COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Network-based prioritization of cancer biomarkers by phenotype-driven module detection and ranking

Haixia Shang, Zhi-Ping Liu *

Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China

ARTICLE INFO

Article history:
Received 1 July 2021
Received in revised form 13 November 2021
Accepted 4 December 2021
Available online 08 December 2021

Keywords:
Phenotype-driven module detection
Block-based module ranking
Network biotechnology
Machine learning
Biomarker discovery

ABSTRACT

This paper describes an ensemble method with supervised module detection and further module prioritization for reliable network-based biomarker discovery. We design a module detection and ranking method called mRank to discover reliable network modules as cancer diagnostic biomarkers, with two procedures: (1) an iterative supervised module detection guided by phenotypic states in a specific network, (2) a block-based module ranking locally and globally via network topological centrality. We validate its effectiveness and efficiency by identifying hepatocellular carcinoma (HCC) network modules on a comprehensive gene regulatory network with specifying gene interactions by HCC RNA-seq data from the Cancer Genome Atlas (TCGA). These top-ranked modules by mRank get a mean AUC of 0.995 on TCGA HCC dataset with 371 tumor samples and 50 controls by cross-validation SVM. Based on the prior knowledge of cancer dysfunctions enriched in top-ranked modules, 69 genes are identified as HCC candidate biomarkers. They are further validated in independent cohorts with a classifier trained on TCGA HCC dataset. A mean AUC of 0.846 is achieved in distinguishing 976 disease samples from 827 controls. Moreover, some known HCC signatures such as AFP and SPP1 are also included in our identified biomarkers. mRank enables us to find more reliable network modules for cancer diagnosis. For a proof-of-concept study, we validate it in identifying HCC network biomarkers and it is generalizable to other cancers or complex disease. The overall results have demonstrated that mRank can find effective network biomarkers for cancer diagnosis which result in less false positives.

1. Introduction

Cancer is a major health-threatening problem and a leading cause of death worldwide [1,2]. Until now, although many driver genes have been found, the pathogenesis of cancer is still difficult to figure out. Generally speaking, cancer is caused by the interactions of genetic factors and environmental reasons [3]. Thus, it is of paramount importance to identify diagnostic biomarkers for its early detection, further for personalized treatment [4]. Currently, some biomarkers for hepatocellular carcinoma (HCC) have been found for diagnosis, prognosis and treatment responses [5], such as Alpha-Fetoprotein (AFP), Des-*r*-Carboxy Prothrombin (DCP), osteopontin (OPN), vascular endothelial growth factor (VEGF), angiopoietin 2 (ANG-2), and Golgi protein 73 (GP73). However, these single-gene-based biomarkers often result in poor specificity and accuracy. An alternative way is to discover big-panel-based biomarkers for achieving better diagnosis. For instance, MammaPrint [6] is such kind of endeavor for breast cancer by using 70 genes as biomarkers. A large gene panel not only increases the detection rates of pathogenic variants, but also discovers variants with uncertain pathogenicity, multiple variants and incidental findings [7]. Thus, generating big-panel-based biomarkers is a pressing solution for cancer diagnosis.

Despite biomarker discovery is full of challenges, the development of high-throughput experimental techniques, e.g., gene expression microarray and next-generation sequencing, have generated amounts of omics data which potentially lead to unveil interesting genes and gene products serving as biomarkers [8,9]. The public availability of data from the cooperative cohort projects, e.g., The Cancer Genome Atlas (TCGA) [10] and International Cancer Genome Consortium (ICGC) [11], provide unprecedented opportunities and requests of developing bioinformatics methods for discovering effective biomarker genes which can distinguish phenotypic states and highlight disease mechanisms.

Changes in phenotype typically involve multiple genes acting in concert. Genes/gene products often work together as a functional module in the form of a biomolecular interaction network to play

* Corresponding author.
E-mail address: zpliu@sdu.edu.cn (Z.-P. Liu).

https://doi.org/10.1016/j.csbj.2021.12.005
2001-0370/

crucial roles in mediating complex biological processes [12]. Thus, it is essential to identify these dysregulated gene sets or modules as biomarkers in cancer research. A typical method is gene set enrichment analysis (GSEA), which calculates an enrichment score to decide whether a predefined gene set shows statistically significant differences between two biological states [13]. However, it may result in too many gene sets related with biological states. Then, some derivatives are proposed to improve the enrichment procedure. For instance, gene set analysis (GSA), which made two modifications based on GSEA. One is the max-mean statistic for defining the differential information underlying the gene set. The other is the re-standardization to achieve more accurate inference [14]. Additionally, Barry et al. proposed a method called significance analysis of function and expression (SAFE) to conduct valid tests of gene categories. SAFE implements a permutation-based procedure for accessing the unknown correlations among genes from their gene expression profiles [15]. These gene-set-based methods account for identifying the statistical significance of differences underlying a gene set across different phenotypes. However, it needs predefined gene sets and the interactions between genes underlying the gene sets have not been considered in the identification process.

