Bioinformatics, 38(2), 2022, 444–452
doi: 10.1093/bioinformatics/btab566
Advance Access Publication Date: 13 September 2021
Original Paper

Systems biology

PATHOME-Drug: a subpathway-based polypharmacology drug-repositioning method

Seungyoon Nam[1,2,3,4,†], Sungyoung Lee [5,6,†], Sungjin Park[1,3], Jinhyuk Lee[7,8], Aron Park [4], Yon Hui Kim[9,*] and Taesung Park [10,11,*]

[1]Department of Genome Medicine and Science, College of Medicine, Gachon University, 21565 Incheon, Korea, [2]Department of Life Sciences, Gachon University, 13120 Seongnam, Korea, [3]Gachon Institute of Genomic Medicine and Science, Gachon University Gil Medical Center, 21565 Incheon, Korea, [4]Department of Health Sciences and Technology, Gachon Advanced Institute for Health Sciences and Technology, Gachon University, 21999 Incheon, Korea, [5]Department of Genomic Medicine, Seoul National University Hospital, 03080 Seoul, Korea, [6]Center for Precision Medicine, Seoul National University Hospital, 03080 Seoul, Korea, [7]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, 34141 Daejeon, Korea, [8]Department of Bioinformatics, University of Sciences and Technology, 34113 Daejeon, Korea, [9]Department of Biomedical Science, Hanyang University, 04763 Seoul, Korea, [10]Interdisciplinary Program in Bioinformatics, Seoul National University, 08826 Seoul, Korea and [11]Department of Statistics, Seoul National University, 08826 Seoul, Korea

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Can Alkan

Received on January 31, 2021; revised on June 10, 2021; editorial decision on July 16, 2021; accepted on September 9, 2021

Abstract

Motivation: Drug repositioning reveals novel indications for existing drugs and in particular, diseases with no available drugs. Diverse computational drug repositioning methods have been proposed by measuring either drug-treated gene expression signatures or the proximity of drug targets and disease proteins found in prior networks. However, these methods do not explain which signaling subparts allow potential drugs to be selected, and do not consider polypharmacology, i.e. multiple targets of a known drug, in specific subparts.

Results: Here, to address the limitations, we developed a subpathway-based polypharmacology drug repositioning method, PATHOME-Drug, based on drug-associated transcriptomes. Specifically, this tool locates subparts of signaling cascading related to phenotype changes (e.g. disease status changes), and identifies existing approved drugs such that their multiple targets are enriched in the subparts. We show that our method demonstrated better performance for detecting signaling context and specific drugs/compounds, compared to WebGestalt and clusterProfiler, for both real biological and simulated datasets. We believe that our tool can successfully address the current shortage of targeted therapy agents.

Availability and implementation: The web-service is available at http://statgen.snu.ac.kr/software/pathome. The source codes and data are available at https://github.com/labnams/pathome-drug.

Contact: yonhuisarahkim@gmail.com or tspark@stats.snu.ac.kr

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

1 Introduction

Drug repositioning (equivalently, drug ‘repurposing’) aims to reveal novel indications for existing drugs (Hopkins, 2008; Sam and Athri, 2019), and is crucial when no drugs are available for specific indications (Sam and Athri, 2019). Also, drug repositioning, via computational approaches using genomics data, can save clinical attrition rates and time (Hopkins, 2008).

One popular computational drug repurposing approach utilizes gene expression signatures (Caroli *et al.*, 2018; Carrella *et al.*, 2014; Duan *et al.*, 2016; Lamb *et al.*, 2006; Subramanian *et al.*, 2017; Wang *et al.*, 2017; Yoo *et al.*, 2015) to set up reference gene expression signatures (or gene sets) related to expression changes, before and after drug treatments.

Subsequently, this approach selects drugs with gene expression signatures, in opposition to such signatures of patients (or disease models), by using gene set analysis (GSA) (de Leeuw *et al.*, 2016; Wang *et al.*, 2017). Many omics-based drug repurposing tools use this approach (Caroli *et al.*, 2018; Carrella *et al.*, 2014; Duan *et al.*, 2016; Lamb *et al.*, 2006; Subramanian *et al.*, 2017; Wang *et al.*, 2017; Yoo *et al.*, 2015).

Other popular approaches (Cheng *et al.*, 2019; Emig *et al.*, 2013; Hsu *et al.*, 2011; Krauthammer *et al.*, 2004; Martinez *et al.*, 2015; Navlakha and Kingsford, 2010; Vanunu *et al.*, 2010; Wang *et al.*, 2014; Wu *et al.*, 2013) for drug repositioning are based on linking drug-target relationships to prior signaling and protein-protein interaction (PPI) databases, including the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Szklarczyk *et al.*, 2015). After connecting biological networks of prior knowledge databases to a list of known disease genes, existing drug-target protein relationships are overlaid on the networks. When the targets of a known drug are connected or proximal to the networks, this approach (i.e. 'proximity approach') reports possible drug repositioning (Sam and Athri, 2019). However, this approach does not utilize transcriptomics information, thus ignoring expression statuses of gene entries in the networks.

These two approaches have other limitations. First, they do not consider what specific signaling pathways, or their regulatory subparts (equivalently, 'subpathways'), are dysregulated in patients or phenotype changes (Koumakis *et al.*, 2016). Also, these approaches do not consider the polypharmacology [i.e. multiple targets of a known drug (Bolognesi, 2019; Jansson *et al.*, 2015)] of dysregulated signaling subpathways, due to their different drug repositioning strategies.

Recently, polypharmacology has been promoted for Food and Drug Administration (FDA)-approved drugs (Bolognesi, 2019; Jansson *et al.*, 2015). Polypharmacology refers to multiple targets of a known drug (Bolognesi, 2019; Jansson *et al.*, 2015). Two advantages of polypharmacology exist. First, from the perspective of network medicine, disease can be considered to result from the systemic dysregulation of signaling networks (Vitali *et al.*, 2013). In that regard, one therapeutic aim is to recover such dysregulated networks by modulating key components of the networks simultaneously (Vitali *et al.*, 2013). Thus, when a drug's multiple targets (i.e. polypharmacology) are revealed in key components of dysregulated networks, the drug is likely to restore those networks' functions (Vitali *et al.*, 2013). Second, polypharmacology of Food and Drug Administration (FDA)-approved drugs has been incorporated into the field of drug repositioning (Bolognesi, 2019; Jansson *et al.*, 2015), since FDA-approved drugs have already been validated in terms of efficacy and safety (Bolognesi, 2019; Jansson *et al.*, 2015).

