bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

KDDANet-a novel computational framework for systematic uncovering hidden gene interactions underlying known drug-disease associations

Hua Yu [1, #*], Lu Lu [2, #], Ming Chen [3], Hu Li [4], Chen Li [2,*], Jin Zhang [1, *]

[1] Center for Stem Cell and Regenerative Medicine & The First Affiliated Hospital, Department of Basic Medical Sciences, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China. Institute of Hematology, Zhejiang University, Hangzhou, Zhejiang, China.

[2] Department of Human Genetics, Zhejiang University School of Medicine, Hangzhou, China.

[3] College of Life Sciences, Zhejiang University, Hangzhou, China.

[4] Center for Individualized Medicine, Department of Molecular Pharmacology & Experimental Therapeutics, Mayo Clinic, Rochester, MN, USA.

[#] These authors contributed equally.

[*]Correspondence:yuhua200886@163.com, chenli2012@zju.edu.cn and zhgene@zju.edu.cn

Abstract

Inferring novel therapeutic indications of known drugs provides an effective method for fast-speed and low-risk drug development and disease treatment. Various computational tools have been developed to accurately predict potential associations between drugs and diseases. Nevertheless, no method has been designed to unveil pharmacological (toxicological) gene interactions underlying Known Drug-Disease Associations (KDDAs). System-level interpretation and elucidation of molecular mechanism underlying KDDAs remains a main challenge. Here, for the first time, we presented a novel and general computational framework, called KDDANet, for systematic uncovering hidden gene interactions underlying KDDAs from the perspective of complex molecular network. KDDANet effectively implemented minimum cost optimization and graph clustering on a unified flow network model. The excellent performance and general applicability of KDDANet on uncovering hidden genes underlying KDDAs were globally demonstrated by enrichment analysis of known and novel KDDA genes against two well-curated databases across broad types of diseases. Case studies on several types of diseases further highlighted that the potential value of KDDANet on revealing hidden gene interactions underlying

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

KDDAs. Particularly, it's worth noting that KDDANet can reveal the shared gene interactions underlying multiple KDDAs. For facilitating biomedical researchers to explore KDDA molecular mechanisms and guiding drug repurposing, we provided an online web server, http://47.94.193.106/kdda/index, to browse, download and analyze the predicted KDDA gene interactions. Our software and usage instruction were freely available at https://github.com/huayu1111/KDDANet/.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 Introduction

The conventional development of novel promising drugs for treating specific diseases is a time-consuming and efforts-costing process, including discovery of new chemical entities, target detection and verification, preclinical and clinical trials and so on (Paul, et al., 2010). Compared with traditional drug development, computational drug repositioning, i.e.predicting the novel indications of existing drugs, offers the possibility for safer and faster drug development because of several steps of tranditional drug development pipeline can be avoided during repurposing efforts (Ashburn and Thor, 2004; von Eichborn, et al., 2011).Various computational tools, including machine-learning and similarity-computation models, have been proposed to predict novel associations between drugs and diseases for facilitating drug repurposing (Iwata, et al., 2015; Lu and Yu, 2018). These methods can efficiently exploit and integrate multi-level omic data sources for discovery of novel drug indications and thus accelerating drug repositioning process. However, they seldomly analyze the molecular mechanisms underlying Known Drug-Disease Associations (KDDAs), which seriously hampered the development of drug repurposing and new therapies.

Theoretically, drug repurposing has been proposed based on two molecular-level facts. On one hand, complex diseases often involve multiple genetic and environmental determinants, including multi-factor driven alterations and dysregulation of a series of genes (Goh, et al., 2007). The dysfunction of these genes will propagate and perturb certain biological processes by the interactions among molecules, leading to the onset of diseases. On the other hand, one drug can exert impacts on many targets and perturb multiple biological processes (Campillos, et al., 2008; Yildirim, et al., 2007). As a result, the shared biological processes manifested in certain disease state and induced by a known drug's treatment suggest potential drug repositioning. Therefore, developing the mechanism-oriented computational tools to unveil these shared molecular interactions is of great significance for understanding

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

disease pathogenesis and guiding drug repurposing. However, few computational tools have been developed to address this need from a perspective of complex molecular network. To our knowledge, two types of methods have been intended to identify drug-gene-disease co-modules. Kutalik et al. and Chen et al. developed Ping-Pong Algorithm (PPA) and Sparse Network-regularized Partial Least Square (SNPLS) to identify co-modules related to specific cancer cell lines of NCI-60 and CGP projects, respectively (Chen and Zhang, 2016; Kutalik, et al., 2008). However, these two methods need to integrate gene-expression and drug-response data of cancer cell lines for constructing models and do not carry out prediction for other types of diseases. The second type of method has been designed to study drugs and disease with known related genes, including comCIPHER and DGPsubNet (Wang, et al., 2014; Zhao and Li, 2012). The common shortcomings of these methods are the identified co-modules including multiple drugs and diseases and thus did not uncover molecular interactions bridging individual KDDA. Moreover, comCIPHER used only 1442 genes to construct learning model, its performance and generalization capability cannot be accurately evaluated.

To fully understanding the molecular mechanism of individual KDDA and the shared molecular processes among multiple KDDAs, it is essential to find routes to bridge drug and its related disease and obtain a true picture of the roadmap of disease state cellular responses with drug administration. To this end, we designed a novel and general computational framework, called KDDANet, which used known interactome network to identify molecular interactions underlying KDDAs, including components that are otherwise hidden (Fig 1). KDDANet first built a unified flow model by integrating drugs, genes and diseases into a heterogeneous network (Fig 1a). Then, the minimum cost flow optimization, a classical algorithm which has been widely used in previous studies (da Rocha, et al., 2016; Huang, et al., 2011; Yeger-Lotem, et al., 2009), was designed and implemented to identify gene subnetwork bridging a given query drug (disease) to its related diseases (drugs) (Fig 1b-c). Finally, Markov

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CLustering (MCL) algorithm was adopted to uncover gene interaction modules bridging the query drug (diseases) and each related disease (drug) (Fig 1d).

The key advantage (strength) of the KDDANet method was its ability to uncover the hidden gene interactions and modules underlying KDDAs through implementing minimum cost optimization and graph clustering on a unified flow network model. Validation experiments against two gold standard databases, including DisGeNet (http://www.disgenet.org/) (Piñero, et al., 2016), and SMPDB (http://smpdb.ca/) (Frolkis, et al., 2009) demonstrated that our method has the excellent performance for uncovering known and novel genes bridging drugs (diseases) and their related diseases (drugs). Case studies on several types of diseases further showed that the potential value of our method on revealing hidden gene interactions of drug actions for treating specific diseases. Particularly, our results demonstrated that KDDANet can reveal the shared gene interactions underlying multiple KDDAs, which provided more valuable guides for drug repurposing. In summary, we developed an effective and universal computational tool for accurate and systematic uncovering molecular modules underlying KDDAs and thus providing novel insights into the mechanism basis of drug repurposing and disease treatments. The computed KDDA subnetworks in this study were provided in an online web server at http://47.94.193.106/kdda/index and the source code of KDDANet was freely available on https://github.com/huayu1111/KDDANet/ for facilitating biomedical researchers to explore KDDA molecular mechanisms and guiding drug repurposing.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

