Personalized single-cell networks: a framework to predict the response of any gene to any drug for any patient

Haripriya Harikumar[1,2+*], Thomas P. Quinn[1+*], Santu Rana[1], Sunil Gupta[1], and Svetha Venkatesh[1]

[1]Applied Artificial Intelligence Institute, Deakin University, Geelong, Australia
[2]Institute for Health Transformation, Deakin University, Geelong, Australia
+ contributed equally, * *h.harikumar@deakin.edu.au; contacttomquinn@gmail.com*

Abstract

Background: The last decade has seen a major increase in the availability of genomic data. This includes expert-curated databases that describe the biological activity of genes, as well as high-throughput assays that measure gene expression in bulk tissue and single cells. Integrating these heterogeneous data sources can generate new hypotheses about biological systems. Our primary objective is to combine population-level drug-response data with patient-level single-cell expression data to predict how any gene will respond to any drug for any patient.

Methods: We take 2 approaches to benchmarking a "dual-channel" random walk with restart (RWR) for data integration. First, we evaluate how well RWR can predict known gene functions from single-cell gene co-expression networks. Second, we evaluate how well RWR can predict known drug responses from individual cell networks. We then present two exploratory applications. In the first application, we combine the Gene Ontology database with glioblastoma single cells from 5 individual patients to identify genes whose functions differ between cancers. In the second application, we combine the LINCS drug-response database with the same glioblastoma data to identify genes that may exhibit patient-specific drug responses.

Conclusions: Our manuscript introduces two innovations to the integration of heterogeneous biological data. First, we use a "dual-channel" method to predict up-regulation and down-regulation separately. Second, we use individualized single-cell gene co-expression networks to make personalized predictions. These innovations let us predict gene function and drug response for individual patients. Taken together, our work shows promise that single-cell co-expression data could be combined in heterogeneous information networks to facilitate precision medicine.

1 Introduction

Advances in high-throughput RNA-sequencing (RNA-Seq) have made it possible to quantify RNA presence in any biological sample [29], producing a gene expression signature that can serve as a biomarker for disease prediction [15, 1, 47] or surveillance [31, 38]. Over the last few years, single-cell RNA-Seq has risen in popularity [14]. Compared with conventional bulk RNA-Seq, which measures the average gene expression for an individual sample, single-cell RNA-Seq (scRNA-Seq) measures gene expression for an individual cell. This new mode of data makes it possible to explore tissue heterogeneity, notably tumor heterogeneity [25], by producing multiple data points per individual (i.e., one for each cell). Since genes are often understood to work in cooperative modules, the analysis of *gene co-expression networks* is commonplace. For bulk RNA-Seq, a gene co-expression network describes how genes co-occur for a population of individual samples. For scRNA-Seq, the network describes gene co-expression for a population of single cells. When these cells belong to an individual patient, the scRNA-Seq network is a kind of **personalized network** that one could use for precision medicine tasks.

Gene co-expression networks can be integrated with outside information to combine **general knowledge** (in the form of a relational database like Gene Ontology [2]) with **specific knowledge** about a sample (in the form of a co-expression network). For example, weighted gene co-expression network analysis is a popular method for functionally characterizing parts of the network, or the network as a whole [23, 24]. Although these coarse descriptions are useful, one could also combine general- and specific knowledge to make finer-level predictions about the behavior of *individual genes*. By representing each modality as a graph, multiple data streams can be combined into a **heterogeneous information network**, and then analyzed under a unified framework based on the principle of "guilt-by-association" [45] (e.g., if "a" is connected to "b" and "b" is connected to "c", then "a" is probably connected to "c). When the general knowledge is **gene-annotation** associations, we can (a) impute the function for genes with no known role or (b) select the most important known function. When the general knowledge is **gene-drug** response, we can predict the response of any gene to any drug. Since these

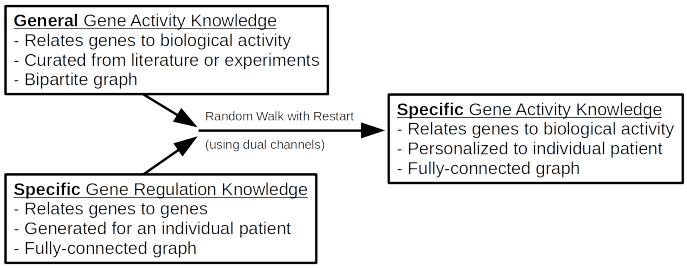

Figure 1: An abstracted schematic of the proposed framework. Expert-curated databases like Gene Ontology (GO) and the Library of Integrated Network-Based Cellular Signatures (LINCS) can provide some **general knowledge** about biological activity. High-throughput single-cell sequencing assays can provide **specific knowledge** for an individual patient. Random walk with restart (RWR) can combine these heterogeneous data sources to provide specific knowledge about biological activity for an individual patient. This framework allows us to predict how any gene will respond to any drug for any patient.

inferences are tailored to the co-expression network used, they can be made personalized by using the single-cell network of an individual patient.

Random walk (RW) is a popular method that offers a general solution to the analysis of heterogeneous information networks [34, 45]. There are many variants to RW, including random walk with restart (RWR), where each step has a probability of restarting from the starting node (or a neighbor of the starting node) [44]. RW and RWR are often used in recommendation systems [3, 8, 19], but can also perform other machine learning tasks like image segmentation [16, 18], image captioning [32], or community detection [36, 22]. One advantage of RW is that it can handle missing data [17], making it a good choice for processing sparse gene annotation databases. RW and RWR have both found use in biology to find associations between genes and another data modality. For example, the "InfAcrOnt" method used an RW-based method to infer similarities between ontology terms by integrating annotations with a gene-gene interaction network [6]. Similarly, the "RWLPAP" method used RW to find lncRNA-protein associations [50], while others have used RW to predict gene-disease associations [51]. Meanwhile, RWR has been used to identify epigenetic factors within the genome [26], key genes involved in colorectal cancer [9], novel microRNA-disease associations [43], infection-related genes [52], disease-related genes [46], and functional similarities between genes [35]. Bi-random walk, another random walk variant, has been used to rank disease genes from a protein-protein interaction network [48].