Despite this potential, the two aforementioned phenomena (i.e. gene expression signatures and proximity approaches) have not been considered important in assessing the polypharmacology of networks. While a few web services for GSA-based drug-gene association methods are freely available (e.g. WebGestalt) (Wang *et al.*, 2017), few network-based polypharmacology web services have so far been offered.

In this study, to address the aforementioned limitations, we develop not only a new subpathway network-based polypharmacology drug repositioning method, but also provide the web services for the new method, PATHOME-Drug. For evaluation, we implemented synthetic expression data in a more biologically realistic manner by considering interdependency between biological entities. Finally, our method performed better than two well-known GSA methods, WebGestalt (Wang *et al.*, 2017) and clusterProfiler (Yu *et al.*, 2012), in terms of detection of signaling pathways and drug repositioning.

2 Materials and methods

2.1 Method of our subpathway network-driven polypharmacology drug repositioning tool, PATHOME-Drug

A schematic overview of PATHOME-Drug, consisting of the four phases is given in Figure 1. PATHOME-Drug takes transcriptomic datasets as input. Our method generates a network by collecting and merging subpathways from previously known signaling KEGG pathways such that regulations between biological entities, within the subpathways, are statistically different, in terms of gene expression profiles between two phenotypes (e.g. normal versus cancer). Then, considering polypharmacology-based drug repositioning (Bolognesi, 2019), drugs affecting multiple targets, in the significant subpathways, are identified by a statistical test. Such drugs' polypharmacologies could change statuses (e.g. gene expression) of biological entities, in the subpathways, by regulating their multiple target components, potentially reversing disease phenotypes.

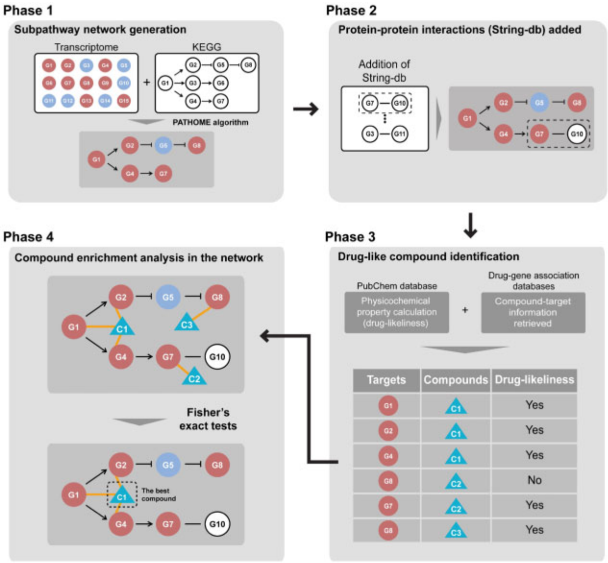
Phase 1
Subpathway network generation
Transcriptome
G1
G2
G3
G4
G5
G6
G7
G8
G9
G10
G11
G12
G13
G14
G15
+
KEGG
PATHOME algorithm
Phase 2
Protein-protein interactions (String-db) added
Addition of String-db
G7
G10
G3
G11
Phase 3
Drug-like compound identification
PubChem database
Physicochemical property calculation (drug-likeliness)
Drug-gene association databases
Compound-target information retrieved
Targets
Compounds
Drug-likeliness
G1 C1 Yes
G2 C1 Yes
G4 C1 Yes
G8 C2 No
G7 C2 Yes
G8 C3 Yes
Phase 4
Compound enrichment analysis in the network
C1
C2
C3
Fisher's exact tests
The best compound

Fig. 1. Overview of PAHOME-Drug. PATHOME-Drug consists of four phases to suggest subpathway network-based polypharmacology drug repositioning. The first phase constructs significant subpathway-networks from KEGG for locating specific subparts of whole signaling pathways, relevant to disease statuses. The second phase then extends reliable PPIs to the networks, while the third phase constructs a compound physicochemical property database to find drug-like compounds, along with drug-target relations. The last phase suggests polypharmacology drugs, regulating multiple targets, within the subpathway-network, using compound enrichment tests

2.1.1 Phase 1: subpathway-network generation

The first phase was implemented for a subpathway-level analysis by using a previous method, PATHOME (Nam *et al.*, 2014). Gene expression data and pathways were input to identify significant subpathways, i.e. linear paths from KEGG pathways, decomposed from top nodes to leaf nodes, by using depth-first search. Statistically significant subpathways were then selected. Subsequently, Cytoscape (Shannon *et al.*, 2003) located the shared nodes among the selected subpathways, to merge them into a single network.

For a pair of adjacent gene nodes i and j in a subpathway graph $S(V, E)$ with nodes V and edges E, the rule was defined as follows (Nam *et al.*, 2014):

$$cor(i,j)reg(i,j) > 0, \forall(i,j) \in E(S) \qquad (1)$$

where $cor(i, j)$ is Pearson's correlation coefficient between gene expression profiles of the nodes i and j. The function $reg(i, j)$ as *a priori* activation ($reg = 1$) or inhibition ($reg = $-1) between the adjacent nodes was obtained from the KEGG pathways. When subpathway S satisfies the rule in control and experimental groups in gene expression data, further statistical tests were performed under no difference of correlation magnitudes of $E(S)$ between the two groups (Nam *et al.*, 2014). Subsequently, statistically significant subpathways were merged into a single network, enabling us to explain which signaling subparts allow compounds and drugs to be selected. Input and output of our web services are described in Supplementary Method S1 and Supplementary Figure S1.

2.1.2 Phase 2: addition of PPIs (PPIs) to the subpathway-network

In the second phase, for revealing further biological relationships, PPIs of first-order neighborhoods of the genes in the subpathways were incorporated into the subpathway-network of the first phase. For this purpose, PPIs (evidence score greater than 0.9) from STRING-db (Szklarczyk *et al.*, 2015) were used.

2.1.3 Phase 3: drug-like compound identification

In the third phase, physicochemical properties of compounds were utilized for integrating the networks of the second phase (see Supplementary Method S2 for details). Curated information of the ligands' (i.e. interacting drug molecules) binding to the proteins, obtained from DrugBank (Knox *et al.*, 2011) and PharmGKB (Thorn *et al.*, 2010), was used to construct our in-house physicochemical property database, for the two databases. We then obtained 'drug-like' compounds by applying a physicochemical property rule, called Lipinski's rule of five (henceforth, RO5 rule) (Lipinski, 2004), to the compounds of our drug database. These drug-like compounds, satisfying the RO5 rule plus the two additional filters (henceforth, RO5 variant), were also considered for drug repositioning within the subpathway-network in the next phase.