In bioinformatics fields, network becomes a powerful tool to model the functional interactions among genes/gene products, in which nodes refer to genes and edges refer to their relationships. It has become ubiquitous in representing gene interactions as pathways. Given knowledge-based signaling pathways, it is expected to combine overrepresentation evidence with positions and interactions of genes to prioritize this pathway in response to certain phenotypic state, e.g., signal pathway impact analysis (SPIA) [16]. SPIA measures the actual perturbation on a given pathway under a specific condition, which only takes the overexpression of differential genes and the topology of the current pathway into consideration. However, it scores the pathways individually and independently without considering the cross-talking between pathways. Recently, CRank is proposed to prioritize network communities, which combines multiple community features to produce a final index to rank communities [17]. It involves the network topology in the identification process, and focuses less on the correspondence between gene sets and phenotypic states.

As mentioned, the genes/gene products often perform their roles in the form of a module/community structure in the network [18]. Originally, the modularity in a network is defined as the eigenvectors of a characteristic matrix, i.e., modularity matrix for the network [19]. For module detection, graph clustering (GC) is often used to group sets of "related" vertices in the network [20]. However, the phenotypic information is not included in this kind of module detection method, especially in the supervision of detecting modules with interconnected genes.

In this paper, we propose a module detection and ranking method, called mRank, to discover reliable network modules as cancer diagnostic biomarkers, including two parts, (1) a phenotype-driven module detection, (2) a block-based module ranking. We validate its effectiveness and efficiency by identifying HCC network biomarkers on a comprehensive gene regulatory network (GRN) with specifying gene interactions by HCC RNA-seq data from TCGA. We firstly build up a phenotype-guided technique to detect modules in a genome-wide GRN with specified gene expression profiling data of HCC. Then, we use a block-based ranking technique to prioritize these identified modules locally and globally by introducing network topology and cross-talking information respectively. These top-ranked modules demonstrate strong associations with phenotypes. For comparison study, GSEA [13], GSA [14], SAFE [15], CRank [17], and GC [20] are implemented with predetermined modules by modularity method [19] for obtaining the top-ranked modules and their phenotype association performances are also calculated at the same way. The comparison results illustrate the efficiency and advantage of mRank.

For justification, the network ontology enrichment analysis identifies the functional implications of discovered modules. Based on the prior knowledge of cancer pathogenesis, 69 genes are identified as our identified HCC biomarkers. These dysfunctions provide more validations for these identified network-based biomarkers of HCC. Moreover, we also compare them with the known HCC-related genes documented in KEGG [21] and MalaCards [22] and further explore their relationships. For a validation purpose, we test the classification performances of network biomarkers in many other independent HCC datasets. Additionally, we also collect the known HCC dysregulated gene sets from MSigDB [13] and compare their classification ability of distinguishing HCC samples from controls with our identified HCC biomarkers. These results further delineate the efficacy and superiority of our proposed method.

The novelty and advantage of our paper can be summarized as follows: A new ensemble module prioritization method called mRank has been proposed to discover module biomarkers on GRN by integrating interactome and transcriptome. mRank firstly conducts an iterative supervised module detection technique guided by phenotypic states to extract the gene community substructure from the global network, then it implements a block-based module ranking technique to prioritize these detected modules by introducing a hypergraph topological centrality. The effectiveness and advantage of mRank have been demonstrated and justified in detecting HCC biomarkers with comparisons to the other alternatives and methods.

