Computers in Biology and Medicine 119 (2020) 103692

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Network-based prioritization of cancer genes by integrative ranks from multi-omics data

Haixia Shang [a], Zhi-Ping Liu [a,b,**]

[a] Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China
[b] Center of Intelligent Medicine, Shandong University, Jinan, Shandong 250061, China

ARTICLE INFO

Keywords:
Cancer gene prioritization
Multi-omics data integration
Multiplex networks
Constrained PageRank

ABSTRACT

Finding disease genes related to cancer is of great importance for diagnosis and treatment. With the development of high-throughput technologies, more and more multiple-level omics data have become available. Thus, it is urgent to develop computational methods to identify cancer genes by integrating these data. We propose an integrative rank-based method called iRank to prioritize cancer genes by integrating multi-omics data in a unified network-based framework. The method was used to identify the disease genes of hepatocellular carcinoma (HCC) in humans using the multi-omics data for HCC from TCGA after building up integrated networks in the corresponding molecular levels. The kernel of iRank is based on an improved PageRank algorithm with constraints. To demonstrate the validity and the effectiveness of the method, we performed experiments for comparison between single-level omics data and multiple omics data as well as with other algorithms: random walk (RW), random walk with restart on heterogeneous network (RWH), PRINCE and PhenoRank. We also performed a case study on another cancer, prostate adenocarcinoma (PRAD). The results indicate the effectiveness and efficiency of iRank which demonstrates the significance of integrating multi-omics data and multiplex networks in cancer gene prioritization.

1. Introduction

Cancer is a leading cause of death worldwide and is a major threat for human health [1]. Cancer's pathogenesis is far from fully understood, but it is regarded as resulting from the complicated interplays between external environmental factors and internal genetic factors [2]. Identifying cancer genes could reveal the heredity of internal factors, which could highly benefit our understanding of the pathogenesis of cancer.

With the emergence and advances of high-throughput technologies, more and more omics data in multiple molecular levels are becoming available such as cancer consortiums of The Cancer Genome Atlas (TCGA) [3] and International Cancer Genome Consortium (ICGC) [4]. The open access multi-omics data have brought about unprecedented opportunities to decipher the pathogenic mechanism of cancer by discovering cancer genes. It is urgent to develop informatics methods and tools to meet the challenge of discovering valuable knowledge from these generated biomedical big data [5], such as cancer gene prioritization.

In recent years, network-based models have been ubiquitous in complex disease research because of their powerful ability to systematically organize and investigate the underlying connectome between biomolecules [6]. For instance, they have been used in the identification of candidate disease genes [7] and diagnostic biomarkers from high-throughput omics data [8]. In network science, the PageRank (PR) algorithm [9] is a discriminative method for ranking webpages. It ranks individual nodes in a network by ranking the PR values of random walkers when they obtain a state of convergence. Using a similar philosophy, it has been used to rank disease genes in bio-molecular interaction networks.

Köhler et al. employed a random walk algorithm with restart to define the similarity between genes within an interaction network. They then ranked candidate disease genes based on their similarities to known disease genes [10]. Li and Parta applied the algorithm in a heterogeneous network to rank genes and their corresponding phenotypes simultaneously [11]. Cowen et al. described the importance of network propagation and summarized three kinds of mature versions of the PR algorithm and their applications [12].

Valdeolivas et al. extended the PR algorithm into multiple networks to rank disease genes [13]. Vanunu et al. proposed a method named PRINCE [14], which identifies disease genes by associating genes and

* Corresponding author. Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China.
E-mail address: zpliu@sdu.edu.cn (Z.-P. Liu).

https://doi.org/10.1016/j.compbiomed.2020.103692
Received 12 December 2019; Received in revised form 10 February 2020; Accepted 29 February 2020
Available online 5 March 2020
0010-4825/

protein complexes with a disease via network propagation. Recently, Cornish et al. proposed a method called PhenoRank, which prioritizes genes through a protein-protein interaction network by integrating different information such as phenotypic types [15].

Molecules perform functions by jointing with others in multi-level organizations, such as gene regulation, protein-protein interaction, and protein modification. A network provides an effective model to describe the complex interactions between these molecules. Current omics data are often focused on profiling single-level molecular information, such as the RNA-seq technique, which is used to measure the abundance of transcripts. Integrating these multi-omics data with multi-level networks could produce more reliable results that can be used to validate identifications.

To date, only a few methods have been proposed to integrate multi-omics data, especially the corresponding samples in multiple molecular levels. Wang et al. proposed a method to stratify cancer subtypes based on the fusion of similarity networks by aggregating data types on genomic scales [16]. Cho et al. proposed a method called Mashup, which is an integrative framework for analyzing the topology of multiple interaction networks from heterogeneous data sources to infer various functional properties of genes or proteins [17]. Similarly, Christos et al. proposed a method called NetICS, which prioritizes cancer genes by integrating diverse types of molecular data in independent samples in a directed functional interaction network to calculate mediator effects [18]. These methods are not for inter-correlated omics datasets, and it is of interest to develop a general framework to integrate multi-level omics data to identify disease genes from multi-level networks.

In this study, we developed a computational method called iRank to prioritize cancer genes by integrating multi-omics data in a network-based framework including RNA-seq, DNA methylation, somatic mutation, miRNA-seq and copy number variant data. We demonstrate the effectiveness of iRank in identifying the cancer genes of HCC by integrating multi-omics data from TCGA. The knowledge-based regulatory network documented in RegNetwork [19] and other kinds of networks are employed to construct the working integrative networks. Then, we prioritize these cancer genes from the weighted interactome specified by the multi-omics data.

The rationality of data integration is demonstrated by comparison with those of individual and partial levels of data. To show the advantage of iRank, We compare the identification results with the other methods in searching cancer genes: RW [10], RWH [11], PRINCE [14], Phenorank [15]. We also verify our method with another cancer by searching cancer genes for PRAD by integrating multi-omics data with multi-networks.