In contrast to the previous works, which make use of population-level graphs, *we apply RWR to patient-level graphs, allowing us to make predictions about gene behavior that are personalized to each patient.* We take 2 approaches to benchmarking "dual-channel" RWR for data integration. First, we evaluate how well RWR can predict known gene functions from single-cell gene co-expression networks. Second, we evaluate how well RWR can predict known drug responses from individual cell networks. We then present two exploratory applications. In the first application, we combine the Gene Ontology database with glioblastoma single cells from 5 individual patients to identify genes whose functions differ between cancers. In the second application, we combine the LINCS drug-response database with the same glioblastoma data to identify genes that may exhibit patient-specific drug responses. Taken together, our work shows promise that single-cell co-expression data could be combined in heterogeneous information networks to facilitate precision medicine.

2 Methods

2.1 Overview

In the medical domain, gene expression can be used as a biomarker to measure the functional state of a cell. One way in which drugs mediate their therapeutic or toxic effects is by altering gene expression. However, the assays needed to test how gene expression changes in response to a drug can be expensive and time consuming. Imputation has the potential to accelerate research by "recommending" novel gene-drug responses. Random walk methods can combine sparse heterogeneous graphs based on the principle of "guilt-by-association" [45]. Figure 1 provides an abstracted schematic of the proposed framework. Figure 2 provides a visualization of the

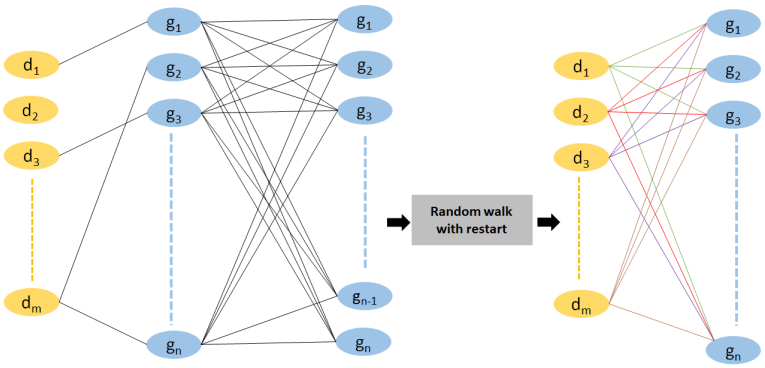

Figure 2: Our goal is to combine a generic gene-based bipartite graph with the auxiliary knowledge of a fully-connected personalized graph. RWR will impute the missing links (and update the existing links) by "walking through" the auxiliary information. The left panel shows a (sparsely-connected) gene-drug graph combined with a (fully-connected) gene-gene graph, where d_i represents the drugs and g_i represents the genes. The example gene-drug network has missing links. The right panel shows the output of RWR: a complete network of newly predicted gene-drug interactions. Here, the missing link between any drug d_i and any gene g_i is replaced with a new link. The method works based on the principle of "guilt-by-association": the value of the new $d_i - g_i$ links will be large if g_i is strongly connected to genes that are also connected to d_i. When the gene-drug graph is fully-connected, RWR will instead "update" the importance of each connection.

input and output for the random walk with restart (RWR) method. Figure 3 presents a bird's-eye view of the data processing, validation, and application steps performed in this study.

2.2 Data acquisition

The gene expression data come from two primary sources. First, we acquired single-cell RNA-Seq (scRNA-Seq) expression data for 5 glioblastoma multiforme tumors [33] using the recount2 package for the R programming language [7] (ID: SRP042161). Since scRNA-Seq data are incredibly sparse, and since the random walk with restart algorithm is computationally expensive, we elected to remove genes that had zero values in more than 25% of cells. This resulted in 3022 genes. Finally, we randomly split the cells into 5-folds per patient so that we could estimate the variability of our downstream analyses. Second, we acquired gene expression data from the Library of Integrated Network-Based Cellular Signatures (LINCS) [20] using the Gene Expression Omnibus (GEO) [10] (ID: GSE70138). We split these LINCS data into smaller data sets based on the cell line ID under study. We *a priori* included the A375 (skin; malignant melanoma), HA1E (kidney; embryonic), HT29 (colon; adenocarcinoma), MCF7 (breast; adenocarcinoma), and PC3 (prostate; adenocarcinoma) cell lines because they were treated with the largest number of drugs.

2.3 Defining the gene co-expression network graphs

Although correlation is a popular choice for measuring gene co-expression, correlations can yield spurious results for next-generation sequencing data [27]. Instead, we calculate the proportionality between genes using the ϕ_s metric from the propr package for the R programming language [39]. Although this does not offer a perfect solution [12], studying gene-gene proportionality has a strong theoretical justification [27] and empirically outperforms other metrics of association for scRNA-Seq [41].

The proportionality metric describes the dissimilarity between any two genes, and ranges from $[0, \infty)$, where 0 indicates a perfect association. We converted this to a similarity measure ϕ_i that ranges from $[0, 1]$ by max-scaling $\phi_i = (\max(\phi_s) - \phi_s)/\max(\phi_s)$, such that $\phi_i = 1$ when $\phi_s = 0$. A gene-gene matrix of ϕ_i scores is analogous to a gene-gene matrix of correlation coefficients, and constitutes our gene co-expression network. We calculated the ϕ_i co-expression network for the entire scRNA-Seq data set (1 network), for each of the 5-folds per-patient (25 networks total), and for each LINCS baseline (drug-free) cell line (5 networks total). All co-expression networks are available from https://zenodo.org/record/3522494.