2.1.4 Phase 4: compound enrichment analysis in the subpathway-network for polypharmacology

In this stage, a statistical test for identifying a 'drug binding to multiple targets', i.e. polypharmacology (Bolognesi, 2019), in the given subpathway-network, was performed (Supplementary Method S3). For measuring the statistical enrichment for compound targets in the network, we used Fisher's exact test by constructing a two by two contingency table for each compound. In the contingency table for a given compound, one factor had two levels (targets versus non-targets), and the other factor two levels (network entries versus non-network entries).

2.2 Performance evaluation of signaling detection in simulated gene expression data, without considering regulations

Regarding signaling pathway detection, we measured powers and false discoveries in the three tools: PATHOME-Drug, WebGestalt (Wang *et al.*, 2017) and clusterProfiler (Yu *et al.*, 2012). To calculate power and false discovery, true positive and true negative pathways should be set. In real gene expression datasets, however, all true positive and true negative pathways have not been experimentally identified, thus requiring simulation. Several true positive pathways have been confirmed by *in vitro* and *in vivo* experiments. It is natural that these experimentally validated pathways could be regarded as true positive pathways. Thus, the gene expression profiles, in themselves, for the gene entries in these validated pathways, can be regarded as true positive data. For true negative pathways, it is often assumed that true negative pathways refer to non-significant pathways between control and experimental groups (e.g. patients versus healthy individuals). These non-significant pathways can be simulated under the null hypothesis that the gene expression profiles in the pathways are not different between the two groups.

We then generated simulated datasets, without considering network structure, using our previous GEO dataset (accession number: GSE36968) (Chang *et al.*, 2016). In brief, this dataset contains 18 890 gene expressions of 6 non-cancerous gastric tissue and 24 gastric cancer (GC) tissue samples. Through experimental validation (Chang *et al.*, 2016), we confirmed that the dataset represented the WNT signaling pathway, which we assumed to be the true causal pathway for KEGG entries for 150 genes. In this respect, we generated 100 simulated datasets, by resampling expressions of the 150 genes within the WNT pathway, and assumed 100 datasets as true positives.

For true negative dataset construction, the gene expressions of the other 18 740 genes (18 890 minus 150; equivalently, non-WNT pathway genes) were designed to have no difference between the non-cancerous samples and the gastric cancer samples, assuming true negatives. In other words, the non-cancerous samples and the gastric cancer samples have the same normal distribution, for which the means and variances for each gene were obtained from the non-cancerous samples. Here, we obtained 100 datasets by simulating the normal distributions of the 18 740 genes, assuming true negatives.

We combined both true and negative datasets to result in the 100 final simulation datasets (Fig. 2). In generating the simulated dataset, the WNT signaling pathway was set as a true positive and non-WNT signaling pathways as true negatives. In other words, the power was defined as how many times the *P*-values of the WNT signaling pathway detections were less than the significance level α (i.e. 0.05), in the 100 (say, M) simulated datasets (Equation 2). The number of false signals (i.e. false discoveries) were defined as how many of the 100 simulated datasets reported any non-WNT signaling pathways having *P*-values less than the significance level α (Equation 3).

$$\text{Power} = \frac{1}{M}\sum_{j=1}^{M} I(p_{WNT,j} < \alpha) \tag{2}$$

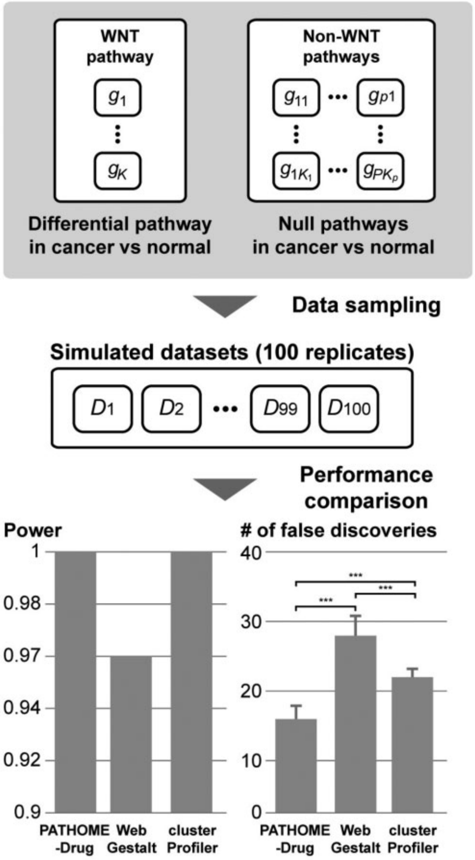

Fig. 2. Performance evaluation of PATHOME-Drug, WebGestalt and clusterProfiler for signaling detection in simulated gene expression data, for gastric cancer versus normal samples, without considering regulation. We constructed simulated datasets (see the Methods section), such that WNT signaling was the differential pathway in cancer versus normal groups, and the genes of other pathways (null pathways) were not different between the two groups. We measured powers (bottom left panel) for the WNT signaling and false discoveries (bottom right panel) for the null pathways, in the 100 simulated datasets, by the two methods, indicating better performance of PATHOME-Drug over WebGestalt for both measurements. PATHOME-Drug also demonstrated less false discoveries than clusterProfiler, even though PATHOME-Drug and clusterProfiler had the same power. (n.s., not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$)

$$\text{offalsediscoveries} = \frac{1}{M}\sum_{j=1}^{M}\sum\nolimits_{i\in S-\{\text{WNTpathway}\}} I(p_{i,j} < \alpha) \quad (3)$$

where $I(\bullet)$ represents the indicator function, and $p_{i,j}$ the P-value of signaling pathway i detection in the jth simulated dataset. Signaling pathway i represents an element of signaling pathway set, $S = \{$WNT pathway, .., MAPK pathway$\}$. For example, $p_{\text{WNT},j}$ indicates the P-value of the WNT pathway in the jth simulated dataset. We obtained powers and false discoveries for the tools in the simulated datasets.

2.3 Comparison among PATHOME-Drug, WebGestalt and clusterProfiler for compound suggestion, in the eight real GC datasets

We applied the tools to the eight GC expression datasets to obtain the numbers of statistically significant compounds. We compared the numbers of the compounds, to inspect which tool suggested more compounds, noting that, regarding compound-target relationships, PATHOME-Drug used DrugBank (Knox *et al.*, 2011) and PharmGKB (Thorn *et al.*, 2010). WebGestalt (Wang *et al.*, 2017) used PharmGKB (Thorn *et al.*, 2010) and a literature-based compound-protein database, GLAD4U (Jourquin *et al.*, 2012). ClusterProfiler used DSigDB (Yoo *et al.*, 2015). For statistical significances, q-values < 0.1, P-values < 0.05 and FDR q-values < 0.1 were used for PATHOME-Drug, WebGestalt and clusterProfiler, respectively.