2. Materials and methods

2.1. Data sets

For a proof-of-concept introduction, we prioritize the cancer genes of HCC compiled from KEGG [20] and Malacards [21]. Altogether, 33 genes were found to be causal in the development and progression of HCC (Supplementary Table S1.1). The multi-omics data of HCC were downloaded from TCGA [22]. The raw data were processed with the tools of TCGABiolinks [23], TCGA Assembler 2 [24] and cBioPortal [25]. We included the following multi-level omics data: DNA methylation, copy number variant, somatic mutation, miRNA-seq and RNA-seq data. Due to the non-availability of matched control and tumor samples, we excluded HCC proteomics data in TCGA. After matching the consistent samples with multi-level omics information, 37 samples with both control and tumor annotations were finally chosen. The details of the data preparation and normalization are available in Supplementary Tables S1.2 and S1.3.

We used a multiplex network to organize the corresponding multi-omics data. In the cross-level interactome, we selected the gene regulatory network (GRN) as a core layer for its crucial importance in biological processes. We downloaded an integrative human GRN from RegNetwork [19] and the miRNA regulatory interactions from miRTarbase [26]. Additionally, a protein-protein interaction network (PPIN) was constructed using several databases: STRING [27], Bind [28], BioGrid [29], HPRD [30], IntAct [31], and MINT [32].

For the cross-level molecular interactions, we linked the direct interactions from the DNA-level nodes to the corresponding RNA-level nodes according to the central dogma of molecular biology. Altogether, a multiplex network was built up in six levels with 94,173 nodes and 907,211 interactions. After combining the multi-networks with the corresponding multi-omics data, 63,205 nodes and 225,490 edges were left (see Supplementary Table S1.4 for statistics). All data and source code used in this study are available at https://github.com/zpliulab/iRank.

2.2. Framework

Fig. 1 shows the framework of iRank, which mainly contains six steps. As shown in Fig. 1(a) and (b), the multi-omics data of HCC are downloaded from TCGA, and the comprehensive bio-molecular networks in multiple levels are constructed from various databases. For clarity, we use GRN as a core-level network and other-level networks provide cross-talking (epi)genetic information. We also build up the interactions across the different levels. We integrate the prior networks with the multi-omics profiling data by weighing the interactions with differential mutual information (DMI) between the control and tumor samples (Fig. 1(c)).

We propose a constrained PageRank (CPR) algorithm on the weighted multiplex network. In each network, the PR values of nodes are achieved via CPR as shown in Fig. 1(d). After aggregating the ranks of multiple networks (Fig. 1(e)), the final rank of each node are obtained as the output of iRank, as shown in Fig. 1(f).

2.3. Integration of multiplex network with multi-omics data

The comprised multiplex network is an integrated human interactome. We employ the adjacent matrix A to describe the intra-network at the individual molecular level and the inter-network across different molecular levels. In both types of networks, if the i-th node has an interaction with the j-th node, we define the corresponding element $A_{ij} = 1$. Some intra-networks are directed, such as, GRN, while others are undirected, such as, PPIN. For the inter-networks, we assign the directions between nodes according to the flow of genetic information within a biological system.

We map the molecular profiles onto the integrated multiplex network by weighing these edges, and we calculate the mutual information (MI) of the measurements between two nodes, such as, X and Y, of an edge. We employ a straightforward approach to estimate MI [33], which firstly partitions the variables into bins with finite size N whose set is dependent on the variable size. This is done by drawing grids on a scatterplot of the two variables. The number of points of X (or Y or X and Y) falling into these girds is counted, and then MI is approximated as:

$$I(X,Y) = \sum_{i,j} \log \frac{p(i,j)}{p_x(i)p_y(j)}$$

where $p_x(i) \approx \frac{n_x(i)}{N}$, $p_y(j) \approx \frac{n_y(j)}{N}$, $p(i,j) \approx \frac{n(i,j)}{N}$, $n_x(i)$ is the number of points falling into the i-th gird of X, $n_y(j)$ is the number of points falling into the j-th gird of Y and $n(i,j)$ is the number of points in their intersections.

For each gene with the two states of control and disease, we identify the DMI value on each edge in the multiplex network:

$$\mathrm{DMI} = |\mathrm{MI}(X_c, Y_c) - \mathrm{MI}(X_d, Y_d)|$$

DMI represents the absolute difference between the two MI values of control ('c') and disease ('d') states.

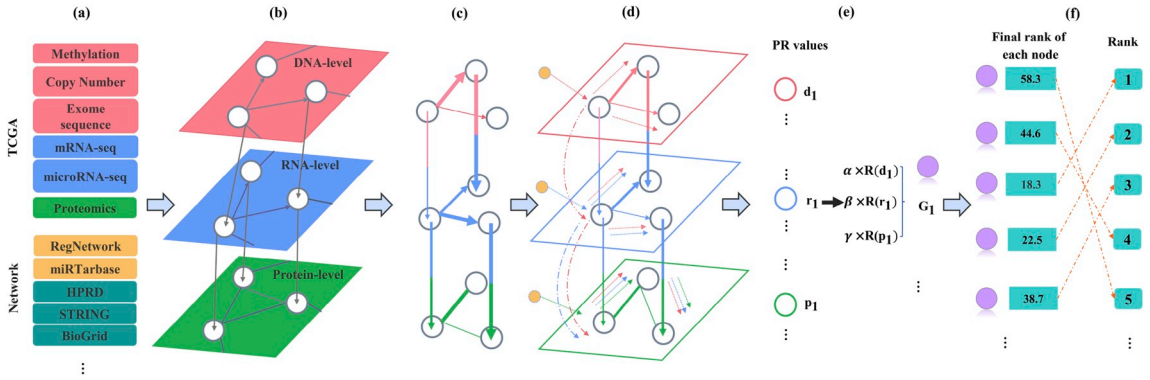

Fig. 1. The framework of iRank. Different colors represent different levels of molecular information, pink: DNA; blue: RNA; and green: protein. $\alpha, \beta,$ and γ represent the corresponding weights of different levels. R (·) represents the rank of a node in its corresponding layer of network. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Next, DMI is normalized by the min-max method. More details about calculating MI in different omics data are available in Supplementary Table S2.1. We also tested other correlation definitions such as Pearson's correlation coefficient to weigh the multi-layer networks for justification purposes (see Supplementary Tables S2.2-S2.5).

2.4. CPR algorithm on weighted multiplex network

We used DMI values to weigh these edges to obtain the weighted multiplex network. The weighted network is a kind of content-specific network after removing the edges whose weights are 0. Then, we obtain the orders of nodes by the CPR algorithm, which is based on PR with the constraints on information flows in the intra-layer networks and inter-layer networks across different levels.