2.4 Defining the bipartite graphs

We constructed two types of bipartite graphs: the **gene-annotation graph** and the **gene-drug graph**. First, we made the gene-annotation graph from the Gene Ontology Biological Process database [2] via the AnnotationDbi and org.Hs.eg.db Bioconductor packages. An edge exists whenever a gene is associated with an annotation. Second, we made the gene-drug graphs using the LINCS data. For each cell line, we computed a gene-drug graph by calculating the log-fold change between the median of the drug-treated cell's expression and

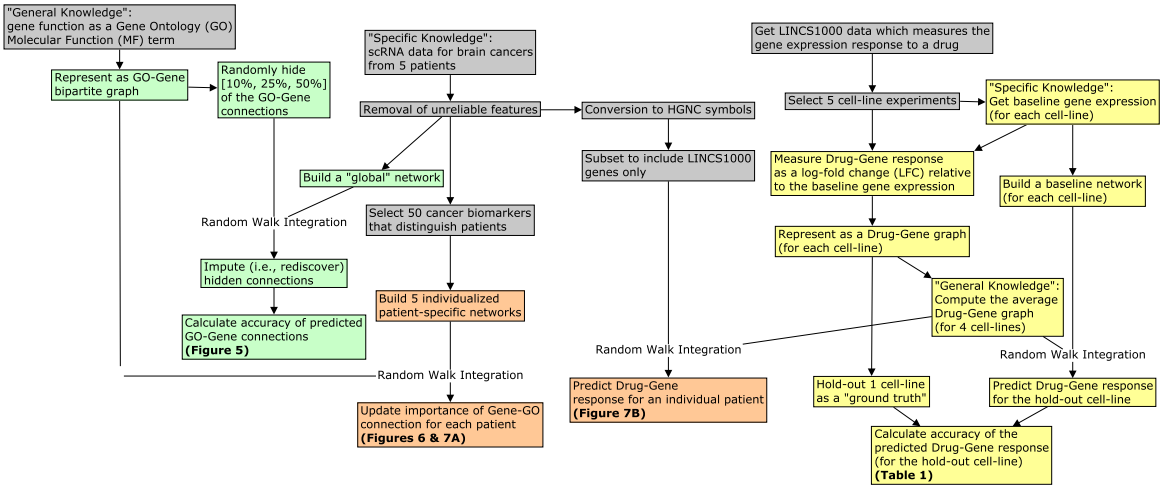

Figure 3: A bird's-eye view of the data collection, integration, and analysis steps performed in this study. We use RWR to combine general knowledge with some specific knowledge about a sample. We separately use gene function and drug-response data as the source of general knowledge. We use co-expression networks as the source of specific knowledge. By combining the drug-response data with an individualized single-cell network, we can make predictions about gene behavior that are personalized to each patient. Concept nodes are colored by activity type: data processing (gray), validation of gene-annotation prediction (green), validation of gene-drug prediction (yellow), and exploratory application (orange).

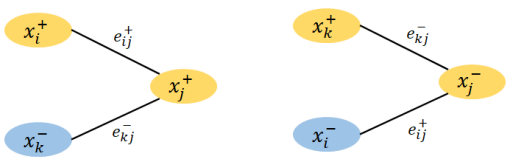

Figure 4: This figure illustrates the flow of information between adjacent nodes. The positive information nodes are yellow and the negative information nodes are purple. A positive edge weight is represented by e_{ij}^+, while a negative edge weight is represented by e_{kj}^-. The sign of the edge weights determines which information (positive or negative) flows from one node to another. The positive information of a node x_j depends on the positive information of x_i when the edge is positively weighted (think: positive times positive is positive). The negative information of a node x_j depends on the negative information of x_i when the edge is positively weighted (think: negative times positive is negative). Similarly, the negative flow of negative information can contribute to positive information. The "dual-channel" RWR algorithm incorporates these edge weights. Note that in practice each node will contain both positive and negative information.

the median of the drug-naive cell's expression. This results in a fully-connected and weighted bipartite graph, where a large positive value means that the drug causes the gene to up-regulate (and *vice versa*). All bipartite graphs are available from https://zenodo.org/record/3522494.

2.5 The combined co-expression and bipartite graph

Consider a graph G with $V = 1...N$ vertices, E^+ positive edges, and E^- negative edges. The graphs used for our analyses are composed for two parts: a (general knowledge) bipartite graph and a (specific knowledge) fully-connected gene co-expression graph. For a bipartite graph, the vertex set V can be separated into two distinct sets, V_1 and V_2, such that no edges exist within either set. For a fully-connected (or complete) graph, there exists an edge between every pair of vertices within one set. For the graph G, the bipartite and fully-connected graphs are joined via the common vertex set V_1 that contains genes and V_2 contains annotations or drugs.

2.6 Dual-channel random walk with restart (RWR)

Traditional RWR methods can only perform a random walk on graphs with positive edge weights [34]. Since the response of a gene to a drug is directional (up-regulated or down-regulated), we chose to use a modified RWR method, proposed by [5], that handles graphs with both positive and negative edge weights. Random walk requires transition probability matrices to decide the next step in the walk. The Chen et al. transition probability matrices can be computed based on the following equations:

$$P(x_j^+|x_i^+) = \frac{|e_{ij}|}{\sum_{l \in N(x_i)} |e_{il}|} \tag{1}$$

$$P(x_j^-|x_i^-) = \frac{|e_{ij}|}{\sum_{l \in N(x_i)} |e_{il}|} \tag{2}$$

when $e_{ij} \geq 0$, and

$$P(x_j^-|x_i^+) = \frac{|e_{ij}|}{\sum_{l \in N(x_i)} |e_{il}|} \tag{3}$$

$$P(x_j^+|x_i^-) = \frac{|e_{ij}|}{\sum_{l \in N(x_i)} |e_{il}|} \tag{4}$$

when $e_{ij} < 0$. For all equations, e_{ij} is the edge weight between nodes x_i and x_j, and $N(x_i)$ is the set of neighbors for node x_i. These equations separate out the positive (and negative) transitions, and are used to calculate the total positive (and negative) information flow for each node. They are fixed for all steps.