2.4 Performance evaluation of signaling detection and drug repurposing in simulated gene expression data, considering network structures

Since traditional simulations do not reflect network structure, we performed more biologically realistic simulation, under a Gaussian Bayesian Network (GBN) model, using a Bayesian learning package, bnlearn (Scutari, 2010) that considers network structure of pathways. The simulated datasets were used to calculate the powers of signaling detection and drug detection (see Supplementary Method S4 for details).

The network structures of four well-established GC signaling pathways (i.e. 'true' pathways) (Chang *et al.*, 2016; Zhang *et al.*, 2014) were selected from KEGG (WNT signaling: KEGG hsa04310; mTOR signaling: KEGG: hsa04151; MAPK signaling: KEGG hsa04010; and JAK-STAT signaling: KEGG hsa04630). The 'true' pathways were then regarded as differential pathways in GC versus control samples. The log2-normalized gene expressions of 30 real GC biological samples (24 GC versus 6 normal-appearing samples), in the GEO dataset GSE36968 (Kim *et al.*, 2012), were used for simulating gene expression. Simulated gene expression data for a given true pathway was generated by a GBN model, using the bnlearn package (Scutari, 2010).

Gene expression data for the other pathways (say, null pathways), not belonging to the given pathway, was obtained by resampling the normal samples, without regard to phenotypes in GSE36968. Null pathways can be regarded as non-differential pathways in GC versus normal samples. So, these genes were not different between the two phenotypes, and it was expected that no pathways except the four 'true' pathways were likely to be significant.

Under diverse GBN parameter settings, we generated 100 simulated datasets for each GBN parameter setting, and subsequently calculated the powers for the tools. The power of differential pathway detection was defined as how many times the P-values of detecting the 'true' differential signaling pathway i were less than the significance level $\boldsymbol{\alpha}$ (e.g. 0.05) in the 100 (say, M) simulated datasets (Equation 4). The true differential pathway i belongs to {WNT pathway, mTOR pathway, MAPK pathway, JAK-STAT pathway} as mentioned earlier. The power of identifying drugs in the 'true' differential pathway i was defined as how often the P-values of the drugs, associated with i, were less than the significance level $\boldsymbol{\alpha}$ (Equation 5).

$$\text{Power of the differential pathway } i \text{ detection} = \frac{1}{M}\sum_{j=1}^{M} I(p_{i,j} < \alpha) \quad (4)$$

$$\text{Power of drug detections in the differential pathway } i = \frac{1}{M}\sum_{j=1}^{M} I\Bigg(\bigg(\sum_{d\in D_i} I(p_{i,j,d} < \alpha)\bigg) > 0\Bigg), \quad (5)$$

where $p_{i,j,d}$ indicates the P-value, reported by the tools, of drug d that is statistically associated with the true signaling pathway, i, in the jth dataset. The set D_i indicates the list of known compounds (drugs) for the ith pathway. The $p_{i,j}$ notation is the same as Equation 3.

3 Results

3.1 Overview

PATHOME-Drug consists of four phases (Fig. 1). The first phase is to extract subpathways from prior signaling KEGG pathways, such that regulations between biological entities within the subpathways are statistically different, in terms of the expression profiles specific to the two phenotypes (e.g. normal versus cancer). Subsequently, the subpathways were merged into a network. This phase was implemented, using our previous method (Nam *et al.*, 2014). The second phase was to extend the network, using PPIs (Szklarczyk *et al.*, 2015). The third phase was to construct a database for compound-target information, and drug-likeness information, of the compounds. Drug-likeness refers to a drug's physicochemical properties necessary to become a successful, efficacious and safe biological agent (Segall, 2012). The last phase was to identify either compounds or drug-like compounds for which multiple targets were enriched in the network. In other words, a drug binding to multiple targets in the given networks, i.e. polypharmacology (Bolognesi, 2019), was reported in this phase.

For performance evaluation, we generated the two types of simulated gene expression datasets. The first type did not consider regulations among genes, while the second type did consider regulations, using Bayesian learning. The second type is more biologically realistic than the first one.

3.2 Performance evaluation of signaling detection in simulated expression data, without considering regulations

We compared the performances of PATHOME-Drug and a classical GSA tool for signaling pathways and drug associations, WebGestalt (Wang *et al.*, 2017). While there are many different drug repositioning methods, few of these implement these methods. One such Web-based method, WebGestalt, was selected as a baseline method for our study. WebGestalt included gene sets for signaling pathways and for drug-target genes. Taking the user's own genes as input, WebGestalt provides statistically enriched gene sets for signaling pathways and drugs, separately. We also considered another popular method, over-representation analysis (ORA) (Boyle *et al.*, 2004), based on a list of differentially expressed genes (DEGs). The ORA, implemented in the R package, clusterProfiler (Yu *et al.*, 2012), was applied to the simulated datasets, for identifying significant pathways and drugs. The clusterProfiler package used the KEGG database and DSigDB (Yoo *et al.*, 2015) as the sources of pathway-gene associations and drug-gene associations, respectively.

We generated 100 simulated gene expression datasets without considering regulations between protein-coding genes (see Methods for details). As a result, our method, WebGestalt and clusterProfiler showed powers of 1, 0.96 and 1, in signal pathway detection, respectively, indicating better or equal power of our method, compared to WebGestalt and clusterProfiler (Fig. 2). Regarding false discoveries of signaling pathway detection, PATHOME-Drug showed less false discoveries (15.84 ± 3.29, average and standard error), when compared to WebGestalt (27.84 ± 5.71) and clusterProfiler (22 ± 1.5) (Fig. 2).

3.3 Comparison among PATHOME-Drug, WebGestalt and clusterProfiler in terms of compound suggestions, in the eight real GC datasets

Given specific disease pathogeneses represented as networks or gene sets, the discovery of compounds that could reverse detrimental phenotypes is critical for restoring physiological homeostasis. Also, since available targeted therapies (or compounds), in cancer in particular, are considerably limited, it is important to suggest more compounds for drug repositioning. To that end, in the first comparison among the three tools, we measured the number of statistically significant compounds reported by PATHOME-Drug, WebGestalt and clusterProfiler, in the eight GC expression datasets. Also, considering the quality of their physicochemical properties (i.e. drug-likenesses), in the second comparison, we regarded drug-like compounds as a reference set, comparing the number of statistically significant drug-like compounds between the tools.

In the first comparison, we compared PATHOME-Drug, WebGestalt and clusterProfiler, by measuring the number of compounds detected from the eight gastric cancer (GC) datasets (see Supplementary Table S1 for dataset descriptions). This assessment showed that PATHOME-Drug identified a greater number of statistically significant compounds, in a majority of the eight datasets, compared to WebGestalt and clusterProfiler (Fig. 3A).