Let T be the transition matrix of the multiplex network. Referring to the genetic flows from DNA to protein, some elements in T will be set to be 0 because there is no information transfer between the corresponding two nodes. We set GRN as the central in bridging genotype and phenotype. The information from the other layers is reflected with gene regulation. We denote the interactions between DNA methylation, somatic mutation, copy number variation, and GRN as DRG, SRG, and CRG, respectively.

We set seeds of random particles at DRG, CRG, SRG, GRN in CPR. When implementing the method in the multiplex network, if a random particle stays in GRN, it has two choices: continually walking in GRN or transitionally walking into the other layers of network. When random walkers spread in the multiplex network, we use two parameters to describe their trajectories. When they spread in GRN, let λ be a jumping probability that quantifies the chance of random particles walking in GRN, and let $(1-\lambda)$ be the probability of the particles walking out of GRN to the other layers of the network. Let $\phi = (\phi_d, \phi_c, \phi_s)^T$ be the cross jumping probability which quantifies the chances of random particles from the other three inter-layer networks walking into GRN, (i.e., DNA methylation (d), somatic mutation (s) and copy number variation (c)). When implementing our method on these multi-omics data, a grid search technique is adapted to search for the best parameters ϕ_d, ϕ_c, ϕ_s. The details about the parameters optimization are presented in Supplementary Tables S2.6-S2.8.

Each element of T can be defined as follows. For a transition matrix in the intra-layer, such as, T_{GG} in GRN, the probability of a random walker ($g_i \to g_j$) at the i-th row and j-th column is defined as:

$$(T_{GG})_{i,j} = \begin{cases} \lambda \times (GW)_{ji} \times G_{i,j} \Big/ \sum_j G_{i,j}, \text{if } \sum_j G_{ij} \neq 0 \\ 0, \quad \text{otherwise} \end{cases}$$

where G is the adjacent matrix of GRN, and GW is G weighted by DMI.

For the inter-layer transition matrix $T_{\bullet G}$, the probability of a random walker ($\bullet_i \to g_j$) at the i-th row and j-th column is defined as:

$$(T_{\bullet G})_{i,j} = \mathrm{Pro}(g_j|\bullet_i) = \begin{cases} \phi_\bullet \times \bullet W_{ij} \times \bullet G_{ij} \Big/ \sum_j \bullet G_{ij}, \quad \text{if } \sum_j \bullet G_{ij} \neq 0 \\ 0, \quad \text{otherwise} \end{cases}$$

Where·refers to each of the other three layers.

The CPR algorithm is summarized in Table 1. Firstly, we define the constraint matrix C based on prior knowledge about the transition matrix T. For the direction and other specifications in genetic information flows, we set the corresponding element in C to be 1, which refers to the constraint transition. Of course, the credibility can also be represented in the constraint matrix. Then, we can formulate the details of CPR on the multi-layer heterogeneous networks.

Let O be the initial probability vector, and let I_i be the vector in which the i-th element holds the probability of finding a random walker at node i at the current step t. The probability vector at step t+1 is given by:

$$I_{i \to j} = r \times C \otimes T^T \times I_i + (1-r) \times O$$

where $j = i+1$ is the next point of i, and $(1-r)$ ($r \in (0,1)$) is the restart probability which represents the chance of random walker going back to the seed nodes, $\otimes$ indicates Hadamard product for a matrix.

After iterating some steps, the probability will reach a steady state, which can be obtained by performing the iterations until the difference between I_i and I_j (measured by the L1 norm) falls below a threshold, such as, 10^{-10}. In this study, we set (1-r) to 0.85 and λ to 0.5 on empirical trails.

2.5. Rank aggregation

After applying CPR to multi-layer networks, each node achieves a PR value in each level and we rank these nodes by their PR values in each level. Then, we implement the mean rank aggregation method to obtain the final rank by:

$$R_f(i) = \sum_{k=1}^{n} w_k R_k(i)$$

where k refers to a single-layer network used in the multiple n layers, w_k represents the weights of k-th layer of networks, $R_k(i)$ represents the rank of the i-th nodes in the k-th layer network, and $R_f(i)$ represents the final rank of the i-th node. Here, we simply set $w_k = \frac{1}{4}$ for 4 layers of networks; i.e., GRN, DRG, CRG, and SRG.

Table 1
The CPR algorithm.

Input:

Network GRN, DRG, CRG, SRG

Calculation:

Calculate DMI for each edge in the multiple networks

The number of nodes in multiplex networks (denoted as N)

Selection:

Select edges with DMI > 0

Construction:

Construct transition matrix T and constrained matrix C of multiple networks matrix

Initialization:

Parameters: $\lambda = 0.5$, $1 - r = 0.85$, φ_d=0.1, φ_c=0.1, φ_s=0.1

PR_0 : a N-length column vector in which each value is 1/N

PR : the same with PR_0

del_P=*inf*

iter=1

CPR:

while del_PR > $1 \times e^{-10}$ || iter < 200

$$PR_{pre} = PR$$

$$PR = r \times C \otimes PR_{pre} + (1 - r) \times PR_0$$

$$del\_PR = norm(PR - PR_{pre})$$

$$iter = iter + 1$$

end while

Rank aggregation:

$$R_f(i) = \sum_{k=1}^{n} w_k R_k(i)$$

w_k : the weights of different-layer networks

$R_k(i)$: the rank of the *i*-th node in the *k*-th layer network

$R_f(i)$: the final rank of the *i*-th node

n: the number of network layers

Output:

Sort genes according to R_f

3. Results and discussion

3.1. Prioritization of cancer genes

We first used iRank to identify the ranks of HCC genes. Fig. 2(a) presents boxplots of the ranks from different strategies of combining networks and omics data. We first rank the disease genes of HCC in GRN due to its centrality in our setting. ORI represents the ranks obtained in the original GRN, which means only the network topology of GRN is considered. The other boxplots show the results of experiments on gene ranking by integrating multi-layer networks with the corresponding multi-omics data. ORIr represents the ranking results of HCC genes in the original GRN weighted by RNA-seq data, in which both the network topology and gene expression profile have been integrated for gene prioritization.