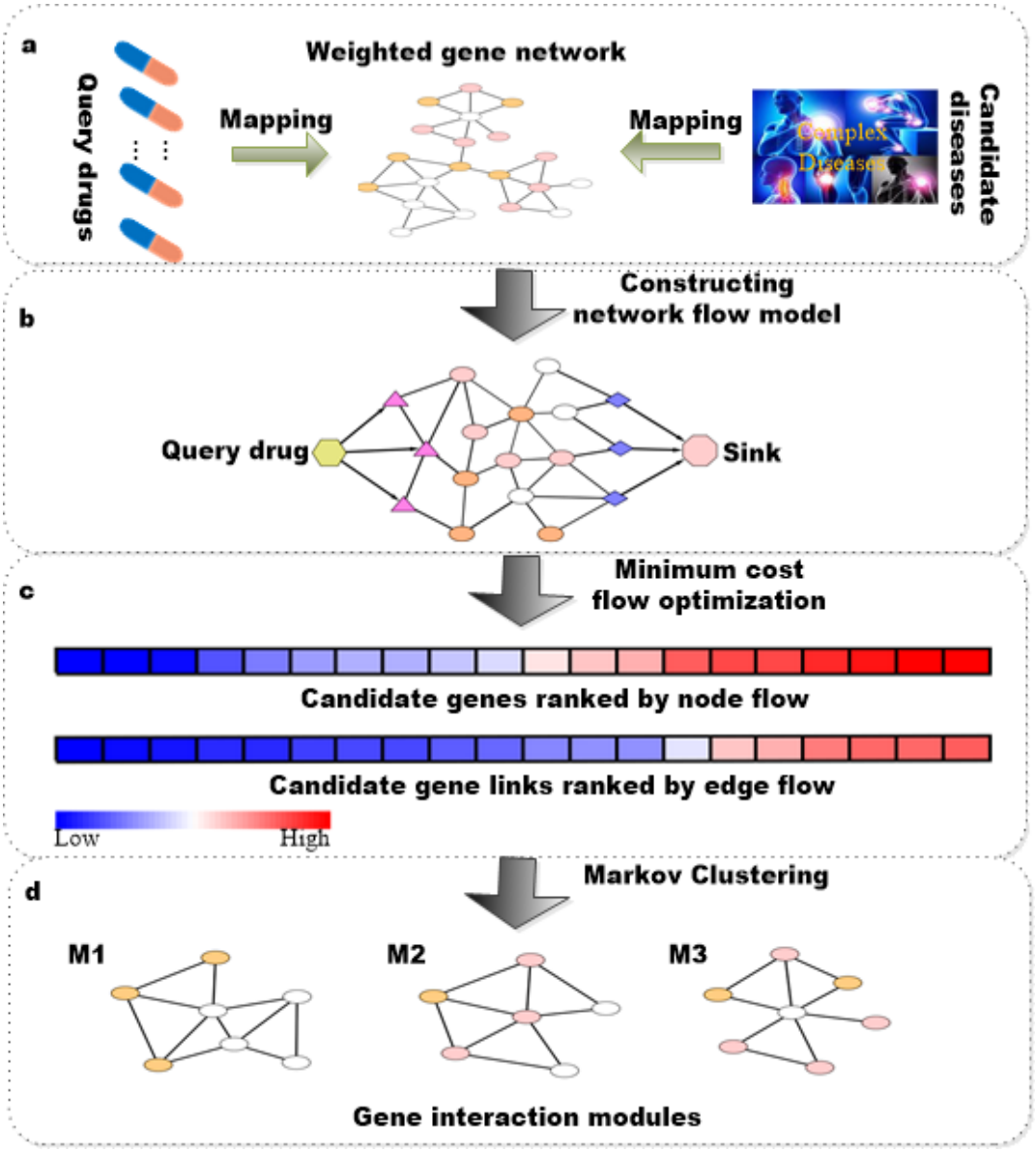


Fig 1. Schematic illustration of KDDANet computational framework. a) Mapping drugs and diseases into the weighted gene network through known drug-target associations and gene-disease associations. b) Constructing a unified flow network. c) Identifying the highest probability gene sub-network connecting query drug and candidate diseases by minimum cost flow optimization algorithm. d) Uncovering hidden gene interaction modules by Markov CLustering (MCL) algorithm.

2 Materials and Methods

2.1 Datasets

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The weighted human gene network, HumanNet, constructed by integrating multiple data sources, was used in our current study (Cho, et al., 2016), in which the nodes were represented by gene ID and connected by bidirectional edges. Drugs and targets with known interactions were obtained from DrugBank database (http://www.drugbank.ca/) on January 22, 2017. Currently, this database contained 9591 drug entries and 4661 non-redundant protein (i.e. drug target/enzyme/transporter /carrier) sequences were linked to these entries. In this study, we selected 4861 drugs with at least one known target which was contained in the weighted human gene networks for further analysis. In total, 2196 known target proteins included in the weighted human gene networks were connected to these drugs by 12014 interactions (Dataset 1). Classification of diseases and disease-related genes were extracted from the previous study (Goh, et al., 2007). We focused on 1441 diseases with at least one related gene which was included in HumanNet for our study. In total, 16712 associations link these diseases to 1521 genes existing in HumanNet (Dataset 2). The known drug-disease associations were extracted from Comparative Toxicogenomics Database (Davis, et al., 2009) (Dataset 3). In this study, we focused only on 53124 associations in which the drug has at least one target gene and the disease has at least one related gene contained in HumanNet. For simplicity and consistency, different types of drug, disease and gene nomenclatures were converted to DrugBank drug ID, OMIM disease ID and Entrez gene ID for subsequent modeling and analysis.

2.2 Construction of a unified flow network model

Our computational framework can be applied to two contexts: 1) uncovering hidden gene interactions bridging a query drug and its related disease (SDrTDi); 2) unveiling hidden gene interactions bridging a query disease and its related drug (SDiTDr). We described below the construction process of unified flow network in the first context. Overall, the unified flow network model was built by integrating the query drug and all its related diseases into the weighted gene network based upon known drug-target relations and gene-disease associations (Fig.1A). For the given query drug, we used it

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as source node (*S*) and integrated it into gene network by introducing the directed edges from it point to its target genes. For each its related disease, we mapped the disease to gene network by introducing the directed edges from its related genes point to it. To obtain a flow network, we incorporated a sink node *T* and introduced a directed edge pointing from each disease to it.

In the integrated heterogeneous network, we assigned each edge a weight value reflecting the probability that two nodes were associated. For the edges linking two genes, their weight values were directly extracted from original literature (Cho, et al., 2016). Since there was not reliable probability knowledge available between different types of nodes (i.e. drug, gene and disease), the weighting scheme was simply designed as follow.

I) For each edge connecting the query drug (*S*) and its target gene *v* or linking a gene *g* and each its related disease *d*, we given it a weight $w_{Sv} = 1$ or $w_{gd} = 1$.

II) For each edge linking each disease *d* to sink node *T*, we assigned it a weight $w_{dT} = 1$.

We further defined for each edge in this heterogeneous network a capacity value that limits the flow. For each edge connecting the query drug *S* to its target gene *v* or linking the disease *d* to sink node *T*, we assigned it a capacity c_{Sv} or c_{dT} that equal positive infinity allowing unlimited flow quantity. For other edges, we assigned them a capacity $c_{ij} = 1$. With these definitions, a unified flow network was constructed as a complex heterogeneous graph $G = (V, E)$, where *V* was the set of vertices and *E* was the set of edges. This graph included two types of edges (bidirectional and directed) and three types of nodes (drugs, genes and diseases). Each edge was assigned with a weight and a capacity. To apply our computational framework in the second context, the query disease and a set of its related drugs need to be mapped into gene network by disease-related genes and drug's target genes. After this, the weights and capacities of network edges can be set using the same method as above for constructing the

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

unified flow network model.

2.3 Minimum cost flow optimization and Markov CLustering (MCL)

The main aim of this study was to identify hidden gene interactions relating a query drug (disease) and its related diseases (drugs). This first step to solve this problem was that finding a highest probability sub-network that can connect a query drug (disease) to each its related disease (drug). This requirement could be solved using "flow algorithm", a type of computational approach that has been employed in our previous study and other labs to predict the disease-related genes and uncovering unknown biological pathways (Chen, et al., 2011; Dasika, et al., 2006; Huang, et al., 2011; Yeger-Lotem, et al., 2009; Yu, et al., 2017).

As mentioned in 2.2 Section, our method can be applied in two contexts by simply reconstructing flow network model, we described below our algorithm in the context of uncovering hidden gene interactions bridging a query drug and its related diseases. Since we want to find the topological structure of hidden gene interactions associating a query drug with a series of its related diseases, we required that flow pass from the query drug through gene network to its related diseases. We formulated our goal as a minimum cost flow optimization problem (Huang, et al., 2011; Yeger-Lotem, et al., 2009) (Fig.1B). Given the unified flow network, this problem can be expressed as a linear programming formula that minimized the total cost of the flow network while diffusing the most flow from source node to sink node. Let w_{ij}, f_{ij} and c_{ij} referred to the weight, flow and capacity from node i to node j, respectively. The linear programming formula can be written as follow.

$$\text{Minimize} \quad \sum_{i\in V, j\in V}\left(-\log\left(w_{ij}\right) * f_{ij}\right) - \left(\gamma * \sum f_{Sv}\right) \tag{1}$$

$$\text{Subject to} \quad \sum_{j\in V} f_{ij} - \sum_{j\in V} f_{ji} \quad \forall i \in V - \{S, T\} \tag{2}$$

$$f_{Sv} - \sum_{i\in D} f_{iT} = 0 \tag{3}$$

$$0 \leq f_{ij} \leq c_{ij} \tag{4}$$

Where V denoted a set of nodes included in the flow network; S denoted the source

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

node (the query drug) and v denoted it target gene. The parameter gamma (γ) controls the size and the quality of the optimized sub-network. The first component of this formula, $\sum_{i\in V, j\in V}\left(-\log\left(w_{ij}\right) * f_{ij}\right)$, ensured minimizing the network cost that given priority to obtain higher probability gene interactions, at the same time, the second component, $-(\gamma * \sum f_{Sv})$, indicated maximizing the total flow across entire network. Expression (2) guaranteed the conservation of flow quantity for each node in the gene network. Expression (3) ensured that all flow out of the source node must arrive at the sink node. Expression (4) guaranteed that the flow via each edge is non-negative and does not exceed its capacity. This optimization problem can be efficiently solved using primal simplex method provided in Mixed Integer Linear Programming (MILP) solver (http://lpsolve.sourceforge.net/). The solution $argmin_{f_{ij}>0}\sum_{i\in V, j\in V}\left(-\log\left(w_{ij}\right) * f_{ij}\right) - (\gamma * \sum f_{Sv})$ got the highest probability sub-network relating the query drug with all its related diseases. For each related disease of query drug, we searched the all possible paths linking query drug to it and used genes contained in these paths to obtain subnetwork bridging query drug and this disease. Once the subnetwork underlying query drug and each its related disease was identified, Markov Clustering (MCL) algorithm (https://micans.org/mcl/) can be adopted to discover hidden gene interaction modules by using the flow quantities through edges of subnetwork as weight values.