Though the transition probabilities are computed separately, the information accumulated in a node depends on both the positive and negative information which flows through the node. For example, the positive information in a node depends on the negative information of any neighboring node connected by a negative edge weight (think: negative times negative is positive). Likewise, negative information in a node depends on the positive information in a neighboring node connected by a negative edge weight, and *vice versa* (think: negative times positive is negative). Figure 4 illustrates the information flow to a node x_j from two neighbors.

The flow of information between the positive "plane" of the graph to the negative "plane" of the graph can be formulated with the equations:

$$P(x_j^+)_k = \left[\sum_{x_i \in N(x_i) \ \& \ e_{ij} \geq 0} P(x_i^+)_{k-1} P(x_j^+ \mid x_i^+) \right] + \left[\sum_{x_i \in N(x_i) \ \& \ e_{ij} < 0} P(x_i^-)_{k-1} P(x_j^+ \mid x_i^-) \right] \tag{5}$$

$$P(x_j^-)_k = \left[\sum_{x_i \in N(x_i) \ \& \ e_{ij} \geq 0} P(x_i^-)_{k-1} P(x_j^- \mid x_i^-) \right] + \left[\sum_{x_i \in N(x_i) \ \& \ e_{ij} < 0} P(x_i^+)_{k-1} P(x_j^- \mid x_i^+) \right] \tag{6}$$

where the probability $P(x_j^+)_k$ is updated at each step $k = 2...10000$.

RWR always considers a probability α to return back to the original nearest neighboring nodes at each step in the random walk. This is used to weigh the importance of node-specific information with respect to the whole graph, including for long walks:

$$P_{rst}(x_j^+)_k = (1 - \alpha) \times P(x_j^+)_{k-1} + \alpha \times P(x_j^+)_2 \tag{7}$$

$$P_{rst}(x_j^-)_k = (1 - \alpha) \times P(x_j^-)_{k-1} + \alpha \times P(x_j^-)_2 \tag{8}$$

where the restart probability $P_{rst}(x_j^+)_k$ is updated at each step $k = 2...10000$, and $P(x_j^+)_2$ is the probability after the first update. These equations find the positive and negative restart information with respect to the node x_j. Each $P_{rst}(x_j^+)_k$ is a vector of probabilities that together sum to 1. This probability has two parts: the

global information and the local information. The local information is the initial probability with respect to the nearest neighbors of node x_j, and is denoted by $P(x_j^+)_2$ [or $P(x_j^-)_2$] (i.e., the probability after the first update). The restart probability α is chosen from the range $[0, 1]$, where a higher value weighs the local information more than the global information. We chose $\alpha = 0.1$ to place a larger emphasis on the global information. Simulations with a toy data set verified this choice.

2.7 Analysis of random walk with restart (RWR) scores

For each gene, the RWR algorithm returns a vector of probabilities that together sum to 1. According to the guilt-by-association assumption, we interpret these probabilities to indicate the strength of the connection between the reference gene and each target. Since we are only interested in gene-annotation and gene-drug relationships, we exclude all gene-gene probabilities. Viewing the probability distribution as a composition (c.f., [11]), we perform a centered log-ratio transformation of each probability vector subset. This transformation normalizes the probability distributions so that we can compare them between samples [4, 37]. We define the RWR score r_{ga}^+ (or r_{ga}^-) for each gene-annotation connection as the transform of its RWR probability:

$$r_{ga}^+ = \log \frac{p_{ga}^+}{\sqrt[A]{\prod_i^A p_{gi}^+}} \tag{9}$$

$$r_{ga}^- = \log \frac{p_{ga}^-}{\sqrt[A]{\prod_i^A p_{gi}^-}} \tag{10}$$

for a bipartite graph describing $g = 1...G$ genes and $a = 1...A$ annotations (or A drugs), where $\mathbf{p}_g^+ = P_{rst}(x_g^+) = [p_{g1}^+, ..., p_{gA}^+]$ (i.e., from the final step). These transformed RWR scores can be used for univariate statistical analyses, such as an analysis of variance (ANOVA) (c.f., analysis of compositional data [13, 28]).

2.8 Benchmark validation

2.8.1 Validation of gene-annotation prediction

Our strategy to validate RWR for gene-annotation prediction involves "hiding" known functional associations and seeing whether the RWR algorithm can re-discover them. This is done by turning 1s into 0s in the bipartite graph, a process we call "sparsification". Our sparsification procedure works in 4 steps. First, we combine the original GO BP (or MF) bipartite graph with the master single-cell co-expression graph. Second, we subset the graph to include 25% of the gene annoations and 25% of the genes (this is done to reduce the computational overhead). Third, we randomly hide $[10, 25, 50]$ percent of the gene-annotation connections from the bipartite sub-graph. Since this random selection could cause a feature to lose all connections, we use a constrained sampling strategy: the subsampled graph must contain at least one non-zero entry for each feature. Fourth, we apply the RWR algorithm to the sparsified and non-sparsified graphs, separately. We repeat this process 25 times, using a different random graph each time. By comparing the RWR scores between the hidden and unknown connections, we can determine whether our method rediscovers hidden connections.

2.8.2 Validation of gene-drug prediction

We use a different strategy to validate RWR for drug-response prediction. Since we have the gene-drug and gene-gene interaction data for 5 cell lines (A375, HA1E, HT29, MCF7 and PC3), we can set aside the known gene-drug responses for 1 cell line (PC3) as a "ground truth" test set. Then, we can use a composite of the remaining 4 gene-drug graphs to predict the gene-drug responses for the withheld cell line.