We also ran iRank in the multi-layer networks sequentially by adding multi-level information, i.e., DNA methylation (d), somatic mutation (s) and miRNA interference (m). ORIrd refers to the ranking results in the multiplex network obtained by integrating RNA-seq and DNA methylation. ORIrds refers to the ranks obtained by integrating RNA-seq, DNA methylation, and somatic mutation. We find that the cancer genes tend to achieve higher ranks when more molecular levels of omics data were integrated in the prioritization. Supplementary Figures S3.1-S3.4 show more boxplots of the ranks of HCC disease genes by integrating combinatorial levels of information.

High ranks indicate the effectiveness of iRank in the prioritization of cancer genes. Fig. 2(a) show that cancer genes gradually achieve higher ranks with smaller mean values and standard deviations when we integrate more and more levels of omics information in most of the experiments. The mean rank of cancer genes has been improved from 3255 in ORI to 1069 in ORIrdms when we used multi-omics data. When we integrated a few types of certain omics information, such as, copy number variation, the ranks slightly decreased. However, more empirical tests show that the cancer genes tend to achieve higher ranks when

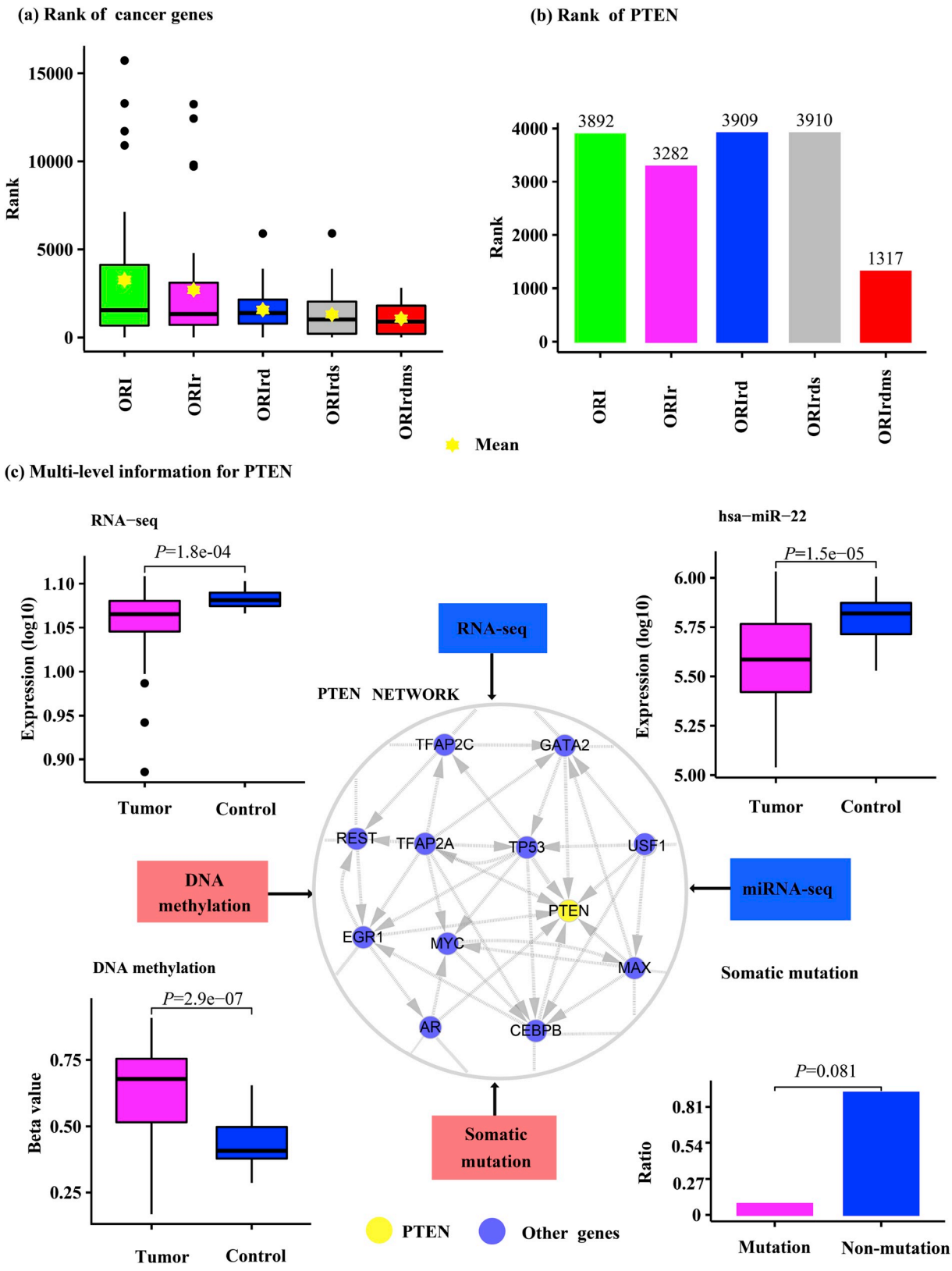

Fig. 2. The ranks for HCC disease genes by integrating different omics data. (a) Boxplot of the ranks of cancer genes. Star point refers to the mean rank; (b) Ranks of PTEN; (c) Multi-molecular levels of information for PTEN. The PTEN network presents partial gene regulations between PTEN and some of its first-order neighbor genes. The difference of multi-omics information for PTEN between control and tumor samples are shown respectively. In the post-transcriptional level, has-miR-22 is a representative miRNA that regulates PTEN. The somatic mutation refers to the ratio of samples with mutated/non-mutated PTEN. *P* value is given by two-side Wilcoxon test.

more levels of omics data are integrated in the prioritization, as shown in Supplementary Figures S3.1-S3.4.