2.5 Performance evaluation

In most of cases, the true gene interactions underlying KDDAs were poorly understood. Consequently, there was no perfect way to assess the prediction results. Here, we systematically evaluated the performance of KDDANet based on fold enrichment of true genes underlying KDDAs against random genome background (full network). The fold of enrichment was calculated by the following formula:

$$\frac{p}{q} \times \frac{M}{N} \quad (1)$$

where, p was the number of true positive genes underlying a KDDA; q was the

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

number of inferred genes in the resulting subnetwork; *M* denoted the number of true positive genes in the gold standards and *N* represented the number of all genes in full network.

Using this method, we conducted two main evaluation experiments: I) Uncovering known genes which directly linked by query drug and it related disease in the unified flow network model; II) Uncovering novel genes which have not directly connected by query drug and its related disease in the unified flow network model, but were to be shared between disease-related genes and drug-related genes. To achieve the second, we collected a comprehensive knowledgebase of recently updated known disease-related genes and known drug-related genes as gold standards. The known disease-related genes were directly downloaded from DisGeNet database, a largest publicly available collections of genes and variants associated to human diseases (Piñero, et al., 2016), which contained 130097 associations between 14345 genes and 5391 diseases (Dataset 4). We obtained known drug-related genes by drug-pathway associations extracted from SMPDB, a well-established small molecule pathway database (Frolkis, et al., 2009; Jewison, et al., 2014). If a pathway was associated with a drug, we considered that all genes in this pathway were associated with this drug. This resulted in 15835 associations connecting 598 drugs and 810 genes (Dataset 5). The network visualization was carried out using Cytoscape (https://cytoscape.org/). KEGG pathway enrichment analysis was performed by clusterProfiler R package (Yu, et al., 2012).

3. Results

3.1 Excellent performance and general applicability of KDDANet computational framework

We first asked whether the true genes underlying Known Drug-Disease Associations (KDDAs) have been found in previous biomedical researches by checking the discrepancy between Known Drug Target Genes (KDTGs) and Known Disease

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Related Genes (KDRGs). To evaluate the discrepancy between KDTGs and KDRGs, we collected the KDTGs and KDRGs of 4861 drugs and 1441 diseases from the recently updated data source of DrugBank and DisGeNet database, respectively. Then, we analyzed the overlap ratio between KDTGs and KDRGs underlying 53124 KDDAs obtained from Comparative Toxicogenomics Database (CTD) (Davis, et al., 2009). We observed that the most gene sets demonstrated extremely low overlap ratio, with the increasing of gene number, the overlap ratio was sharply decreased (Fig 2a). We further checked the overlap percentage in different types of diseases and found the overlaps between KDTGs and KDRGs were all small for 19 disease types (Fig 2b). This discrepancy between KDTGs and KDRGs indicated that each gene set alone provides only a limited and biased view of KDDAs. Meanwhile, these outcomes also suggest that the current biomedical researches in drug development are not in well conjunction with disease pathogenesis research across broad types of diseases. To address this, we designed a novel computational framework, KDDANet, which effectively integrated network flow optimization and graph clustering analysis to bridge this gap (Fig.1, see Methods for details).

To comprehensively evaluate the capability of KDDANet framework on uncovering the hidden genes underlying KDDAs, we first checked whether the KDTGs and KDRGs were contained in the resulting subnetworks by performing enrichment analysis with a range of γ settings from 4 to 12 with a step of 1 (see Methods for details). For comparing with background enrichment, we carried out a permutation test by producing random subnetworks with the same number of genes. We found that, in SDrTDi and SDiTDr contexts, the KDTGs and KDRGs were significantly enriched in the resulting subnetworks against random subnetworks (Fig S1a-d). This indicated that KDDANet can effectively capture the genes associated with query drugs (diseases) and their related diseases (drugs). To check whether the known true KDDA genes used to construct network flow model were captured in our prediction results, we then performed the enrichment analysis of shared genes between KDTGs

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and KDRGs underlying the same KDDAs (see Methods for details). As shown in Fig 2c and 2d, we can observe that, in SDrTDi and SDiTDr contexts, the subnetworks outputted by KDDANet have obviously higher enrichment of the shared genes when compared with random permutation subnetwork. For further demonstrating whether our computational framework can uncover novel true KDDA genes which were not be shared between KDTGs and KDRGs, we collected a set of novel shared genes between drug related genes and disease related genes from recently updated SMPDB and DisGeNet database, respectively. We performed enrichment analysis and found that the resulting subnetworks have still prominently higher enrichment of these novel shared genes than random permutation subnetwork. Get together; these results suggested that our model was a powerful tool for achieving the goal of systematic uncovering hidden genes underlying KDDAs.

As the size and quality of KDDANet output subnetwork depended on a parameter γ. Higher γ values will identify more links between the source node and the sink node but with lower confidence. To obtain suitable values for γ, we run KDDANet with γ values ranging between 4 and 12 with a step of 1. For each of the output subnetwork, we computed the percentage of input, namely KDTGs and KDRGs, that were incorporated into the network, as well as the percentage of low probability edges with weight smaller than 0.3. As shown in Fig S1e and Fig S1f, we observed that the percentages for KDTGs, KDRGs and low probability edges contained in resulting subnetworks became stable when $\gamma = 6$ and 8 for SDrTDi and SDiTDr, respectively. We therefore selected $\gamma = 6$ for SDrTDi and $\gamma = 8$ for SDiTDr as they were the minimal values with which a significant fraction of the input was incorporated while the percentage of low probability edges remained small in respective contexts. After confirming the excellent performance and obtaining the optimal γ value of KDDANet framework by global and systematic analysis, we further determined whether KDDANet has the general application value across a variety of disease types. Using the same enrichment analysis strategy on 19 different types of diseases, we found that

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

KDDANet can make effective capturing of true genes underlying KDDAs for all 19 types of diseases (Fig 2g-h and Fig S1g-h, Dataset 6 and Dataset 7). Collectively, these results indicated that KDDANet was an excellent and general computational tool which can be applied to uncover hidden genes underlying KDDAs across broad types of diseases.