In the second comparison, we considered compounds' drug-likenesses (Daina *et al.*, 2017). Drug-likeness refers to the physicochemical properties of a molecule as an oral drug regarding bioavailability, solubility, ligand efficiency, etc. (Daina *et al.*, 2017) (see details in the methods). Also, in this comparison, our method suggested more drug-like compounds, in five of the eight datasets, than WebGestalt (Fig. 3B) and more than clusterProfiler, in all the datasets (Fig. 3B).

3.4 Performance evaluation of signaling detection and drug repurposing, in simulated gene expression data, while considering network structure

Since our proposed method showed comparable performance to the existing methods, in the traditional simulation approach that does not consider complex structure of pathways, we performed further simulations, to reflect network structures of pathways.

Using the network structures of four GC-related KEGG pathways (WNT signaling, mTOR signaling, MAPK signaling and JAK-STAT signaling) (Chang *et al.*, 2016; Zhang *et al.*, 2014), and a GC dataset (GEO accession: GSE36968) (Kim *et al.*, 2012), 100 simulated replicate datasets for GC versus normal-appearing samples were generated for each pathway, using a GBN (Fig. 4A; Supplementary Method S4 in details). The network structure of each pathway was determined using a Bayesian learning tool, bnlearn.

The performance results (Fig. 4B) showed that the simulation considering pathway network structure was crucial for evaluating statistical power. First, the powers of identifying the 'true' pathways were substantially higher in PATHOME-Drug than in WebGestalt. The powers of PATHOME-Drug showed 0.72–0.99 in all the simulated pathways, regardless of simulation parameters, while those of WebGestalt were modest and highly variable (0.03–0.19, Fig. 4B). The powers of clusterProfiler were zeros in all the settings (Fig. 4B).

Next, in aspect of detecting compounds associated with a pathway, PATHOME-Drug (power 0–0.95), outperformed WebGestalt (power 0–0.05) and clusterProfiler (power 0) in the four true pathways (Fig. 5). Also, we inspected the scenario of identifying drug-like compounds (a subset of compounds) in a pathway, and the powers of PATHOME-Drug (power 0–0.58) were consistently higher than those of WebGestalt (power 0–0.03) and clusterProfiler (power 0) (Fig. 5).

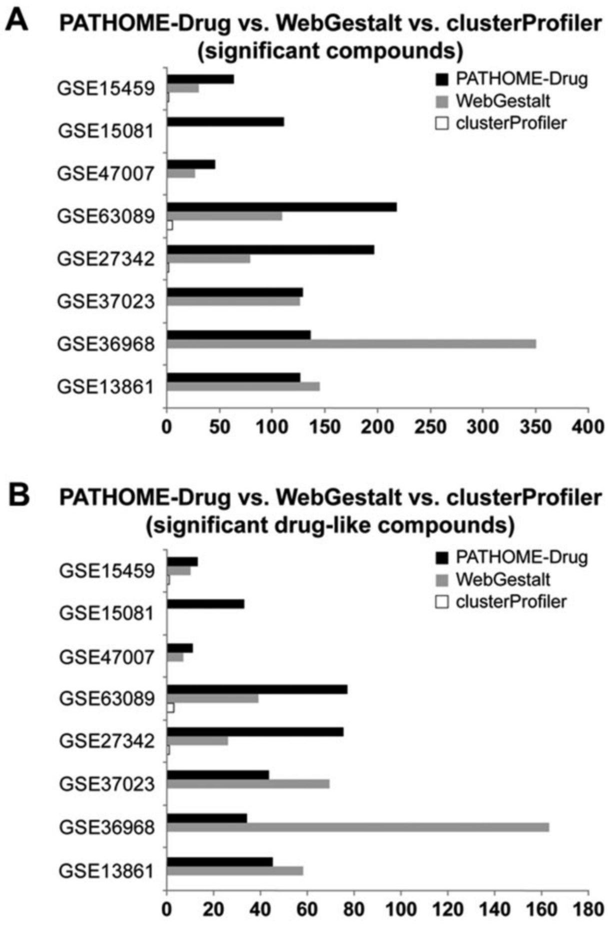

Fig. 3. Comparison of PATHOME-Drug, WebGestalt and clusterProfiler in terms of the number of reported compounds in the eight GC datasets. (**A**) We compared the number of compounds identified by the three methods from the datasets. PATHOME-Drug, in comparison to WebGestalt, reported more compounds in six of the eight datasets. PATHOME-Drug, in comparison to clusterProfiler, suggested more compounds in all the datasets. (**B**) Since compounds contained non-drug-like compounds, as well as drug-like compounds, we compared the number of 'drug-like' compounds, reported by PATHOME-Drug and WebGestalt, from the eight GC datasets. PATHOME-Drug reported more drug-like compounds, in five of the eight datasets, than WebGestalt, and more drug-like compounds, in all the datasets, than clusterProfiler. Overall, the two panels imply that our tool suggested more compounds for GC

3.5 Application of PATHOME-Drug and WebGestalt to pancreatic cancer cell lines

We next applied PATHOME-Drug and WebGestalt to dasatinib-insensitive versus -sensitive pancreatic cancer cell lines (GEO accession: GSE59357) (Chien *et al.*, 2015). As a result, 10 compounds were detected only by PATHOME-Drug, but not by WebGestalt. One such compound, regorafenib, associated with ABL1, a gene upregulated in the dasatinib-insensitive cells. Regorafenib, a known metastatic colorectal cancer therapeutic, is a multi-kinase inhibitor targeting RET, VEGFR1, FGFR1, TIE2 and ABL1 (Food and Drug Administration, 2012). Interestingly, in a recent bioactivity study (Mayer *et al.*, 2017), 2 μM regorafenib reduced cell viability in two dasatinib-insensitive pancreatic cancer cell lines, Panc1 and MiaPaca2, and also reduced tumor volume in an in-vivo xenotransplantation model of pancreatic cancer in fertilized chicken eggs (Mayer *et al.*, 2017).