To illustrate the results specifically, Fig. 2(b) shows the ranks of a cancer gene, PTEN. As shown, its rank has been improved from 3892 by only using the ORI information to 1317 by integrating the multi-omics data of ORIrdms. We note that it achieves slightly lower rank when iRank uses the information of ORIrd or ORIrds. However, it obtains higher ranks when we combined more molecular levels of omics data. In details, Fig. 2(c) demonstrates the multiple molecular levels of information for PTEN, i.e., GRN, RNA-seq, DNA methylation, miRNA, and somatic mutation. In some molecular level, e.g., RNA-seq, PTEN contains significantly differential information between control and tumor samples. When we integrated this information into iRank, PTEN obtains a higher rank compared to the usage of original GRN. As shown in Fig. 2 (a), the ranks of cancer genes have globally been improved when we integrated multi-omics data. For a cancer gene, its multi-level information, as shown in Fig. 2(c) for PTEN, provides a comprehensive characterization of its profiles. When we combined them in the integrated framework of iRank, they provided cross-molecular-level validation for the final prioritization.

At this point, the networks used in iRank are not complete. To test the robustness of our method, we added perturbations by randomly adding 30,000, 60,000, and 90,000 edges to the original networks. We add these permutation edges in the different-level of networks for different networks combined with different omics data (ORI, ORIr, and ORIrd). The ranks of cancer genes under different permutations are shown in Fig. 3. The results show that iRank can stably obtain the ranks of HCC genes in the network permutation experiments. Furthermore, the top-ranked genes are also not affected by these network permutations. We also integrated the top-100 ranked genes in different cases of adding permutation edges. The genes share significant intersections in the three permutations (Supplementary Tables S3.1-S3.3).

3.2. Classification results

To show the discrimination of achieved ranks in cancer genes, we used the final PR values obtained by different strategies of data integration as the decision values for classifying disease genes and normal ones. For a comparison study, we ran iRank in the different levels of weighted networks individually. To obtain a baseline, we randomly chose the same number of normal genes and obtained their PR values via iRank. We repeated the experiment 1000 times and obtained the performance metrics in the classifications. The mean receiver operating characteristic (ROC) curves and boxplots of the area under the curve (AUC) values are shown in Fig. 4. Table 2 shows the details of classification metrics about sensitivity (SE), specificity (SP), accuracy (ACC), F1 score (F1) from the 1000 experiments.

In Fig. 4, when iRank runs on ORI, it obtains a mean AUC value of 0.834. By integrating the original network with the multi-omics data, iRank obtains slightly better classification performances with mean AUC value of 0.839 for ORIr, 0.895 for ORIrd, 0.911 for ORIrds, and 0.910 for ORIrdms. This indicates the effectiveness of classifying disease genes and normal genes by integrating multi-omics data with the multi-layer heterogeneous networks.

Additionally, we obtained the classification results of combinational integration of these omics data at different levels. For the addition of single-level omics data with ORI in GRN, RNA-seq data are the best information for improving the AUC value (0.839). For the two-level data, the combination of RNA-seq and DNA methylation data obtains the highest AUC value (0.895). The results indicate the importance of DNA methylation when compared with somatic mutation, miRNA expression and copy number variant information. Supplementary Figures S3.5-S3.8 show more detailed results about classification results by integrating different omics information.

In Table 2, ORIrds achieves a quite high AUC value of 0.911 and high SE value of 0.925. Moreover, ORIrdms achieves a high AUC value of 0.910, a high SE of 0.984, F1 score of 0.893 and ACC of 0.881, but it has a low SP of 0.778. The results provide more evidence for the effectiveness of integrating multi-omics information in prioritizing cancer genes.

According to these AUC values, the RNA-seq data obtains the best classification results when only one type of data is integrated with the original network. Then, the DNA methylation improves the classification most effectively in comparison to the other levels of information, such as, miRNA, somatic mutation and copy number variation. This indicates the crucial role of DNA methylation in the epigenetic regulation of gene expression, which has been proven in experiments. These identified cancer genes and related variants could potentially be employed as specific diagnostic biomarkers for HCC [34]. We also calculated the classification performance of the other combinatorial strategies of data integration. Supplementary Tables S3.4-S3.7 show more results about classification performances in different strategies of data integration.

3.3. Comparisons with other methods

Other methods are also available for prioritizing disease genes, such as RW [10], RWH [11], PRINCE [14], and PhenoRank [15]. These methods are not designed for integrating multi-omics data, so we compared iRank with them in the prioritization of HCC genes in GRN only. Fig. 5 shows the mean ROC curves and boxplots of AUC values of these methods in 1000 experiments and Table 3 gives a detailed comparison of the results of different indexes. From the classification results, iRank achieves the AUC of 0.839 that is marginally better than the newly available method PRINCE and significantly better than the other methods. Note that the performance of iRank is based on the information of ORIr. As shown in Fig. 4, it performs much better by using multi-omics data, e.g., AUC of 0.911 in ORIrds.

3.4. More case study on PRAD

To validate the generality of iRank, we also performed all the steps on another cancer, PRAD. All processes were run in the same manner for HCC. Fig. 6 illustrates the mean ROC curves and boxplots of AUC values of the classification performance with the obtained PR values. In Fig. 6,

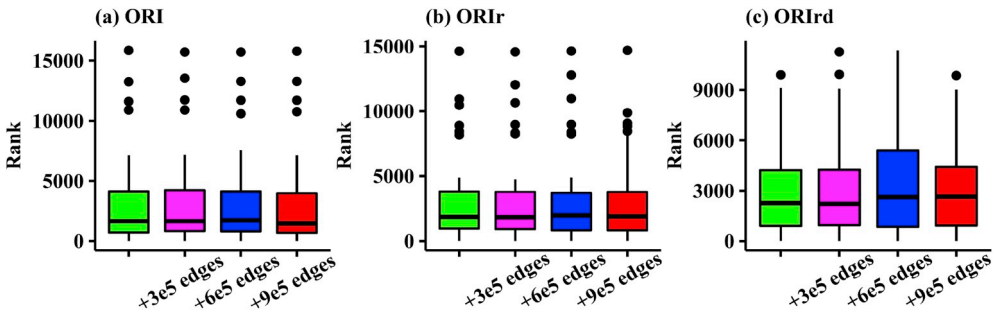

Fig. 3. Boxplots of ranks of cancer genes under different network perturbations when adding more edges.