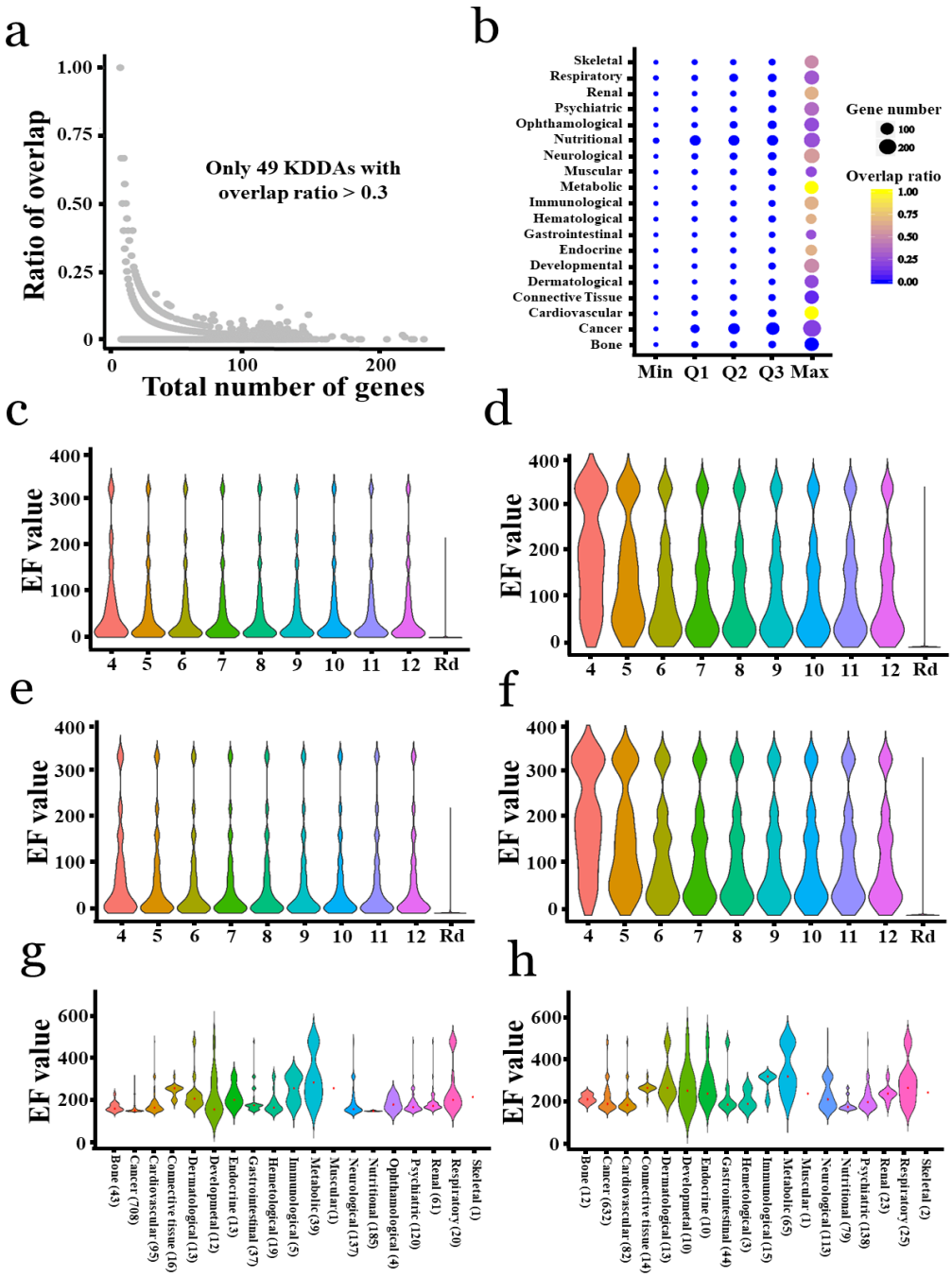


Fig 2. Excellent performance and general applicability of KDDANet method. a) Scatter plot demonstrated the distribution of overlap ratio between KDTGs and KDRGs underlying the same

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

KDDAs; The x-axis denoted the total number of KDTGs and KDRGs underlying the same KDDAs. b) Colored scatter plot demonstrated the Minimum (Min), 25th Quantile (Q1), 50th Quantile (Q2), 75th Quantile (Q3) and Maximum (Max) of overlap ratio distribution in different types of diseases; The size and color of point denoted the gene number and overlap ratio, respectively; c) Violin plot demonstrated the enrichment of known shared genes between KDTGs and KDRGs underlying same KDDAs with different γ settings in SDrTDi context, using random permutation as control (Rd); d) Similar to c) demonstrating the enrichment of known shared genes in SDiTDr context; e) Similar to c), demonstrating the enrichment of novel shared genes between drug related genes collected from SMPDB database and disease related genes obtained from DisGeNet database; f) Similar to d), demonstrating novel shared genes between drug related genes collected from SMPDB database and disease related genes obtained from DisGeNet database; g) Violin plot demonstrated the enrichment of known shared genes between KDTGs and KDRGs underlying the same KDDAs across different types of diseases in SDrTDi context; Values in parentheses denoted the number of KDDAs. h) Similar to g), demonstrating the enrichment of known shared genes between KDTGs and KDRGs in SDiTDr context.

3.2 Global functional analysis and detailed case studies highlight that the potential value of KDDANet for revealing hidden gene interactions underlying individual KDDA

The comprehensive analysis above demonstrated that KDDANet was powerful for predicting the possible gene links underlying KDDAs. Here, we further evaluated the rationality of functions of the KDDA gene interaction subnetworks by carrying out a global enrichment of all predicted KDDA subnetworks against 53 classical KEGG pathways. We calculated Enrichment Score (ES) for each pathway in all 19 types of diseases (Fig S2). The obtained enrichment results can be validated by existing scientific knowledge. For example, we observed that p53 signaling pathway has relative higher ES value in cancer when comparing with other pathways for both SDrTDi and SDiTDr resulting modules (see Fig S2). For neurological disease, we found that the notch signaling pathway was significantly enriched in both SDrTDi and

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SDiTDr resulting modules (Fig S2). This was consistent with the previous discoveries that demonstrated the association of notch-related pathways with neurological disorders and the potential of agents target notch signaling as therapeutic interventions for several different central nervous system disorders (Lathia, et al., 2008). Interestingly, we found that ophthamological, immunological, endocrine, developmental and cardiovascular diseases have consistent less enrichment on all 53 known KEGG pathways than other types of diseases (Fig S2). This indicated that the known drugs developed to cure these types of diseases cannot be well explained by these known pathways and the uncovered molecular interactions underlying KDDAs in these types of diseases is great value for further experimental studies.

In addition to global function analysis KDDA genes, we performed the detailed case studies to further demonstrate the potential merits of KDDANet for uncovering gene interactions underlying individual KDDA. For simplicity, we randomly selected the predicted subnetworks of 4 KDDAs including DB00489 (sotalol)-176807 (prostate cancer) association, DB01109 (heparin)-601665 (obesity) association, DB03147 (flavin adenine dinucleotide)-114480 (breast cancer) association and DB01022 (phylloquinone)-104300 (Alzheimer disease) association for detailed analysis. Sotalol was normally used to treat life threatening ventricular arrhytmias and maintained normal sinus rhythm in patients with atrial fibrillation. It has been reported that sotalol was associated with decreased prostate cancer risk (Kaapu, et al., 2015). For sotalol-prostate cancer association, we predicted a subnetwork consisting of 31 genes and 28 links (Fig 3a). All three known target genes, KCNH2, ADRB1 and ADRB2, were included in this subnetwork. Meanwhile, this subnetwork also captured 12 disease-related genes. Among genes included in this subnetwork, KCNH2, PRKACA and PRKAR1A were the novel shared genes between recently updated drug-related genes and disease-related genes, which were not used as KDTGs and KDRGs in our flow network model. The top 10 enriched KEGG terms were demonstrated in Fig 3b. As expected, PI3K-Akt signaling pathway, MAPK signaling pathway and FoxO signaling

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pathway were the classic signaling pathways related to cancer formation and development. The associations between EGFR tyrosine kinase inhibitor resistance, Relaxin signaling pathway, AGE-RAGE signaling pathway and prostate cancer have also been widely investigated and reported in the previous works (Bao, et al., 2015; Neschadim, et al., 2015; Ozvegy-Laczka, et al., 2005). Moreover, autophagy and focal adhesion were two widely observed biological processes in cancer (Eke and Cordes, 2015; Farrow, et al., 2014). By applying MCL with default parameters, this subnetwork was further decomposed into 3 gene modules, M1, M2 and M3. Obviously, the enriched KEGG terms of M1 and M3 were closely related to cancer formation and development (Fig 3c). Surprisingly, M2 was not enriched to any KEGG term, indicating that it might be a novel module explaining this association. DB01022 (phylloquinone)-104300 (Alzheimer disease) association has been reported in previous studies (Alisi, et al., 2019). A subnetwork including 46 genes and 44 links were uncovered for this association (Fig 3d). All two known target genes of phylloquinone and 18 Alzheimer disease related genes were identified in this subnetwork. Interestingly, for this association, two separated gene modules were detected by two known targets of phylloquinone. The enriched KEGG terms of these two gene modules demonstrated strong closeness with Alzheimer disease (Fig 3e-f).