3.6 Application of PATHOME-Drug to GC: dasatinib and imatinib for potential drug repositioning in GC

To address the current conundrum of limited available targeted GC drugs, we applied PATHOME-Drug to the eight GC datasets

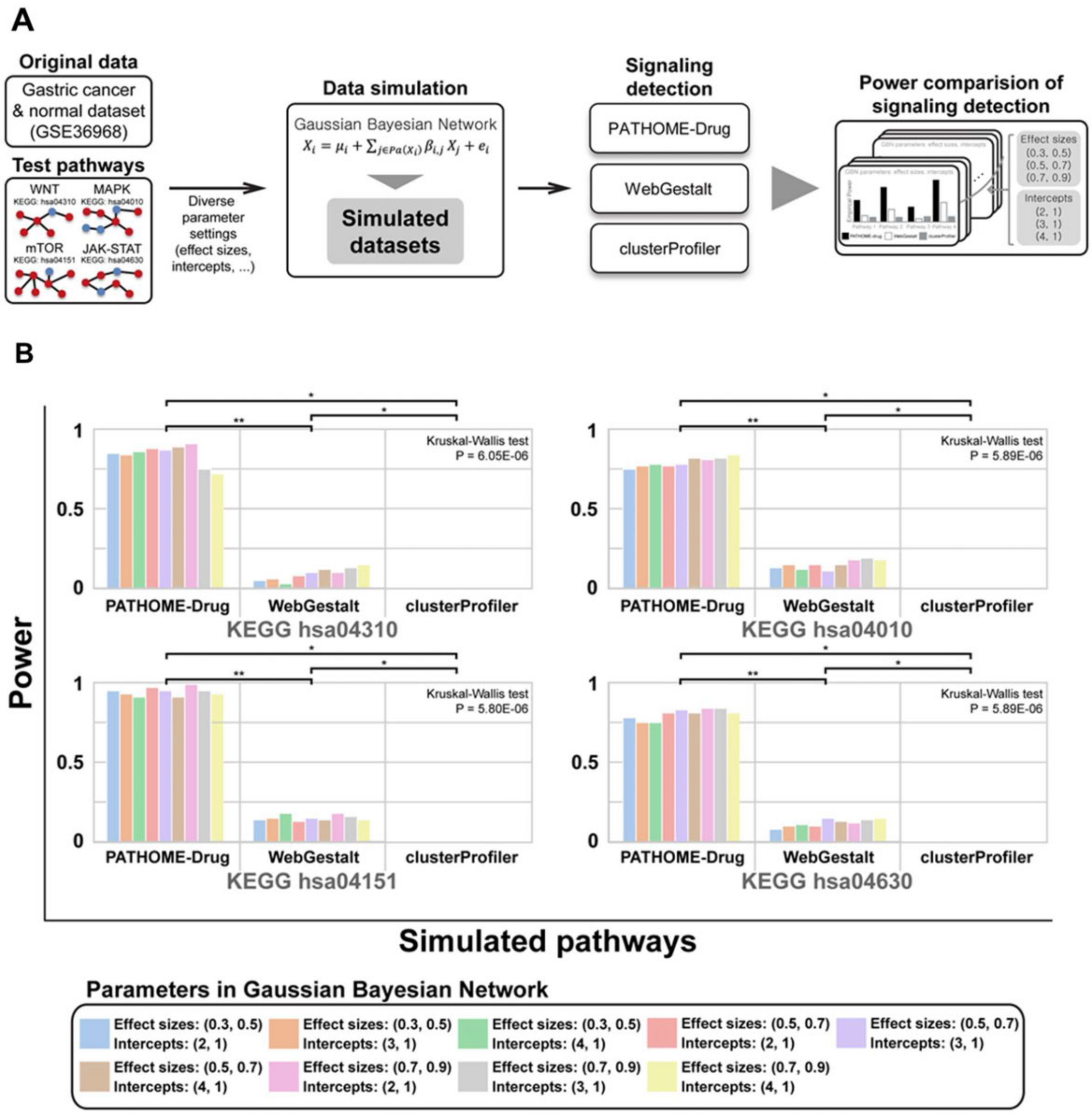

Fig. 4. Performance evaluation of signaling detection in simulated gene expression data, with considering prior network structure. (**A**) We simulated four KEGG pathways (hsa04310, hsa04010, hsa04151 and hsa04630) by using a GBN (detailed in Supplementary Method S4). This simulation was designed to reflect prior network structures for the four KEGG pathways. (**B**) Vertical and horizontal axes represent empirical power (proportion of how many times the *P*-value of the given pathway was significant for all replicates) of the signaling detection methods, PATHOME-Drug (PD), WebGestalt (WG) and clusterProfiler (CP), in each simulated pathway, respectively. According to various parameter settings of gene-wise effects (effect sizes), and sample-wise effects, (intercepts) in GBN, empirical powers of both methods were obtained. PATHOME-Drug showed better performance over WebGestalt, and clusterProfiler (n.s., not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$)

(Supplementary Table S1), obtaining the eight networks and the compounds under FDR *q*-values < 0.1. As a result, 17 compounds were enriched in ≥5 of the 8 networks (Supplementary Table S2). Of interest, 4 of the 17 compounds belonged to the tyrosine kinase inhibitor (TKI) drug class for treating several cancer types. Considering our datasets were also GC, we focused on TKIs. The attention was paid to dasatinib and imatinib, because the two inhibitors were frequently observed in the GC subpathway networks, compared to the other kinase inhibitors (Supplementary Table S2). In fact, dasatinib and imatinib were effective in GC cell and animal models (Kim *et al.*, 2019; Wang *et al.*, 2018). For imatinib and dasatinib, we collected the adjacent genes of their targets throughout the eight GC networks generated by PATHOME-Drug, revealing specific connections to EPHA3 (Supplementary Fig. S2). EPHA3 was upregulated in three (GSE37023, GSE15459, GSE63089) of the eight GC datasets. High EPHA3 expression clinically associated with GC (Nasri *et al.*, 2017), and its family proteins, EPH receptors, play context-dependent roles in tumor progression, implicating their therapeutic potential in cancer (Boyd *et al.*, 2014).

Of interest, recent inhibitor biochemical assays suggest that dasatinib has a strong binding affinity with EPHA3 (London and Gallo, 2020), strongly indicating its potential drug repositioning for GC, as reported by PATHOME-Drug. Thus, our application of PATHOME-Drug to GC encourages further study on imatinb and dasatinb, through EPHA3, for drug repurposing in GC.