This is done in two steps. First, we use the averaged gene-drug data for 4 cell lines (a general drug graph) and the gene-gene data for PC3 (a specific gene graph) to impute the gene-drug response for PC3 (a specific drug graph). In the second step, we use the gene-drug data for PC3 (a specific drug graph) and its corresponding gene-gene data (a specific gene graph) to calculate the "ground truth" RWR scores for PC (a specific drug graph). The "ground truth" is the RWR scores when all PC3 drug-response experiments have been performed. With these two outputs, we can calculate the agreement between the imputed and "ground truth" RWR scores (using Spearman's correlation, MSE, and accuracy).

2.9 Exploratory application of gene-drug prediction

Having demonstrated that RWR can perform well for single-cell co-expression networks, and can make meaningful drug-response predictions from composite LINCS data, we combine these heterogeneous data sources to make personalized drug-response predictions for individual single-cell networks. This requires some data munging. First, we transform the ENGS features used by the single-cell data into the HGNC features used by LINCS (only including genes with a 1-to-1 mapping, resulting in 181 genes). Second, we build an HGNC co-expression

network with ϕ_i (for 5 folds of 5 patients, yielding 25 networks total). Third, we combine the composite LINCS gene-drug bipartite graph with each of the 25 HGNC single-cell networks. Fourth, we use our RWR algorithm to predict how 181 genes would respond to 1732 drugs for each patient fold. As above, we perform an analysis of variance (ANOVA) to detect inter-patient differences.

3 Results and Discussion

3.1 Why use single cells?

In this study, we analyze a previously published single-cell data set that measured the gene expression for 5 glioblastoma patients. A principal components analysis of these data show that the major axes of variance tend to group the cells according to the patient-of-origin. Indeed, an ANOVA of gene expression with respect to patient ID reveals that 2204 of the 3022 genes have significantly different expression in at least one patient (FDR-adjusted $p < .05$). This suggests that single-cell gene expression is unique to each patient.

Although it is possible to obtain sample-specific *gene expression* using bulk RNA-Seq, our approach to data integration exploits the graphical structure of sample-specific *gene co-expression networks*. scRNA-Seq, generating multiple measurements per individual, makes the computation of sample-specific co-expression networks straightforward (though others have proposed ways to estimate these from bulk RNA-Seq [21, 30]).

3.2 Validation of gene-annotation prediction

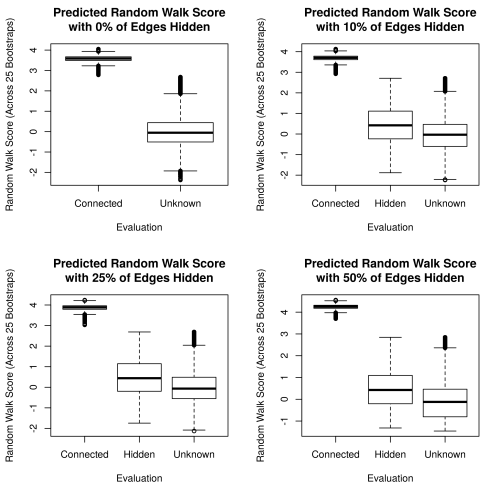

Figure 5: The RWR scores for the hidden and unknown gene-annotation connections (faceted by the amount of sparsity). When known connections are hidden, RWR tends to give higher scores than when connections are unknown, suggesting that RWR can discover real gene-annotation relationships. Note that GO is an incomplete database: the absence of a gene-annotation connection is not the evidence of absence. For this reason, we do know whether the high scoring "Unknown" connections are false positives or previously undiscovered connections. All t-tests comparing "Hidden" and "Unknown" connections have $p < 10^{-15}$.

The Gene Ontology (GO) project has curated a database which relates genes to biological processes (BP) and molecular functions (MF) (called **annotations**). The GO database has widespread use in bioinformatics for assigning "functional" relevance to sets of gene biomarkers [42]. Although GO organizes the semantic relationships between annotations as a directed acyclic graph, we could more simply represent the relationships as a bipartite graph. By combining a (fully-connected) gene co-expression graph with a (sparsely-connected) gene-annotation bipartite graph, RWR can predict new gene-annotation connections.

To test whether the RWR predictions are meaningful, we "hid" a percentage of known gene-annotation links (by turning 1s into 0s in the bipartite graph), and compared the RWR scores for the *hidden* gene-annotation links against those for the *unknown* links (see Methods for a definition of the RWR score). Figure 5 shows that the RWR scores for hidden connections are appreciably larger than for the unknown connections, confirming that RWR can discover real gene-annotation relationships from a single-cell gene co-expression network.

3.3 Exploratory application of gene-annotation prediction

Since single-cell RNA-Seq assays measure RNA for multiple cells per patient, we can use these data to build a personalized graph that describes the gene-gene relationships for an individual patient. In order to estimate the

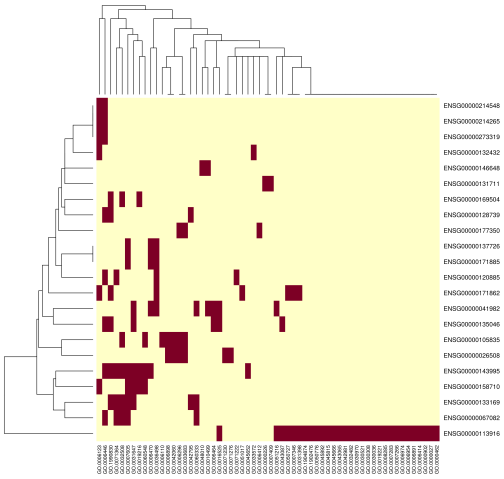

Figure 6: This figure shows a heatmap of the predicted gene-annotation connections that are significantly different between patients (dark red indicates a gene-wise FDR-adjusted $p < .05$). Out of the 50 genes tested, 22 appear to have some form of patient-specific activity.