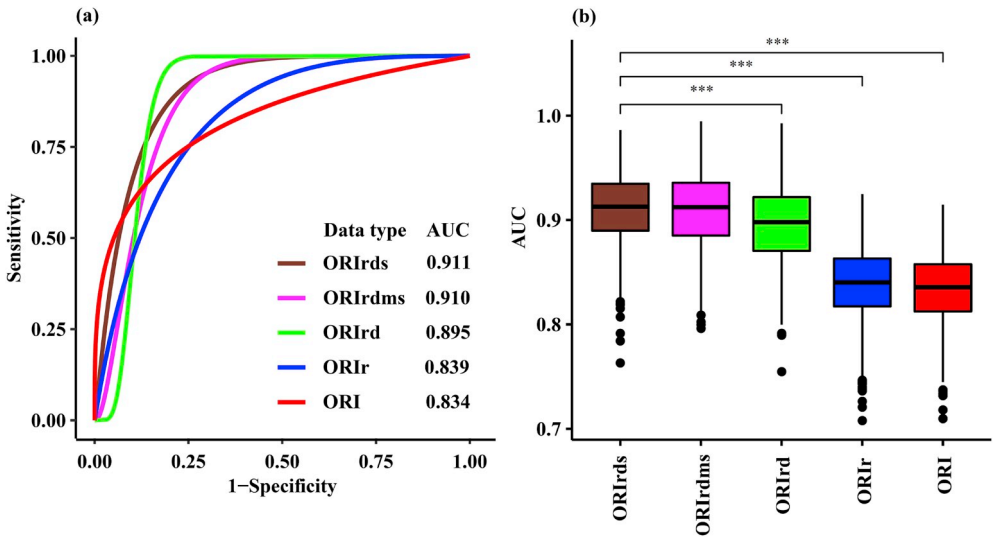

Fig. 4. Classification results of integrating different omics information. (a) Smoothened mean ROC curves of 1000 experiments. (b) Boxplots of AUCs in the corresponding 1000 experiments. *** P-value < e−10, **P < e−05 and * P < 0.05 by two-side Wilcoxon test.

Table 2
Classification performances of different strategies of data integration (ordered by AUC). Values are the mean ± standard deviation.

Data	SE	SP	ACC	F1	AUC
ORIrds	0.925 ± 0.041	0.814 ± 0.073	0.869 ± 0.032	0.877 ± 0.028	0.911 ± 0.033
ORIrdsm	0.984 ± 0.030	0.778 ± 0.073	0.881 ± 0.036	0.893 ± 0.031	0.910 ± 0.033
ORIrd	0.934 ± 0.037	0.804 ± 0.072	0.869 ± 0.033	0.877 ± 0.028	0.895 ± 0.038
ORIr	0.827 ± 0.050	0.801 ± 0.078	0.814 ± 0.033	0.817 ± 0.029	0.839 ± 0.034
ORI	0.792 ± 0.071	0.802 ± 0.083	0.797 ± 0.030	0.796 ± 0.032	0.834 ± 0.033

different colors represent different strategies of prioritizing cancer genes of PRAD with different strategies of data integration. Table 4 shows the details of the classification metrics of iRank on PRAD.

From Fig. 6 and Table 4, ORIrdm achieves the higher AUC value than those of ORIrd and ORIrm. The results confirm the effectiveness of iRank in cancer gene prioritization by data integration. Moreover, ORIrd obtains better results than ORIrm and ORIrc. This also indicates the importance of DNA methylation information in improving the identification of cancer genes. Notably, the results highlight the necessity of

Table 3
Classification performance of different methods.

Methods	SE	SP	ACC	F1	AUC
iRank	0.827 ± 0.050	0.801 ± 0.079	0.814 ± 0.033	0.817 ± 0.029	0.839 ± 0.034
PRINCE [14]	0.800 ± 0.085	0.768 ± 0.097	0.784 ± 0.034	0.786 ± 0.037	0.838 ± 0.032
RW [10]	0.792 ± 0.072	0.802 ± 0.083	0.797 ± 0.030	0.796 ± 0.032	0.834 ± 0.033
RWH [11]	0.754 ± 0.063	0.776 ± 0.090	0.765 ± 0.034	0.763 ± 0.032	0.800 ± 0.037
Phenorank [15]	0.819 ± 0.051	0.714 ± 0.088	0.767 ± 0.037	0.778 ± 0.031	0.795 ± 0.038

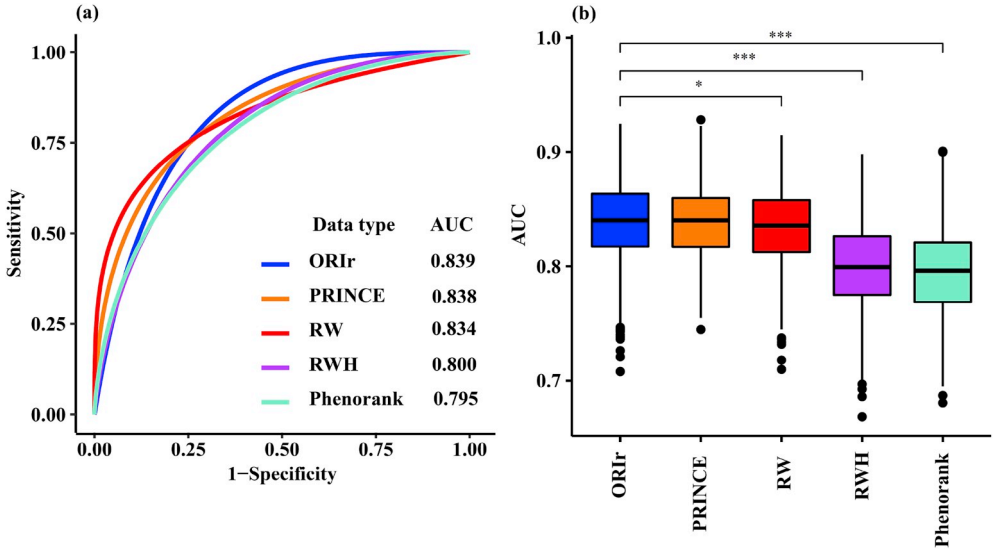

Fig. 5. ROC curves and boxplots of AUC values achieved by different methods. (a) Smoothened mean ROC curves of 1000 experiments of different methods. (b) Boxplots of AUC values of different methods in the corresponding 1000 experiments. *** P-value < e−10, **P < e−05, and * P < 0.05.

