic target gene only. Using the GO-CC and the TFT databases, the tool also proved to perform comparably or better than the LR method, which had been previously reported to be the most accurate across 13 different approaches.

We experimentally validated Gene2Drug in two different settings: targeting a metabolic enzyme (GPT) and a Transcription Factor (TFEB). We showed that in both cases, Gene2Drug was effective at identifying small molecules with the desired effects, at least in

cell lines, whose known direct targets where either unknown (e.g. tomatidine) or completely unrelated to the desired effect (e.g. deptropine).

Gene2Drug can be easily extended to larger collections of gene-expression profiles, such as the new LINCS database. However the L1000 platform, which LINCS is based on, actually measures the expression of ~1000 genes, with all the others being computationally inferred. While methods based on the overall similarity of transcriptional responses may not be significantly impacted by this limitation, the effectiveness of pathway enrichment analysis on L1000 data remains to be investigated. Indeed, two alternative paths to the conversion of LINCS data to pathway-based data are possible: (i) LINCS actually provides 'virtual' genome-wide gene expression data, where expression values of missing genes is inferred from the 1000 measured. However, the inference process introduces artificial correlations among genes which may alter the significance of the gene set enrichment analysis. Unfortunately a comprehensive analysis of these effects is currently missing; (ii) use only a subset of pathways sufficiently covered by the 1000 genes. In this case however, the number of pathways would be strongly reduced thus impacting on the usefulness of the tool.

In conclusion, Gene2Drug's approach to rational drug repositioning combines transcriptomic data with prior knowledge in the form of pathway databases, and provides and effective alternative to those methods based on protein interaction networks.

Funding

This study was supported by Fondazione Telethon (grant TGM11SB1 to D.d.B.); the European Research Council (IEMTx), and the Hyperoxaluria and Oxalosis Foundation to N.B.-P.

Conflict of Interest: none declared.

References

Ashburn,T.T. and Thor,K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.*, **3**, 673–683.

Bakheet,T.M. and Doig,A.J. (2009) Properties and identification of human protein drug targets. *Bioinformatics*, **25**, 451–457.

Bultinck,J. *et al.* (2012) Protein–protein interactions: network analysis and applications in drug discovery. *Curr. Pharm. Des.*, **18**, 4619–4629.

Chen,B. and Butte,A. (2016) Leveraging big data to transform target selection and drug discovery. *Clin. Pharmacol. Therap.*, **99**, 285–297.

Csermely,P. *et al.* (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Therap.*, **138**, 333–408.

Emig,D. *et al.* (2013) Drug target prediction and repositioning using an integrated network-based approach. *Plos One*, **8**, e60618.

Hase,T. *et al.* (2009) Structure of protein interaction networks and their implications on drug design. *PLOS Comput. Biol.*, **5**, e1000550.

Iorio,F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Pro. Natl. Acad. Sci. USA*, **107**, 14621–14626.

Isik,Z. *et al.* (2015) Drug target prioritization by perturbed gene expression and network information. *Sci. Rep.*, **5**, 17417.

Kibble,M. *et al.* (2015) Network pharmacology applications to map the unexplored target space and therapeutic potential of natural products. *Nat. Prod. Rep.*, **32**, 1249–1266.

Kunkel,S.D. *et al.* (2011) mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell Metabolism*, **13**, 627–638.

Laenen,G. *et al.* (2013) Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Mol. BioSyst.*, **9**, 1676–1685.

Lamb,J. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Law,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.

Ma,D.-L. *et al.* (2013) Drug repositioning by structure-based virtual screening. *Chem. Soc. Rev.*, **42**, 2130–2141.

Malcomson,B. *et al.* (2016) Connectivity mapping (ssCMap) to predict A20-inducing drugs and their antiinflammatory action in cystic fibrosis. *Proc. Natl. Acad. Sci. USA*, **113**, E3725–E3734.

Medina,D.L. *et al.* (2015) Lysosomal calcium signalling regulates autophagy through calcineurin and TFEB. *Nat. Cell Biol.*, **17**, 288–299.

Musa,A. *et al.* (2017) A review of connectivity map and computational approaches in pharmacogenomics. *Brief. Bioinf.*, **18**, 903.

Napolitano,F. *et al.* (2016) Drug-set enrichment analysis: a novel tool to investigate drug mode of action. *Bioinformatics*, **32**, 235–241.

Overington,J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.

Pagliarini,R. *et al.* (2016) In silico modeling of liver metabolism in a human disease reveals a key enzyme for histidine and histamine homeostasis. *Cell Rep.*, **15**, 2292–2300.

Pesce,E. *et al.* (2016) Evaluation of a systems biology approach to identify pharmacological correctors of the mutant CFTR chloride channel. *J. Cystic Fibrosis*, **15**, 425–435.

Ramsey,J.M. *et al.* (2013) Entinostat prevents leukemia maintenance in a collaborating oncogene-dependent model of cytogenetically normal acute myeloid leukemia. *Stem Cells (Dayton, Ohio)*, **31**, 1434–1445.

Settembre,C. *et al.* (2011) TFEB links autophagy to lysosomal biogenesis. *Science (New York, N.Y.)*, **332**, 1429–1433.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Szklarczyk,D. *et al.* (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.

Szklarczyk,D. *et al.* (2016) STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.*, **44**, D380–D384.

Vidović,D. *et al.* (2014) Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. *Front. Genet.*, **5**, 342.

Yuan,Q. *et al.* (2016) DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics*, **32**, i18–i27.

Zhu,M. *et al.* (2009) The analysis of the drug–targets based on the topological properties in the human protein–protein interaction network. *J. Drug Target.*, **17**, 524–532.