	Correlation	MSE	Top 5% (ACC)	Top 10% (ACC)	Top 25% (ACC)	Top 50% (ACC)
Positive Channel	0.7173	0.2022	0.9279	0.8857	0.8574	0.8427
Negative Channel	0.5502	0.2176	0.9450	0.8946	0.7578	0.7053

Table 1: Overall agreement (Spearman's correlation and MSE) and the accuracy of the overlap (for the top 5%, 10%, 25%, and 50% predicted scores), as calculated separately for the positive and negative channels. Overall, agreement is high, especially for the top up-regulation and down-regulation events.

variation in these personalized graphs, we divided the cells from each sample into 5 folds (giving us 5 networks per-patient). Above, we show that RWR can discover real gene-annotation relationships. By combining the personalized graph (a kind of *specific knowledge*) with a gene-annotation bipartite graph (a kind of *general knowledge*), the RWR algorithm will score the gene-annotation connections for a given patient. From this, we can identify genes that may have a different functional importance in one cancer versus the others.

Taking a subset of the 50 genes with the largest inter-patient differences, we use RWR to compute personalized RWR scores. This results in 25 matrices (for 5 folds of 5 patients), each with 50 rows (for genes) and 369 columns (for BP annotations). Performing an ANOVA on each gene-annotation connection results in a matrix of 50x369 p-values. Figure 6 shows a heatmap of the significant gene-annotation connections (dark red indicates a gene-wise FDR-adjusted $p < .05$). Figure 7A plots the per-patient RWR scores for 4 annotations of the BCL-6 gene that significantly differ between patients. BCL-6 is an important biomarker whose increased expression is associated with worse outcomes in glioblastoma [49]. Our analysis suggests that BCL-6 could have a larger role in inflammation for patients 3 and 5, but a larger role in cartilage development and translational elongation in patient 1. Of course, this hypothesis requires experimental validation.

3.4 Validation of gene-drug prediction

The NIH LINCS program has generated a large amount of data on how the gene expression signatures of cell lines change in response to a drug. By conceptualizing the baseline (drug-free) gene co-expression network as a complete graph of *specific knowledge*, and by re-factoring the average gene-drug response as a (weighted) bipartite graph of *general knowledge*, we can apply the same RWR algorithm to predict a cell's gene expression response to any drug. Since the modified RWR algorithm contains two channels–a positive and negative channel–we can predict up-regulation or down-regulation events separately.

To test whether RWR can make accurate predictions about how a gene in a cell would respond to a drug, we ran the RWR algorithm on the baseline (drug-free) gene co-expression graph of the PC3 cell line using a composite gene-drug graph of 4 different cell lines. We then compared these RWR scores with a "ground truth" (i.e., the RWR scores for when all PC3 drug-response experiments have been performed). The agreement between the composite gene-drug RWR scores and the "ground truth" gene-drug RWR scores tells us how well the composite gene-drug map generalizes to new cell types. Table 1 shows that agreement is high, especially for the top up-regulation and down-regulation events. This confirms that our composite gene-drug graph is useful for drug-response prediction.

3.5 Exploratory application of gene-drug prediction

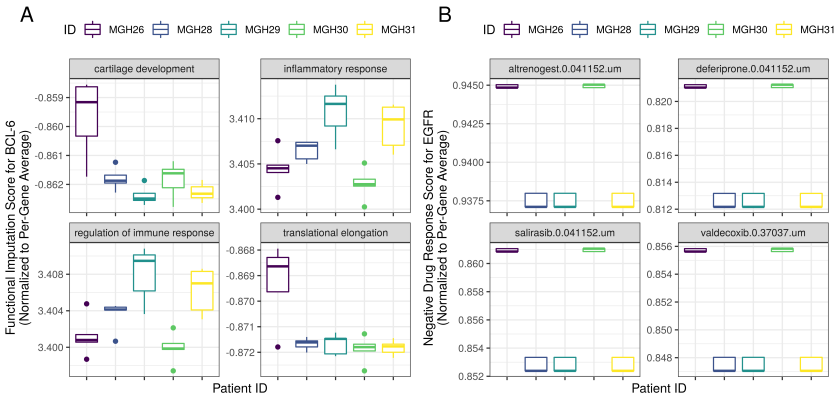

Figure 7: The personalized RWR scores for 4 biological functions of the BCL-6 gene (left panel) and for the EGFR response to 4 drugs (right panel). Panel A suggests that BCL-6 may have a larger role in inflammation for patients 3 and 5, but a larger role in cartilage development and translational elongation in patient 1. Panel B suggests that the anti-inflammatory drug valdecoxib and the anti-neoplastic drug salirasib may cause a stronger down-regulation of EGFR in patients 1 and 4 versus the others.

The RWR algorithm can combine specific knowledge and general knowledge from disparate sources to make personalized recommendations. This makes RWR a potentially valuable tool for precision medicine.

As an exploratory analysis, we combine the personalized gene co-expression networks with the composite gene-drug graph from LINCS. By running the RWR algorithm on these two data streams, the RWR scores will now suggest how the expression of any gene might change in response to any drug for each of the 5 glioblastoma patients. Using an ANOVA, we identify hundreds of gene-drug connections with RWR scores that differ significantly between patients (gene-wise FDR-adjusted $p < .05$).

Figure 7B shows an example of drugs that have different (negative channel) RWR scores for EGFR. It suggests that the anti-inflammatory drug valdecoxib and the anti-neoplastic drug salirasib may cause a stronger down-regulation of EGFR (a pan-cancer oncogene [40]) in patients 1 and 4 versus the others. The Supplementary Information includes a complete table of the unadjusted ANOVA p-values for the gene-drug inter-patient differences available in https://zenodo.org/record/3743897.

4 Limitations

We deployed our framework on only 5 individual patients. As such, we lack a sufficient sample size to test whether any inter-patient differences could be explained by known demographic or clinical phenotypes. It is worth noting that cancer cells are very heterogeneous and, depending on the location of the sample collection, the composition of cell types (and thus gene expression profiles) can change dramatically. As such, factors other than the patient-specific tumour profile could account for differences in the sample-specific gene co-expression networks. Such differences may be difficult to account for without careful experimental design and standardization. Although the scope of this paper is to prove the concept, we wish to remind the readers that much care should be taken when translating the methodology to real-world clinical problem-solving.