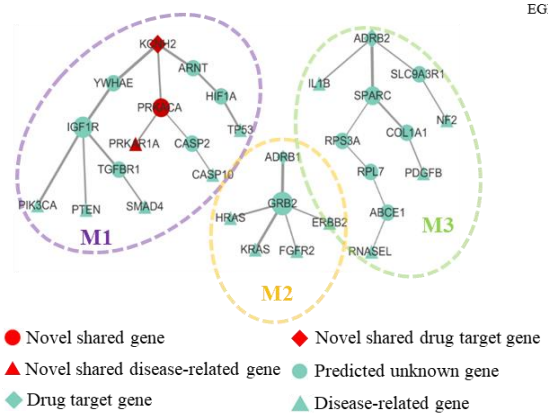
The association between DB01109 (heparin) and 601665 (obesity) was inferred by multiple genes as described in CTD database. For this association, KDDANet predicted a subnetwork containing 168 edges connecting 169 genes (Fig S3a). We found that all 3 known drug target genes and 67 disease-related genes were captured in this subnetwork. Particularly, 9 genes were the novel shared genes between recently updated drug-related genes and disease-related genes which were not used as KDTGs and KDRGs in our flow network model. Interestingly, 3 genes used to infer this KDDA, including ARK1, PARP1 and TNF, were also effectively captured in the resulting subnetwork. The top 10 enriched functions of this subnetwork were demonstrated in Fig S3b. As expected, Insulin resistance, Type II diabetes mellitus,

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

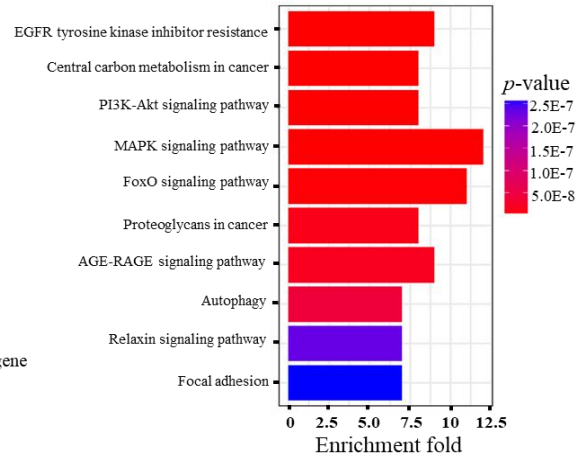
Insulin signaling pathway and Adipocytokine signaling pathway were the frequently reported molecular processes associated with obesity (Kahn and Flier, 2000; Taylor and Macqueen, 2010). In addition, Proteoglycans, Lipolysis and AMPK signaling pathway were also highly correlated with Insulin resistance (Langin, et al., 2005; Olsson, et al., 2001; Viollet, et al., 2009). For some KEGG terms, though we did not observe their associations with obesity, they might be the true biological processes underlying this association. We delineated this subnetwork into gene modules. The top 3 gene modules and their enriched functions were demonstrated in Fig S3a and Fig S3c. The major enriched functions of these modules were related to obesity. For DB03147 (flavin adenine dinucleotide)-114480 (breast cancer) association, we obtained a subnetwork consisting of 148 genes and 137 interactions (Fig S3d). As expected, CAT, a gene used to infer this association, was included in this subnetwork. In consistent with this, among 86 targets of DB03147, 25 were contained in this subnetwork. Meanwhile, 44 breast cancer related genes were also captured in this subnetwork. Further analysis demonstrated that the mainly enriched functions of this subnetwork and inferred gene modules were closely related to cellular basis of breast cancer (Fig S3e-f). All these results above indicated that the molecular basis of KDDAs can be revealed by our proposed approach.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a



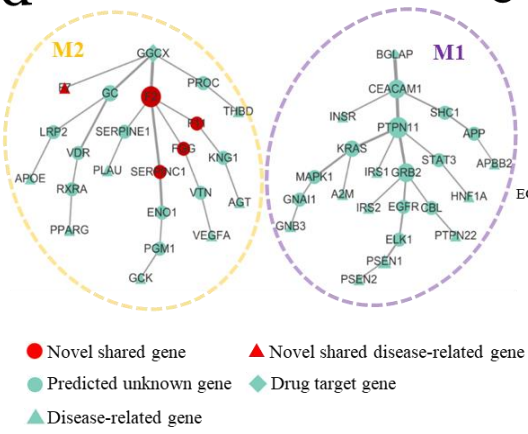
b



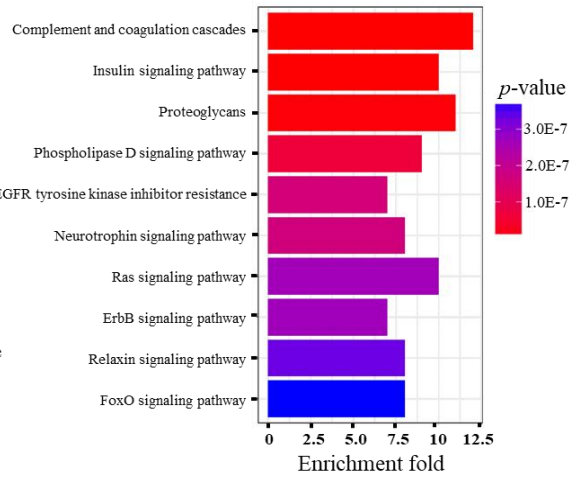
c

Module ID	Representative KEGG terms
M1	FoxO signaling pathway; Ras signaling pathway; PI3K-Akt signaling pathway; MAPK signaling pathway
M3	B cell receptor signaling pathway; ErbB signaling pathway

d



e



f

Module ID	Representative KEGG terms
M1	Phospholipase D signaling pathway; Chemokine signaling pathway; Prolactin signaling pathway; Estrogen signaling pathway
M2	Complement and coagulation cascades; Cholesterol metabolism

Fig 3. Case studies on DB00489 (Flavin adenine dinucleotide)-176807 (Prostate Cancer) association and DB01022 (Phylloquinone)-104300 (Alzheimer disease) association. a) Uncovered gene interaction subnetwork for DB00489-176807 association in SDrTDi context; b) Top 10 enriched KEGG terms of subnetwork genes underlying DB00489-176807 association; c) Representatively enriched KEGG terms of 3 identified gene modules underlying DB00489-

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

176807 association; d) Uncovered gene interaction subnetwork underlying DB01022-104300 association in SDrTDi context; e) Enriched KEGG terms of subnetwork genes underlying DB01022-104300 association; f) Representatively Enriched KEGG terms of two uncovered gene modules underlying DB01022-104300 association.

3.3 KDDANet uncovers the shared gene interactions underlying multiple KDDAs

Analysis on above fully demonstrated that KDDANet can uncover hidden gene interactions underlying individual KDDA. Here, we further asked whether KDDANet can reveal the shared gene interactions underlying multiple KDDAs. We answered this from two aspects: I) Multiple Diseases associating with One Drug (MDiODr); II) Multiple Drugs associating with One Disease (MDrODi). Considering the practical merits of first one analysis, we required multiple diseases belongs to same type of diseases. To perform these two evaluations, we produced meta-subnetworks by combining multiple KDDA subnetworks for 12386 MDiODr combinations and 773 MDrODi combinations produced in SDrTDi context. The weight of an edge in the meta-subnetwork was defined as the number of KDDA subnetworks containing this edge divided by the total number of KDDA subnetworks. The higher weight value of a link in meta-subnetwork indicated more conservation and commonality. Thus, we used the weight value to evaluate the capacity of KDDANet in unveiling the shared gene interactions underlying multiple KDDAs. We carried out a permutation test by producing random meta-subnetworks with the same number of edges for comparing with random background. As shown in Fig 4a-d, the weights of KDDANet meta-subnetworks were significantly higher than random meta-subnetworks across different types of diseases. This indicated the gene interactions tend to be shared in KDDANet meta-subnetworks than random background. We also conducted the same analysis in SDiTDr context for 12189 MDiODr combinations and 773 MDrODi combinations and obtained the similar results (Fig S4a-d). All these results indicated that KDDANet can effectively uncover the shared gene interactions underlying multiple KDDAs.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next, we presented some examples for intuitively describing the capability of KDDANet in identifying shared gene interactions underlying multiple KDDAs. For MDiODr, we selected two cases: I) DB00370 (mirtazapine)-cancer associations and II) DB00392 (profenamine)-neurological diseases associations, for detailed descriptions. Considering the biomedical significance and conciseness, we selected the edges of meta-subnetwork with weights larger than 0.6 for presentation. As recorded in CTD database, mirtazapine was associated with 9 cancers, including Colorectal Neoplasms, Breast Neoplasms, Neuroblastoma, Glioma, Urinary Bladder Neoplasms, Stomach Neoplasms, Esophageal Neoplasms, Lung Neoplasms and Prostatic Neoplasms. Fig 4e showed the shared meta-subnetwork including 25 genes and 24 edges. Among these genes, HTR2C, ADRA2A, OPRK1 and HRH1 were mirtazapine's target genes and TP53, BRAF, KRAS and EGFR were cancer related genes. A KEGG term enrichment demonstrated that these genes play roles in various cancers, including 4 cancers associated with mirtazapine as described on above (Fig 4f). Consistent with this, the enriched pathways, such as Sphingolipid signaling pathway, were closely related to cancer formation and development (Ogretmen, 2018). As reported in CTD database, profenamine has associations with three neurological diseases, including Parkinsonian Disorders, Multiple Sclerosis and Alzheimer Disease. The shared meta-subnetwork contained 10 edges linking 13 genes with 3 profenamine's target genes, CHRM1, CHRM2 and GRIN3A (Fig 4g). The top 10 enriched KEGG terms were demonstrated in Fig 4h. A majority of enriched KEGG terms, such as Cholinergic synapse and Glutamatergic synapses, have been reported dysfunction in neurological disorders and diseases (Moretto, et al., 2018; Tata, et al., 2014)