4 Discussion

PATHOME-Drug is the significantly advanced version for subpathway network-based polypharmacology drug repurposing. However, neither PATHOME (Nam *et al.*, 2014), our recently described subpathway-network analysis, nor its web services, considered such repositioning, at that time. Also, in PATHOME, no further

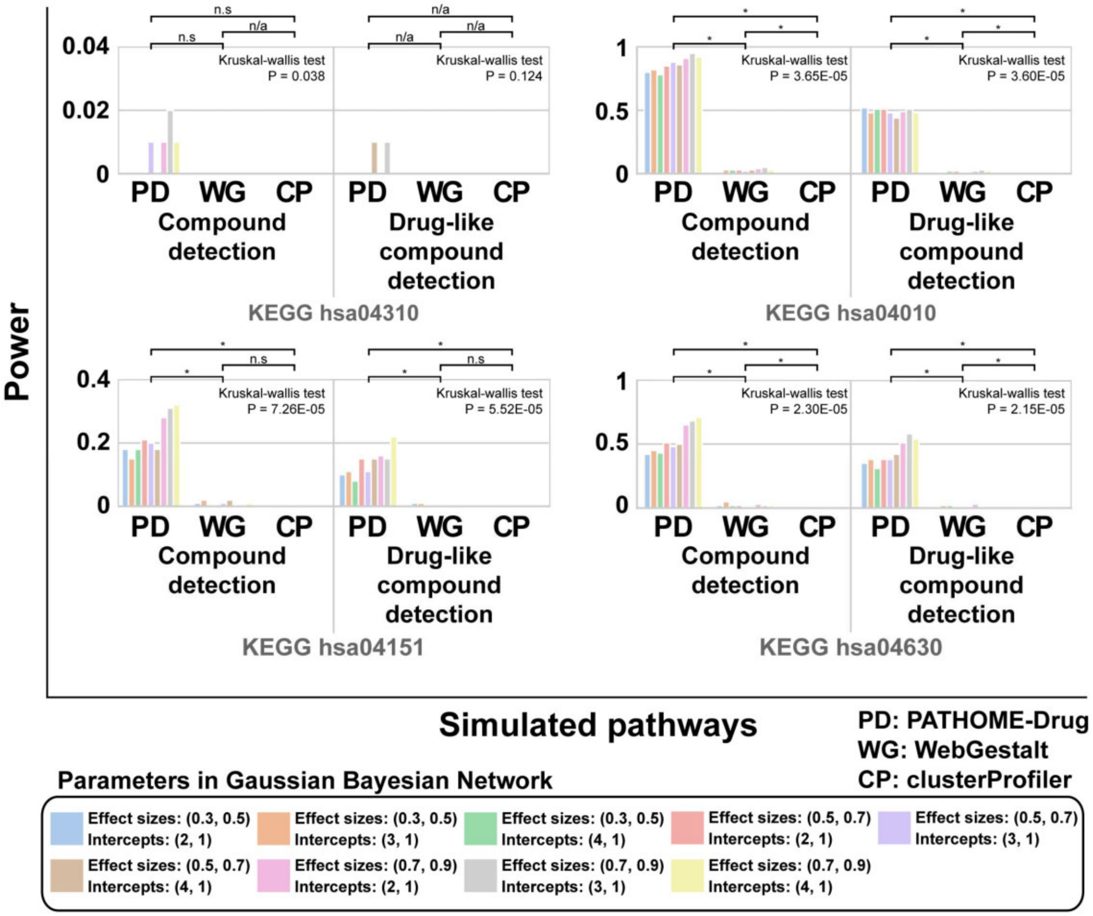

Fig. 5. Performance evaluation of drug repositioning, using synthetic gene expression data with considering prior network structure. The power values of compounds detected by PATHOME-Drug, WebGestalt and clusterProfiler, were measured in the true pathways (KEGG pathways hsa04310, hsa04010, hsa04151 and hsa04630). Power was defined as how frequently statistically significant drugs, targeting proteins in the true simulated pathway, were observed in simulated datasets. Since every compound is not drug-like, drug-like compounds were chosen, based on physicochemical properties. So, 'Compounds' on the x-axis in each panel include non-'drug-like' compounds, as well as drug-like compounds, i.e. those satisfying drug-like physicochemical properties. Each panel represents the powers of compound detection, and drug-like compound detection, in the true simulated pathway, by the two methods. The simulation datasets were equivalent to those of Figure 4. Each bar color indicates a GBN simulation parameter setting. (n.s., not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; n/a: not available)

extension to network-based drug repurposing was proposed, until PATHOME-Drug implemented new functions in its second through the fourth phases, while also publishing user-friendly web services.

We deliberately designed more biologically realistic simulated gene expression data considering regulations among gene entries, allowing us to rigorously measure performance evaluation. In the current work, PATHOME-Drug showed better performance over WebGestalt and clusterProfiler, in terms of both signaling detection and drug repurposing.

For the eight GC datasets, the gene entries reported by each method were matched to KEGG pathway networks, as prior signaling pathways. As a result, we inspected how the gene entries of each method connected to each other (Supplementary Fig. S3). Visual inspection of the projections revealed substantially different topologies. While PATHOME-Drug did not have orphan nodes (i.e. genes), WebGestalt did, largely because the first phase of PATHOME-Drug considered interactions among gene entries, and because GSA-based approaches, by nature, did not.

We also inspected the statistical significances of these performance differences (Figs 2, 4 and 5). In Figure 2, the number of false discoveries by the tools were significantly different (t-test; P-values < 0.001), and those of PATHOME-Drug were lower than those of the other tools. In Figure 4, performance differences of PATHOME-Drug versus WebGestalt versus clusterProfiler were statistically significant for detecting the four true differential GC pathways (Kruskal-Wallis test; the second column in Supplementary Table S3). In terms of compound detection (Fig. 5), the performance differences of PATHOME-Drug versus WebGestalt versus clusterProfiler were statistically significant in the four true pathways (Kruskal-Wallis test; P-values < 0.05; the second column in Supplementary Table S4). Regarding drug-like compound detection (Fig. 5), performance differences of PATHOME-Drug versus WebGestalt versus clusterProfiler were statistically significant (Kruskal-Wallis tests; the second column in Supplementary Table S5) in three of the four true pathways, except hsa04310.

Figure 3 shows some interesting results for GSE36968 and GSE15081. For GSE36968, the compounds identified by WebGestalt were three to four times more than those by PATHOME-Drug. In contrast, for GSE15081, the number of compounds identified by WebGestalt was zero. Since WebGestalt inputs statistically significant genes, we speculated that these were more DEGs, and thus reporting more compounds. So, we inspected the ratio of the number of the reported compounds to the number of the statistically significant genes (i.e. DEGs) in each dataset. Interestingly, GSE36968 showed the highest ratio (0.39) in all the datasets (Supplementary Fig. S4A), even though GSE36968 did not have the largest number of DEGs. The scatter plot of the number of the compounds versus the number of the significant genes (i.e. the number of WebGestalt input genes) confirmed this speculation, showing that GSE36968 is an outlier in the overall linear trend line (Supplementary Fig. S4B; R^2 of 0.12). In fact, excluding GSE36968, the scatter plot (Supplementary Fig. S4C) indicated the strong linear association between the number of the compounds versus the number of the significant genes with R^2 of 0.65. Thus, GSE36968 is considered as an outlier and WebGestalt is prone to provide the number of compounds proportional to the number of the input genes.