In the absence of experimental validation, we support our analyses using 2 forms of *in silico* validation, which together demonstrate that RWR can integrate sparse heterogeneous data to discover real biological activity. Although we find the *in silico* validation encouraging, we acknowledge that RWR is merely a prediction tool that recommends hypotheses, and that these predictions may change when the source of general knowledge changes. Experimental validation is needed to determine whether these hypotheses prove true in practice. Further work is needed to validate the clinical relevance of the proposed framework.

5 Conclusions

This manuscript describes a framework for combining patient-specific single-cell networks with public drug-response data to make personalized predictions about drug response. Importantly, our approach makes use of a generic framework, and so can be applied to combine many kinds of data. We think the targeted analysis of personalized single-cell networks is promising, and could offer a new direction for precision medicine research.

We conclude with some perspectives on what the future of personalized network analysis may hold. Although RWR can handle sparse heterogeneous data, the positive and negative information obtained for each node can

be infinitesimally small. One might address this by first transforming the RWR probabilities. Otherwise, we note that RWR is computationally expensive, making the analysis of high-dimensional data prohibitively slow. One might address this by pre-training a deep neural network to provide an approximate RWR solution. These improvements could help scale personalized predictions to larger graphs.

Abbreviations

- RW: random walk
- RWR: random walk with restart
- RNA-Seq: RNA sequencing
- MSE: mean square error
- scRNA-Seq: single-cell RNA sequencing
- LINCS: Library of Integrated Network-Based Cellular Signatures
- GO: Gene Ontology
- BP: Biological Process
- MF: Molecular Function
- ANOVA: analysis of variance

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The raw data are publicly available from the resources described in the Methods. All gene co-expression and bipartite graphs used in these analyses are available from https://zenodo.org/record/3522494.

Competing interests

No authors have competing interests.

Acknowledgements

Not applicable.

Authors' contributions

HH implemented the RWR algorithm and applied it to the graphical data. TPQ prepared the graph data and performed the analysis of the resultant RWR scores. HH and TPQ reviewed the literature, designed the experiments, and drafted the manuscript. All authors helped conceptualize the project and revise the manuscript.

Acknowledgements

Not applicable.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12):6745–6750, June 1999.

[2] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[3] Toine Bogers. Movie recommendation using random walks over the contextual graph. In *Proc. of the 2nd Intl. Workshop on Context-Aware Recommender Systems*, 2010.

[4] K. Gerald van den Boogaart and Raimon Tolosana-Delgado. Fundamental Concepts of Compositional Data Analysis. In *Analyzing Compositional Data with R*, Use R!, pages 13–50. Springer Berlin Heidelberg, 2013.

[5] Yu-Chih Chen, Yu-Shi Lin, Yu-Chun Shen, and Shou-De Lin. A modified random walk framework for handling negative ratings and generating explanations. *ACM transactions on Intelligent Systems and technology (tISt)*, 4(1):12, 2013.

[6] Liang Cheng, Yue Jiang, Hong Ju, Jie Sun, Jiajie Peng, Meng Zhou, and Yang Hu. Infacront: calculating cross-ontology term similarities using information flow by a random walk. *BMC genomics*, 19(1):919, 2018.

[7] Leonardo Collado-Torres, Abhinav Nellore, and Andrew E. Jaffe. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. *F1000Research*, 6:1558, August 2017.

[8] Colin Cooper, Sang Hyuk Lee, Tomasz Radzik, and Yiannis Siantos. Random walks in recommender systems: exact computation and simulations. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 811–816. ACM, 2014.

[9] Xiaofeng Cui, Kexin Shen, Zhongshi Xie, Tongjun Liu, and Haishan Zhang. Identification of key genes in colorectal cancer using random walk with restart. *Molecular medicine reports*, 15(2):867–872, 2017.

[10] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, January 2002.

[11] Ionas Erb and Nihat Ay. The information-geometric perspective of Compositional Data Analysis. *arXiv preprint arXiv:2005.11510*, 2020.

[12] Ionas Erb and Cedric Notredame. How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, 135:21–36, 2016.

[13] Andrew D. Fernandes, Jennifer Ns Reid, Jean M. Macklaim, Thomas A. McMurrough, David R. Edgell, and Gregory B. Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:15, 2014.

[14] Charles Gawad, Winston Koh, and Stephen R. Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188, March 2016.

[15] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, 286(5439):531–537, October 1999.

[16] Leo Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1768–1783, 2006.

[17] Sander Greenland, Mohammad Ali Mansournia, and Douglas G Altman. Sparse data bias: a problem hiding in plain sight. *bmj*, 352:i1981, 2016.

[18] Sonu Kumar Jha, Purnendu Bannerjee, and Subhadeep Banik. Random walks based image segmentation using color space graphs. *Procedia Technology*, 10:271–278, 2013.

[19] Anne-Marie Kermarrec, Vincent Leroy, Afshin Moin, and Christopher Thraves. Application of random walks to decentralized recommender systems. In *International Conference On Principles Of Distributed Systems*, pages 48–63. Springer, 2010.

[20] Amar Koleti, Raymond Terryn, Vasileios Stathias, Caty Chung, Daniel J. Cooper, John P. Turner, Dušica Vidović, Michele Forlin, Tanya T. Kelley, Alessandro D'Urso, Bryce K. Allen, Denis Torre, Kathleen M. Jagodnik, Lily Wang, Sherry L. Jenkins, Christopher Mader, Wen Niu, Mehdi Fazel, Naim Mahi, Marcin Pilarczyk, Nicholas Clark, Behrouz Shamsaei, Jarek Meller, Juozas Vasiliauskas, John Reichard, Mario Medvedovic, Avi Ma'ayan, Ajay Pillai, and Stephan C. Schürer. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Research*, 46(D1):D558–D566, January 2018.