For MDrODi, we observed that the shared gene interactions were significant less than MDiODr (Fig 4a-d and Fig S4a-d). Nevertheless, we also observed some interesting cases. For example, GRACILE syndrome, a metabolic disease, was associated with 13 drugs as recorded in CTD database. A shared meta-subnetwork including 9 genes

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and 8 edges were obtained for this disease (Fig S4e). This meta-subnetwork contained 6 genes which were known related to this disease. All 9 genes in this subnetwork were involved in Oxidative phosphorylation. This was expected as GRACILE syndrome was a fatal inherited disorder caused by a mutation in an oxidative phosphorylation related gene, BCS1L, necessary for providing cells with energy. Another example was the Keratoconus, an ophthamological disease, was associated with 3 drugs, acetaminophen, valproic acid and theophylline. Keratoconus and these drugs shared a meta-subnetwork consisting 15 genes and 11 edges (Fig S4f). This meta-subnetwork included 5 Keratoconus related genes, i.e., PIKFYVE, TACSTD2, TGFBI, KERA and COL8A2, and both valproic acid and theophylline's target gene, HDAC2. It is also expected that SNW1 and HDAC2 were contained in this meta-subnetwork as they were involved in notch signaling pathway which is down-regulated in keratoconus (You, et al., 2018). Interestingly, consistent with the associations between collagen genes and keratoconus (Bykhovskaya, et al., 2016), a novel collagens coding gene, COL1A1, was captured in this meta-subnetwork. Taken together, these shared gene interactions provided more valuable and general guides for drug repositioning and disease treatment as they characterized the common molecular mechanism among multiple KDDAs.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

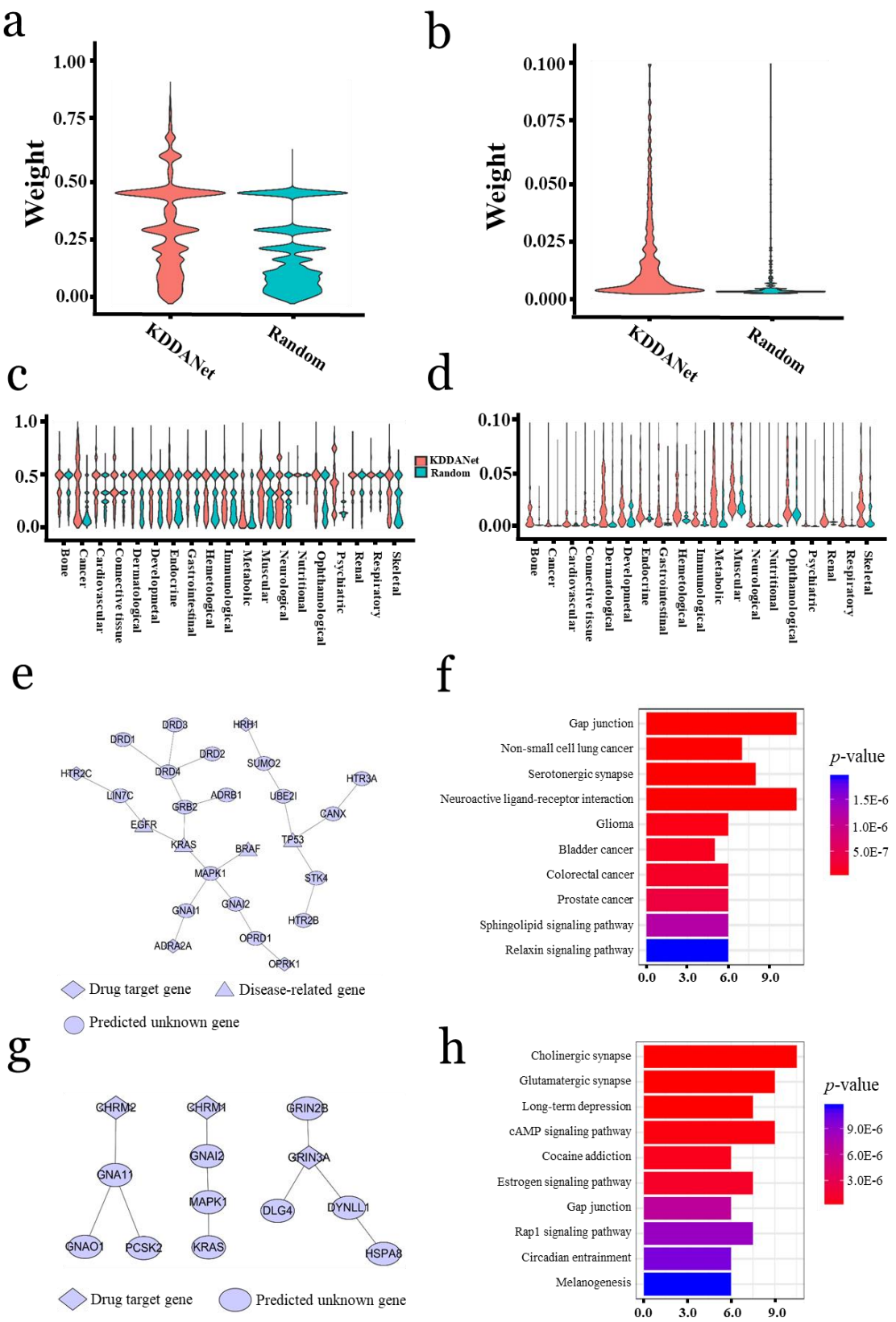


Fig 4. KDDANet uncovered the shared gene interactions underlying multiple KDDAs.

a) Violin plot demonstrates the distributions of weight values in KDDANet meta-subnetwork and random meta-subnetwork for MDiODr in SDrTDi context. b) Violin plot demonstrates the

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distributions of weight values in KDDANet meta-subnetwork and random meta-subnetwork for MDrODi in SDrTDi context. c) Violin plot demonstrates the distributions of weight values in KDDANet meta-subnetwork and random meta-subnetwork for MDiODr in SDrTDi context across different types of diseases; d) Violin plot demonstrates the distributions of weight values in KDDANet meta-subnetwork and random meta-subnetwork for MDrODi in SDrTDi context across different types of diseases; e) Uncovered gene interaction subnetwork for DB00370 -cancer associations in SDrTDi context; f) Top 10 enriched KEGG terms of shared meta-subnetwork genes underlying DB00370-cancer association; g) Uncovered gene interaction subnetwork for DB00392-neurological diseases associations in SDrTDi context; h) Top 10 enriched KEGG terms of shared meta-subnetwork genes underlying DB00392-neurological diseases associations.

3.4 A web server for querying and browsing gene interactions underlying KDDAs

To help biomedical researchers to understand the pathogenesis of disease and promote drug development, we developed an online web server, http://47.94.193.106/kdda/subnet, for facilitating the researchers to browsing and analyzing hidden gene interactions underlying KDDAs. Our website was divided into 3 parts (Fig 5a): I) querying and browsing gene interactions underlying individual KDDA (iKDDA); II) querying and browsing shared gene interactions underlying multiple KDDAs connecting multiple diseases to one drug (MDiODr) and III) querying and browsing shared gene interactions underlying multiple KDDAs connecting multiple drugs to one disease (MDrODi). For Part I, hidden gene interactions underlying 52878 and 51745 KDDAs for SDrTDi and SDiTDr were provided for visualizing and downloading, respectively (Fig 5b). Users can input a query DrugBank drug id or drug name and its associated OMIM disease id or disease name. When users selected the corresponding optimization context (SDrTDi or SDiTDr) and clicked the search button, the subnetwork underlying this KDDA was displayed on a panel. In the panel, the size of node was proportional to the node degree, the thickness of edge denoted the flow quantity and the known drug target

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genes and disease related genes were filled with different colors. Users can click any gene node to skip the NCBI Gene website to learn about the knowledge of this gene. The content of subnetwork underlying this KDDA can directly be downloaded for performing downstream analysis, such as module identification and function analysis.