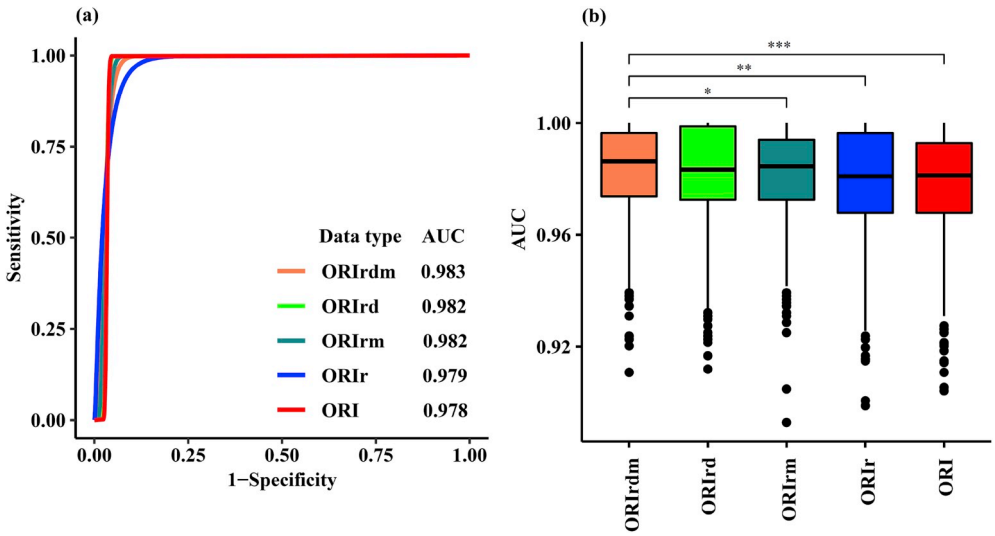

Fig. 6. ROC curves and boxplots of AUC values of different strategies of data integration in PRAD. (a) Smoothened mean ROC curves of 1000 experiments. (b) Boxplots of AUCs in the corresponding 1000 experiments. *** P-value < e−10, **P < e−05, and * P < 0.05.

Table 4
Classification performance of different data integration strategies in PRAD.

Data	SE	SP	ACC	F1	AUC
ORIrdm	0.946 ± 0.039	0.986 ± 0.020	0.966 ± 0.023	0.965 ± 0.024	0.983 ± 0.016
ORIrm	0.944 ± 0.040	0.966 ± 0.024	0.955 ± 0.024	0.954 ± 0.024	0.982 ± 0.016
ORIrd	0.952 ± 0.039	0.993 ± 0.015	0.973 ± 0.022	0.972 ± 0.022	0.982 ± 0.018
ORIr	0.945 ± 0.041	0.979 ± 0.020	0.962 ± 0.023	0.961 ± 0.024	0.979 ± 0.018
ORI	0.948 ± 0.041	0.961 ± 0.017	0.954 ± 0.023	0.954 ± 0.023	0.978 ± 0.017

investigating epigenetic regulations in cancer pathogenesis [35]. Supplementary Tables S3.8-S3.11 show more results in the cancer gene prioritization of PRAD and more detailed information about PRAD data preparation [36].

3.5. Functional enrichments analysis

We performed a gene ontology (GO) enrichment analysis to detect the dysfunctional implications of HCC disease genes. Specifically, we identified the enriched GO terms of biological processes by g:profiler (https://biit.cs.ut.ee/gprofiler/index.cgi) in the cancer genes. We found the enriched functions are related to the detection of a carbohydrate stimulus and the abnormality of the pancreas (Supplementary Table S3.12). Some functions are included in the cancer genes, such as TERT, TP53, and PTEN, which have been verified to be related to HCC in independent experiments [37].

For the top 100 ranked genes, we also identified the enriched functions underlying them (Supplementary Tables S3.13-S3.15). The enriched GO biological processes were mostly house-keeping functions, such as DNA-templated (GO:0006351), nucleic acid-templated transcription (GO:0097659) and RNA biosynthetic processes (GO:0032774). In the enriched KEGG pathways, some pathways related to HCC have also been enriched, such as the pathways of hepatitis C, hepatitis B, and hepatocellular carcinoma. The functional analysis indicates that the top-ranked genes are highly related to the disorder of fundamental functions in HCC. In this work, we only evaluated the performance of iRank in identifying the known casual cancer genes. These top-ranked genes might also potentially be the novel cancer genes after further validations.

4. Conclusions

We have proposed iRank to prioritize cancer genes from multi-layer networks by integrating multi-omics data. The multiple molecular information provides cross-level validations for the variants in cancer and is expected to improve the identification of disease-causing genes more accurately and confidently.

Specifically, we extracted the DMIs by comparing tumor samples with controls and used them to weigh the corresponding multi-layer networks. We then improved the PR algorithm with the constraints of genetic flow and prior knowledge. In our case studies, we prioritized the cancer genes in two complex diseases, HCC and PRAD, from TCGA. The ranks and classifications in these known cancer genes proved the effectiveness of iRank. The comparison studies with the available methods also provided evidence of the advantages of iRank.

Our results also proved the necessity and benefit of integrating multi-omics data in the prioritization of cancer genes. Different level of data gives different characteristics of genes. It is rational to organize them according to the central dogma. In most cases, the ranks of cancer genes can be improved by integrating multi-omics data. Besides, the classification performance was improved when we empirically integrated more levels of omics data in the method. As expected, the better performance in the strategies of data integration generates more discriminative ranks for cancer genes. Among these multi-omics data, DNA methylation significantly improved the prioritization of cancer genes based on the transcriptomic level when compared to the other levels of information. However, this is also the limitation of iRank. It needs the corresponding molecular levels of omics data for both control and tumor samples. The strict requirement of inter-correlated data specifies the application scope and scenario of iRank. Moreover, some characteristics of iRank, such as the PR algorithm with constraints and the employment of GRN centrality, also provide more options and hints for data integration in biomedicine.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC) under grant numbers 61973190, 61572287 and 61533011; Key Research and Development Project of Shandong Province, China under grant number 2018GSF118043; the Innovation Method Fund of China (Ministry of Science and Technology of China) under grant number 2018IM020200; the Program of Qilu Young Scholars of Shandong University.