[21] Marieke Lydia Kuijjer, Matthew George Tung, GuoCheng Yuan, John Quackenbush, and Kimberly Glass. Estimating Sample-Specific Regulatory Networks. *iScience*, 14:226–240, April 2019.

[22] Zhana Kuncheva and Giovanni Montana. Community detection in multiplex networks using locally adaptive random walks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1308–1315. ACM, 2015.

[23] Peter Langfelder and Steve Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology*, 1:54, 2007.

[24] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[25] Devon A. Lawson, Kai Kessenbrock, Ryan T. Davis, Nicholas Pervolarakis, and Zena Werb. Tumour heterogeneity and metastasis at single-cell resolution. *Nature Cell Biology*, 20(12):1349–1360, December 2018.

[26] JiaRui Li, Lei Chen, ShaoPeng Wang, YuHang Zhang, XiangYin Kong, Tao Huang, and Yu-Dong Cai. A computational method using the random walk with restart algorithm for identifying novel epigenetic factors. *Molecular genetics and genomics*, 293(1):293–301, 2018.

[27] David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: A Valid Alternative to Correlation for Relative Data. *PLoS Computational Biology*, 11(3), March 2015.

[28] Siddhartha Mandal, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26, May 2015.

[29] Michael L. Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, January 2010.

[30] Thin Nguyen, Samuel C. Lee, Thomas P. Quinn, Buu Truong, Xiaomei Li, Truyen Tran, Svetha Venkatesh, and Thuc Duy Le. Personalized Annotation-based Networks (PAN) for the Prediction of Breast Cancer Relapse. *bioRxiv*, page 534628, May 2019.

[31] Keith Noto, Saeed Majidi, Andrea G. Edlow, Heather C. Wick, Diana W. Bianchi, and Donna K. Slonim. CSAX: Characterizing Systematic Anomalies in eXpression Data. *Journal of Computational Biology*, 22(5):402–413, May 2015.

[32] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. Gcap: Graph-based automatic image captioning. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 146–146. IEEE, 2004.

[33] Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev, and Bradley E. Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, June 2014.

[34] Karl Pearson. The problem of the random walk. *Nature*, 72(1867):342, 1905.

[35] Jiajie Peng, Xuanshuo Zhang, Weiwei Hui, Junya Lu, Qianqian Li, Shuhui Liu, and Xuequn Shang. Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. *BMC systems biology*, 12(2):18, 2018.

[36] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, 2005.

[37] Thomas P. Quinn, Ionas Erb, Mark F. Richardson, and Tamsyn M. Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, August 2018.

[38] Thomas P. Quinn, Thin Nguyen, Samuel C. Lee, and Svetha Venkatesh. Cancer as a Tissue Anomaly: Classifying Tumor Transcriptomes Based Only on Healthy Data. *Frontiers in Genetics*, 10, 2019.

[39] Thomas P. Quinn, Mark F. Richardson, David Lovell, and Tamsyn M. Crowley. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Scientific Reports*, 7(1):16252, November 2017.

[40] Sara Sigismund, Daniele Avanzato, and Letizia Lanzetti. Emerging functions of the EGFR in cancer. *Molecular Oncology*, 12(1):3–20, January 2018.

[41] Michael A. Skinnider, Jordan W. Squair, and Leonard J. Foster. Evaluating measures of association for single-cell transcriptomics. *Nature Methods*, 16(5):381–386, May 2019.

[42] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005.

[43] Jie Sun, Hongbo Shi, Zhenzhen Wang, Changjian Zhang, Lin Liu, Letian Wang, Weiwei He, Dapeng Hao, Shulin Liu, and Meng Zhou. Inferring novel lncrna–disease associations based on a random walk model of a lncrna functional similarity network. *Molecular BioSystems*, 10(8):2074–2081, 2014.

[44] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 613–622. IEEE, 2006.

[45] Koki Tsuyuzaki and Itoshi Nikaido. Biological Systems as Heterogeneous Information Networks: A Mini-

review and Perspectives. December 2017.

[46] Alberto Valdeolivas, Laurent Tichit, Claire Navarro, Sophie Perrin, Gaelle Odelin, Nicolas Levy, Pierre Cau, Elisabeth Remy, and Anaïs Baudot. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35(3):497–505, 2018.

[47] Laura J. van 't Veer, Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards, and Stephen H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, January 2002.

[48] Maoqiang Xie, Taehyun Hwang, and Rui Kuang. Prioritizing disease genes by bi-random walk. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 292–303. Springer, 2012.

[49] Liang Xu, Ye Chen, Marina Dutra-Clarke, Anand Mayakonda, Masaharu Hazawa, Steve E. Savinoff, Ngan Doan, Jonathan W. Said, William H. Yong, Ashley Watkins, Henry Yang, Ling-Wen Ding, Yan-Yi Jiang, Jeffrey W. Tyner, Jianhong Ching, Jean-Paul Kovalik, Vikas Madan, Shing-Leng Chan, Markus Müschen, Joshua J. Breunig, De-Chen Lin, and H. Phillip Koeffler. BCL6 promotes glioma and serves as a therapeutic target. *Proceedings of the National Academy of Sciences of the United States of America*, 114(15):3981–3986, April 2017.

[50] Qi Zhao, Dan Liang, Huan Hu, Guofei Ren, and Hongsheng Liu. Rwlpap: Random walk for lncrna-protein associations prediction. *Protein and peptide letters*, 25(9):830–837, 2018.

[51] Zhi-Qin Zhao, Guo-Sheng Han, Zu-Guo Yu, and Jinyan Li. Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. *Computational biology and chemistry*, 57:21–28, 2015.

[52] Liucun Zhu, Fangchu Su, YaoChen Xu, and Quan Zou. Network-based method for mining novel hpv infection related genes using random walk with restart algorithm. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1864(6):2376–2383, 2018.