For Part II, 12386 and 12189 shared meta-subnetworks were contained in our web server for SDrTDi and SDiTDr, respectively. When user input the query drug and disease type, all its associated diseases under this disease type and the shared meta-subnetwork were instantly visualized. User can click any related disease to link to OMIM database for acquainting disease related knowledge. To facilitate visualizing and browsing, we selected edges with weight larger than 0.1 for displaying. The weight of interaction of meta-subnetwork was characterized by thickness of edge. Similar to Part I, drug target genes and disease-related genes were displayed with different colors. Any genes in the shared meta-subnetwork can be clicked to link to NCBI Gene website for obtaining related information. The shared meta-subnetwork can be also downloaded for user specific downstream analysis. (Fig 5d). For Part III, 773 shared gene interaction meta-subnetworks were contained in our web server (Fig 5c). The functions of this part were similar to Part II. Finally, we designed a bulk download interface for facilitating user to obtain all subnetworks mentioned above. Overall, we provided a simply and easy-to-use web server for exploring the molecular mechanism underlying KDDAs and thus help researchers to understand the pathogenesis of disease, accelerate drug repurposing and disease treatment.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

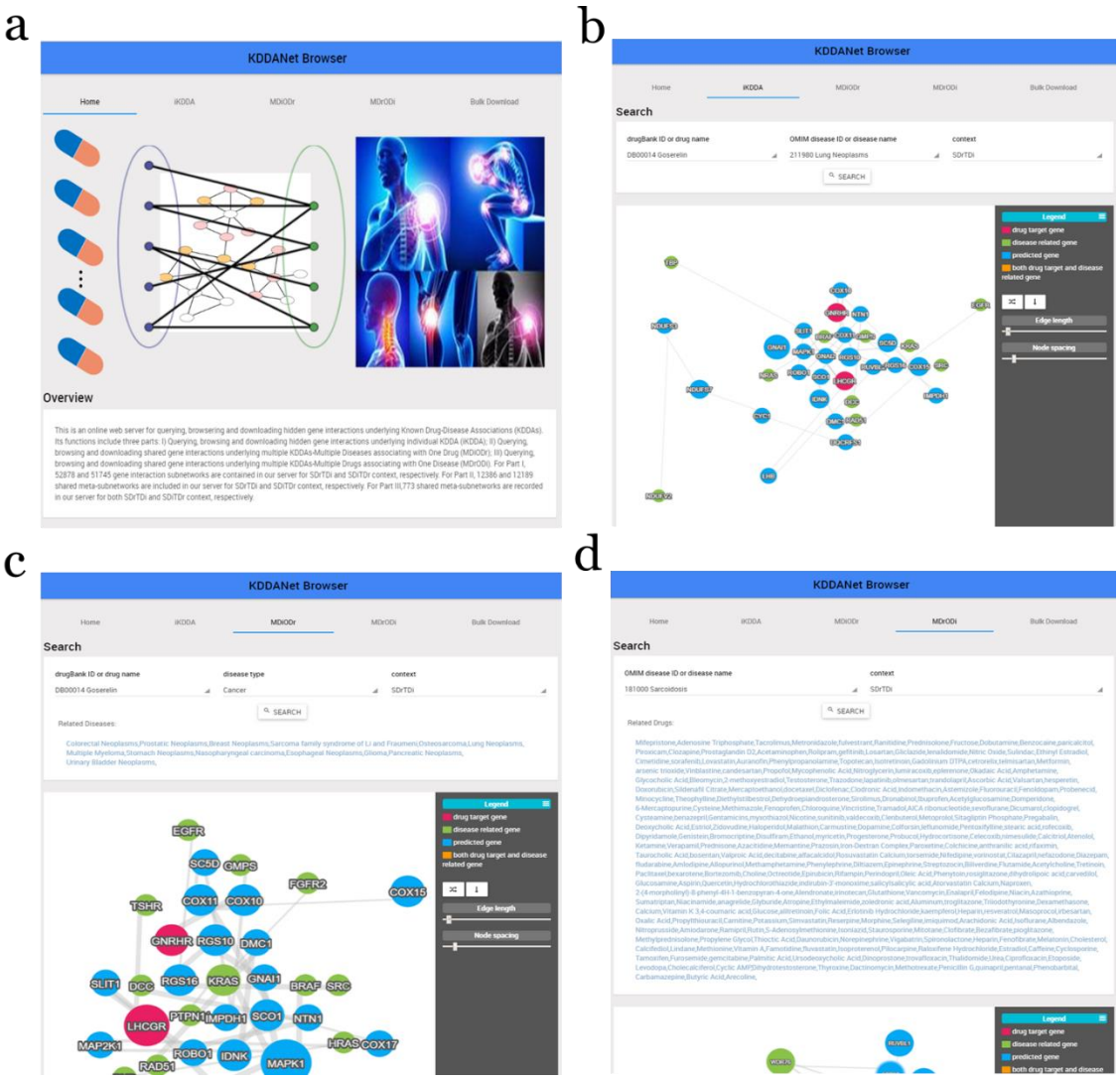


Fig 5. An online web server for visualizing gene interactions underlying KDDAs. a) Home page of our web server; b) Web page for visualizing hidden gene interaction subnetwork for iKDDA; c) Web page for visualizing the shared meta-subnetwork for MDiODr; d) Web page for visualizing the shared meta-subnetwork for MDrODI.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4. Discussion

For drug repurposing, various computational methods have been developed to infer novel drug-disease associations (Iwata, et al., 2015; Lu and Yu, 2018). However, the molecular mechanisms underlying KDDAs have been still not well explored. This was confirmed by the extremely low overlap ratio between KDTGs and KDRGs. The discordance between KDTGs and KDRGs also suggested the existence of hidden gene interactions underlying KDDAs. Therefore, uncovering the hidden gene interactions underlying KDDAs became an important scientific problem for guide drug repurposing. Minimum cost network flow optimization is a classical graph theory algorithm, has been adopted to uncover unknown molecular processes of cell responses to biotic/abiotic stresses and candidate disease genes (Huang, et al., 2011; Yeger-Lotem, et al., 2009). It is worth mentioned that Hu Li, Ph.D., and his research team developed a cutting-edge method, called NetDecoder, which used minimum cost network flow optimization to dissect context-specific biological subnetwork (da Rocha, et al., 2016). This method enables researchers to uncover context-dependent drug targets, which has considerable application value for personalized medicine. The successful practices of minimum cost network flow optimization in these studies implicated that it is helpful in biological and medical researches. Inspired by these works, we integrated minimum cost network flow optimization and graph clustering algorithm and developed a novel computational framework, called KDDANet, to reveal the hidden gene interactions and modules underlying KDDAs. KDDANet allowed for a global and systematic exploration the molecular mechanisms underlying known drug-disease associations, which can effectively complement to the existing computational methods for novel drug-disease association prediction, for example, a published method in our previous study (Lu and Yu, 2018). We applied KDDANet to systematically unravel the hidden gene interactions underlying 53124 KDDAs. For SDrTDi and SDiTDr context, the enrichment results of known and novel KDDA genes of 52878 and 51745 KDDA subnetworks fully demonstrated the powerful prediction ability and general applicability of our computational framework. In order

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to intuitively demonstrate that gene interactions identified by KDDANet were highly reliable and useful, we conducted global KEGG pathway enrichment analysis and detailed case studies to further illustrate this. An important capability of KDDANet method was that it can reveal the shared gene interaction underlying multiple KDDAs as showed in our results. This can provide more valuable guides for drug repositioning and disease treatment as the shared gene interactions were closely related to the molecular basis of drug repurposing.

As described in Results section, the global gene interaction roadmaps obtained by KDDANet demonstrated the power of integrative approaches to illuminate underexplored molecular processes underlying KDDAs. However, there are some limitations that influence the KDDANet application in precision medicine and personalized care. Firstly, the gene interactome network used in our flow network model was static and did not capture network rewiring events under different circumstances. Indeed, the true biological network are highly dynamic and undergo continuous rewiring events (da Rocha, et al., 2016). For a KDDA in specific context, constructing a context-specific interactome network might be more appropriate for accurate uncovering hidden KDDA gene interactions. Secondly, the used gene interactome did not contain enough information of gene interaction. In further studies, we will integrate other non-coding genome elements, especially long non-coding RNA and microRNA for constructing a comprehensive gene interactome network in a context-dependent method. In addition, integrating biological networks from other omics layers, such as epigenetics, can further enhance the accuracy of KDDANet in discovering KDDA subnetworks and the associated key genes and thus help us better understand KDDA at multi-omics levels. Indeed, the ability of KDDANet to analyze large-scale heterogeneous interactome data containing tens of thousands of nodes and edges make it can well be suited to analyzing the accumulating data from ‘multi-omics’ technologies and biomedical research. Thirdly, KDDANet cannot carry out prediction for drug-disease associations when a drug’s target genes were unknown or

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

when a disease has not been related to a set of genes. For this, we plan to calculate the similarity scores between drugs and the similarity scores between diseases, and then integrate drugs without any target genes and diseases without any related genes to our flow network model by similarity scores. Finally, due to the limitations of current experimental conditions, we were not able to do wet experiments to verify the predictions. Once the experimental conditions allowed, we will focus on some specific diseases and conducted molecular experiments, such as pooled CRISPR screen combined with single cell transcriptome sequencing, to systematically verify the predicted KDDA gene interactions.