There exist limitations to our study. For performance comparisons, providing higher numbers of compounds is not necessarily a better approach. Also, providing many more compounds that obey simple drug-likeness rules (e.g. Lipinski's RO5) reveals little, since many approved drugs, long safely used, do not necessarily obey Lipinski's RO5.

In conclusion, PATHOME-Drug successfully translated a polypharmacology drug repositioning process by utilizing Big Data in network biology (Barabasi *et al.*, 2011), which has largely been considered as a 'bottleneck' for precision medicine.

Acknowledgement

The authors thank Curt Balch for editing the draft.

Funding

This research was supported by Basic Science Research Program [NRF-2016R1D1A1B03933145 to S.N.] and the Bio-Synergy Research Project [NRF-2013M3A9C4078158 to T.P.] through the National Research Foundation of Korea (NRF) funded by the Ministry of Education and the Ministry of Science, ICT and Future Planning. This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [HI16C2037010016 to T.P.].

Conflict of Interest: none declared.

Data availability

The data underlying this article are available in GitHub, at https://github.com/labnams/pathome-drug. The web service is available at http://statgen.snu.ac.kr/software/pathome.

References

Barabasi,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Bolognesi,M.L. (2019) Harnessing Polypharmacology with Medicinal Chemistry. *ACS Med. Chem. Lett.*, **10**, 273–275.

Boyd,A.W. *et al.* (2014) Therapeutic targeting of EPH receptors and their ligands. *Nat. Rev. Drug Discov.*, **13**, 39–62.

Boyle,E.I. *et al.* (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Caroli,J. *et al.* (2018) GDA, a web-based tool for Genomics and Drugs integrated analysis. *Nucleic Acids Res.*, **46**, W148–W156.

Carrella,D. *et al.* (2014) Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis. *Bioinformatics*, **30**, 1787–1788.

Chang,H.R. *et al.* (2016) HNF4alpha is a therapeutic target that links AMPK to WNT signalling in early-stage gastric cancer. *Gut*, **65**, 19–32.

Cheng,F. *et al.* (2019) A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat. Commun.*, **10**, 3476.

Chien,W. *et al.* (2015) Activation of protein phosphatase 2A tumor suppressor as potential treatment of pancreatic cancer. *Mol. Oncol.*, **9**, 889–905.

Daina,A. *et al.* (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.*, **7**, 42717.

de Leeuw,C.A. *et al.* (2016) The statistical properties of gene-set analysis. *Nat. Rev. Genet.*, **17**, 353–364.

Duan,Q. *et al.* (2016) L1000CDS(2): LINCS L1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl*, **2**, 16015.

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

Food and Drug Administration. (2012) Application number: 203085orig1s000. *Pharmacol. Rev.*

Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–690.

Hsu,C.L. *et al.* (2011) Prioritizing disease candidate genes by a gene interconnectedness-based approach. *BMC Genomics*, **12**, S25.

Jansson,P.J. *et al.* (2015) The renaissance of polypharmacology in the development of anti-cancer therapeutics: inhibition of the "Triad of Death" in cancer by Di-2-pyridylketone thiosemicarbazones. *Pharmacol. Res.*, **100**, 255–260.

Jourquin,J. *et al.* (2012) GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics*, **13**, S20.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kim,J.L. *et al.* (2019) Imatinibinduced apoptosis of gastric cancer cells is mediated by endoplasmic reticulum stress. *Oncol. Rep.*, **41**, 1616–1626.

Kim,Y.H. *et al.* (2012) AMPKalpha modulation in cancer progression: multilayer integrative analysis of the whole transcriptome in Asian gastric cancer. *Cancer Res.*, **72**, 2512–2521.

Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.

Koumakis,L. *et al.* (2016) MinePath: mining for phenotype differential sub-paths in molecular pathways. *PLoS Comput. Biol.*, **12**, e1005187.

Krauthammer,M. *et al.* (2004) Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. USA*, **101**, 15148–15153.

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

Lipinski,C.A. (2004) Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Tod. Technol.*, **1**, 337–341.

London,M. and Gallo,E. (2020) Critical role of EphA3 in cancer and current state of EphA3 drug therapeutics. *Mol. Biol. Rep.*, **47**, 5523–5533.

Martinez,V. *et al.* (2015) DrugNet: network-based drug–disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.*, **63**, 41–49.

Mayer,B. *et al.* (2017) A marginal anticancer effect of regorafenib on pancreatic carcinoma cells in vitro, ex vivo, and in vivo. *Naunyn Schmiedebergs Arch. Pharmacol.*, **390**, 1125–1134.

Nam,S. *et al.* (2014) PATHOME: an algorithm for accurately detecting differentially expressed subpathways. *Oncogene*, **33**, 4941–4951.

Nasri,B. *et al.* (2017) High expression of EphA3 (erythropoietin-producing hepatocellular A3) in gastric cancer is associated with metastasis and poor survival. *BMC Clin. Pathol.*, **17**, 8.

Navlakha,S. and Kingsford,C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, **26**, 1057–1063.

Sam,E. and Athri,P. (2019) Web-based drug repurposing tools: a survey. *Brief. Bioinf.*, **20**, 299–316.

Scutari,M. (2010) Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Softw.*, **35**, 1–22.

Segall,M.D. (2012) Multi-parameter optimization: identifying high quality compounds with a balance of properties. *Curr. Pharm. Des.*, **18**, 1292–1310.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Subramanian,A. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.e1417.

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

Thorn,C.F. *et al.* (2010) Pharmacogenomics and bioinformatics: pharmGKB. *Pharmacogenomics*, **11**, 501–505.

Vanunu,O. *et al.* (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.

Vitali,F. *et al.* (2013) Network-based target ranking for polypharmacological therapies. *J. Biomed. Inform.*, **46**, 876–881.

Wang,J. *et al.* (2017) WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.*, **45**, W130–W137.

Wang,W. *et al.* (2014) Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, **30**, 2923–2930.

Wang,X. *et al.* (2018) Dasatinib promotes TRAIL-mediated apoptosis by upregulating CHOP-dependent death receptor 5 in gastric cancer. *FEBS Open Biol.*, **8**, 732–742.

Wu,C. *et al.* (2013) Computational drug repositioning through heterogeneous network clustering. *BMC Syst. Biol.*, **7**, S6.

Yoo,M. *et al.* (2015) DSigDB: drug signatures database for gene set analysis. *Bioinformatics*, **31**, 3069–3071.

Yu,G. *et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.*, **16**, 284–287.

Zhang,Z. *et al.* (2014) MicroRNA and signaling pathways in gastric cancer. *Cancer Gene Ther.*, **21**, 305–316.