Declaration of competing interest

None declared.

Acknowledgements

Thanks are due to the editor-in-chief and anonymous reviewers for their valuable comments and suggestions which greatly improve our paper. We thank all the other members of our lab at Shandong University for their assistance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2020.103692.

Appendix B. Abbreviations

Table B.1 lists all the abbreviations used in the main text.

Table B.1
Abbreviations used in this paper

Original words	Abbreviations
hepatocellular carcinoma	HCC
prostate adenocarcinoma	PRAD
random walk	RW
random walk with restart on heterogeneous networks	RWH
PageRank	PR
gene regulatory network	GRN
protein-protein interaction network	PPIN
differential mutual information	DMI
constrained PageRank	CPR
mutual information	MI
interactions between DNA methylation and gene regulatory network	DRG
interactions between somatic mutation and gene regulatory network	SRG
interactions between copy number variation and gene regulatory network	CRG
DNA methylation information	d
copy number variation information	c
somatic mutation information	s
miRNA regulation information	m
receiver operating characteristics	ROC
area under curve	AUC
sensitivity	SE
specificity	SP
accuracy	ACC
F1 score	F1

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA A Cancer J. Clin. 69 (1) (2019) 7–34.
[2] P. Lichtenstein, et al., Environmental and heritable factors in the causation of cancer — analyses of cohorts of twins from Sweden, Denmark, and Finland, N. Engl. J. Med. 343 (2) (2000) 78–85.
[3] J.N. Weinstein, et al., The cancer Genome Atlas pan-cancer analysis project, Nat. Genet. 45 (10) (2013) 1113–1120.
[4] J. Zhang, et al., International cancer genome consortium data portal–a one-stop shop for cancer genomics data, Database (2011) bar026.
[5] A.R. Sonawane, et al., Network medicine in the age of biomedical big data, Front. Genet. 10 (2019) 294.
[6] A.-L. Barabási, Z.N. Oltvai, Network biology: understanding the cell's functional organization, Nat. Rev. Genet. 5 (2004) 101–103.
[7] Y. Moreau, L.-C. Tranchevent, Computational tools for prioritizing candidate genes: boosting disease gene discovery, Nat. Rev. Genet. 13 (2012) 523–536.
[8] Z.-P. Liu, Identifying network-based biomarkers of complex diseases from high-throughput data, Biomarkers Med. 10 (6) (2016) 633–650.
[9] L. Page, et al., The PageRank Citation Ranking: Bringing Order to the Web, Technical Report. Stanford InfoLab, 1999.
[10] S. Köhler, et al., Walking the interactome for prioritization of candidate disease genes, Am. J. Hum. Genet. 82 (4) (2008) 949–958.
[11] Y. Li, J.C. Patra, Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network, Bioinformatics 26 (9) (2010) 1219–1224.
[12] L. Cowen, et al., Network propagation: a universal amplifier of genetic associations, Nat. Rev. Genet. 18 (2017) 551–562.
[13] A. Valdeolivas, et al., Random walk with restart on multiplex and heterogeneous biological networks, Bioinformatics 35 (3) (2019) 497–505.
[14] O. Vanunu, et al., Associating genes and protein complexes with disease via network propagation, PLoS Comput. Biol. 6 (1) (2010) e1000641.
[15] A.J. Cornish, A. David, M.J.E. Sternberg, PhenoRank: reducing study bias in gene prioritization through simulation, Bioinformatics 34 (12) (2018) 2087–2095.
[16] B. Wang, et al., Similarity network fusion for aggregating data types on a genomic scale, Nat. Methods 11 (2014) 333–337.
[17] H. Cho, B. Berger, J. Peng, Compact integration of multi-network topology for functional analysis of genes, Cell Syst. 3 (6) (2016) 540–548.
[18] C. Dimitrakopoulos, et al., Network-based integration of multi-omics data for prioritizing cancer genes, Bioinformatics 34 (14) (2018) 2441–2448.
[19] Z.-P. Liu, et al., RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse, Database (2015), bav095.
[20] M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[21] C.P. Morrey, et al., MalaCards: an integrated compendium for diseases and their annotation, Database (2013), bat018.
[22] A. Ally, et al., Comprehensive and integrative genomic characterization of hepatocellular carcinoma, Cell 169 (7) (2017) 1327–1341, e23.
[23] A. Colaprico, et al., TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data, Nucleic Acids Res. 44 (8) (2016) e71.
[24] L. Wei, et al., TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data, Bioinformatics 34 (9) (2017) 1615–1617.
[25] E. Cerami, et al., The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Canc. Discov. 2 (5) (2012) 401–404.
[26] C.-H. Chou, et al., miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions, Nucleic Acids Res. 46 (D1) (2018) D296–D302.

[27] D. Szklarczyk, et al., The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible, Nucleic Acids Res. 45 (D1) (2017) D362–D368.
[28] C. Alfarano, et al., The biomolecular interaction network database and related tools 2005 update, Nucleic Acids Res. 33 (D1) (2005) 418–424.
[29] C. Stark, et al., BioGRID: a general repository for interaction datasets, Nucleic Acids Res. 34 (D1) (2006) 535–539.
[30] S. Peri, et al., Human protein reference database as a discovery resource for proteomics, Nucleic Acids Res. 32 (D1) (2004) D497–D501.
[31] S. Kerrien, et al., IntAct–open source resource for molecular interaction data, Nucleic Acids Res. 35 (D1) (2007) 561–565.
[32] A. Chatr-aryamontri, et al., MINT: the Molecular INTeraction database, Nucleic Acids Res. 35 (D1) (2007) D572–D574.
[33] P.E. Meyer, F. Lafitte, G. Bontempi, minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information, BMC Bioinf. 9 (2008) 461.
[34] J. Cheng, et al., Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers, Genome Med. 10 (1) (2018) 42.
[35] X. Zhou, et al., Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser, Nat. Biotechnol. 33 (4) (2015) 345–346.
[36] A. Abeshouse, et al., The molecular taxonomy of primary prostate cancer, Cell 163 (4) (2015) 1011–1025.
[37] D.-C. Lin, et al., Genomic and epigenomic heterogeneity of hepatocellular carcinoma, Canc. Res. 77 (9) (2017) 2255–2265.