In summary, we presented a novel network-based computational framework that has broad utility and application value in biomedical studies. Our method allowed the researchers to explore hidden gene interactions underlying KDDAs. Although some challenges are still existing for enhancing the prediction accuracy of our method, the inferred KDDA subnetworks and related genes discovered can serve as roadmaps for guiding drug repurposing and disease treatment. Insights learned from our predicted subnetworks will enable to help researchers to design novel drugs to reverse disease phenotypes via targeting hub genes in the KDDA subnetwork. For helping drug repurposing in wet lab, we provided an online web server to facilitate researchers to browse and analyze the predicted KDDA gene interaction subnetworks. Our online web server was flexible allowing researchers to query and visualize the predicted subnetwork for KDDAs consist of several thousands of drugs and several hundreds of diseases extracted from CTD database. As the huge amount of biological data was becoming available in the biomedical study, we believed that the predictions of KDDANet could be more reliable and useful.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Alisi, L., *et al.* (2019) The Relationships Between Vitamin K and Cognition: A Review of Current Evidence, *Frontiers in Neurology*, **10**, 239.

Ashburn, T.T. and Thor, K.B. (2004) Drug repositioning: Identifying and developing new uses for existing drugs, *Nature Reviews Drug Discovery*, **3**, 673-683.

Bao, J.M., *et al.* (2015) AGE/RAGE/Akt pathway contributes to prostate cancer cell proliferation by promoting Rb phosphorylation and degradation, *American Journal of Cancer Research*, **5**, 1741-1750.

Bykhovskaya, Y., Margines, B. and Rabinowitz, Y.S. (2016) Genetics in Keratoconus: where are we?, *Eye and Vision*, **3**, 16.

Campillos, M., *et al.* (2008) Drug target identification using side-effect similarity, *Science*, **321**, 263-266.

Chen, J. and Zhang, S. (2016) Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data, *Bioinformatics*, **32**, 1724-1732.

Chen, Y., Jiang, T. and Jiang, R. (2011) Uncover disease genes by maximizing information flow in the phenome-interactome network, *Bioinformatics*, **27**, i167.

Cho, A., *et al.* (2016) MUFFINN: cancer gene discovery via network analysis of somatic mutation data, *Genome Biology*, **17**.

da Rocha, E.L., et al. (2016) NetDecoder: a network biology platform that decodes context-specific biological networks and gene activities, *Nucleic Acids Research*, **44**, e100.

Dasika, M.S., Burgard, A. and Maranas, C.D. (2006) A computational framework for the topological analysis and targeted disruption of signal transduction networks, *Biophysical Journal*, **91**, 382-398.

Davis, P.M., Cynthia G, *et al.* (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks, *Nucleic Acids Research*, **37**, D786.

Eke, I. and Cordes, N. (2015) Focal adhesion signaling and therapy resistance in cancer, *Seminars in Cancer Biology*, **31**, 65-75.

Farrow, J.M., Yang, J.C. and Evans, C.P. (2014) Autophagy as a modulator and target in prostate cancer, *Nature Reviews Urology*, **11**, 508-516.

Frolkis, A., *et al.* (2009) SMPDB: The Small Molecule Pathway Database, *Nucleic Acids Research*, **38**, D480-D487.

Goh, K.I., *et al.* (2007) The human disease network, *Proceedings of the National Academy of Sciences of the United States of America,* **104**, 8685-8690.

Huang, J., *et al.* (2011) eResponseNet: a package prioritizing candidate disease genes through cellular pathways, *Bioinformatics*, **27**, 2319-2320.

Iwata, H., *et al.* (2015) Systematic drug repositioning for a wide range of diseases with integrative analyses of phenotypic and molecular data, *Journal of Chemical Information and Modeling*, **55**, 446-

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

459.

Iwata, H., *et al.* (2015) Systematic Drug Repositioning for a Wide Range of Diseases with Integrative Analyses of Phenotypic and Molecular Data, *Journal of Chemical Information and Modeling*, **55**, 446-459.

Jewison, T., *et al.* (2014) SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database, *Nucleic Acids Research*, **42**, 478-484.

Kaapu, K.J., *et al.* (2015) Sotalol, but not digoxin is associated with decreased prostate cancer risk: A population-based case-control study, *International Journal of Cancer*, **137**, 1187-1195.

Kahn, B.B. and Flier, J.S. (2000) Obesity and insulin resistance, *The Journal of Clinical Investigation*, **106**, 473-481.

Kutalik, Z., Beckmann, J. and Bergmann, S. (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data, *Nature Biotechnology.*, **26**, 531.

Langin, D., *et al.* (2005) Adipocyte lipases and defect of lipolysis in human obesity, *Diabetes*, **54**, 3190-3197.

Lathia, J.D., Mattson, M.P. and Cheng, A. (2008) Notch: from neural development to neurological disorders, *Journal of Neurochemistry*, **107**, 1471-1481.

Lu, L. and Yu, H. (2018) DR2DI: a powerful computational tool for predicting novel drug-disease associations, *Journal of Computer-Aided Molecular Design*, **32**, 633-642.

Moretto, E., *et al.* (2018) Glutamatergic synapses in neurodevelopmental disorders, *Progress in Neuropsychopharmacology & Biological Psychiatry*, **84**, 328-342.

Neschadim, A., Summerlee, A.J. and Silvertown, J.D. (2015) Targeting the relaxin hormonal pathway in prostate cancer, *International Journal of Cancer*, **137**, 2287-2295.

Ogretmen, B. (2018) Sphingolipid metabolism in cancer signalling and therapy, *Nature Reviews Cancer*, **18**, 33-50.

Olsson, U., *et al.* (2001) Changes in matrix proteoglycans induced by insulin and fatty acids in hepatic cells may contribute to dyslipidemia of insulin resistance, *Diabetes*, **50**, 2126-2132.

Ozvegy-Laczka, C., *et al.* (2005) Tyrosine kinase inhibitor resistance in cancer: role of ABC multidrug transporters, *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy*, **8**, 15-26.

Paul, S.M., *et al.* (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge, *Nature Review Drug Discovery*, **9**, 203-214.

Piñero, J., *et al.* (2016) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, *Nucleic Acids Research*, **45**, D833-D839.

Tata, A.M., *et al.* (2014) Cholinergic system dysfunction and neurodegenerative diseases: cause or effect?, *CNS & neurological disorders drug targets*, **13**, 1294-1303.

Taylor, V.H. and Macqueen, G.M. (2010) The Role of Adipokines in Understanding the Associations

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted August 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between Obesity and Depression, *Journal of Obesity*, **2010**.

Viollet, B.*, et al.* (2009) Targeting the AMPK pathway for the treatment of Type 2 diabetes, *Frontiers in Bioscience*, **14**, 3380-3400.

von Eichborn, J.*, et al.* (2011) PROMISCUOUS: a database for network-based drug-repositioning, *Nucleic Acids Research*, **39**, D1060-D1066.

Wang, L.*, et al.* (2014) *Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks*, *CPT Pharmacometrics System Pharmacology,* **3**, e146

Yeger-Lotem, E.*, et al.* (2009) Bridging the gap between high-throughput genetic and transcriptional data reveals cellular pathways responding to alpha-synuclein toxicity, *Nature Genetics*, **41**, 316.

Yildirim, M.A.*, et al.* (2007) Drug-target network, *Nature. Biotechnology.*, **25**, 1119-1126.

You, J.*, et al.* (2018) RNA-Seq analysis and comparison of corneal epithelium in keratoconus and myopia patients, *Scientific Reports*, **8**, 389.

Yu, G.*, et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters, *Omics: A Journal of Integrative Biology*, **16**, 284-287.

Yu, H., Chen, X. and Lu, L. (2017) Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm, *Scientific Report*, **7**, 43792.

Zhao, S. and Li, S. (2012) A co-module approach for elucidating drug–disease associations and revealing their molecular basis, *Bioinformatics*, **28**, 955-961.