2. Materials and methods

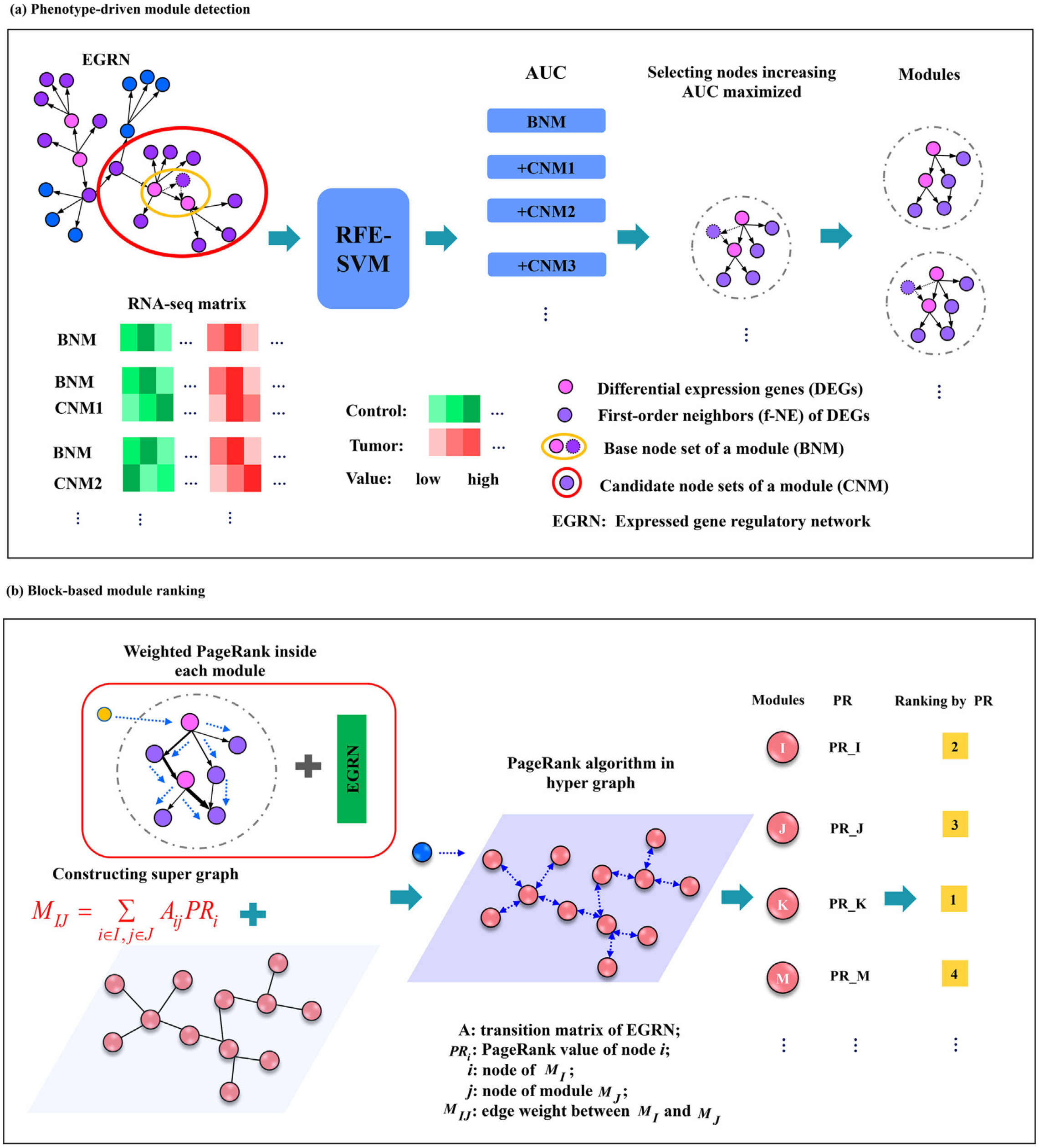
2.1. Datasets

For a-proof-of concept study, we perform mRank for prioritizing network biomarkers of HCC. We compile an integrative GRN from RegNetwork [23], which contains gene regulations documented in more than 20 databases. In total, it contains 169,039 regulations between 18,335 genes. For specifying these gene regulations, we use the context-specific gene expression profiling data form TCGA HCC dataset [9], including 371 tumor samples and 50 control samples. We map the gene expression data to the GRN by differential mutual information (*DMI*), which results in a weighted GRN with 12,343 nodes and 120,069 edges. For justifying our identified biomarkers, we validate the classification performances in other independent datasets, such as HCC gene expression data from NCBI GEO database [24], with the entry IDs GSE14520 [25] with 225 tumor and 220 control samples, GSE25097 [26] with 268 tumor and 243 control samples, GSE45436 [27] with 95 tumor and 39 control samples, GSE64041 [28] with 60 tumor and 60 control samples, GSE63898 [29] with 228 tumor and 168 control samples, and GSE22058 [30] with 100 tumor and 97 control samples. Altogether, there are 976 tumor samples and 827 controls in the independent validation datasets. The detailed descriptions about gene expression datasets used in this study are summarized in **Supplementary Table S1.1**.

2.2. Framework

Fig. 1 summarizes the framework of mRank, which consists of two components: phenotype-driven module detection (Fig. 1(**a**)) and block-based module ranking (Fig. 1(**b**)). The detailed description is as follows:

We firstly preprocess HCC RNA-seq data and identify the differential expressed genes (DEGs) between disease and control samples. Then, we employ a comprehensive GRN from RegNetwork


(a) Phenotype-driven module detection

(b) Block-based module ranking
EGRN
AUC
Selecting nodes increasing AUC maximized
Modules
BNM
+CNM1
+CNM2
+CNM3
RFE-SVM
RNA-seq matrix
BNM
BNM
CNM1
BNM
CNM2
Control:
Tumor:
Value: low high
Differential expression genes (DEGs)
First-order neighbors (f-NE) of DEGs
Base node set of a module (BNM)
Candidate node sets of a module (CNM)
EGRN: Expressed gene regulatory network
Weighted PageRank inside each module
EGRN
Constructing super graph
$M_{IJ} = \sum_{i \in I, j \in J} A_{ij} PR_i$
PageRank algorithm in hyper graph
Modules
PR
Ranking by PR
I
PR_I
2
J
PR_J
3
K
PR_K
1
M
PR_M
4
A: transition matrix of EGRN;
PR_i: PageRank value of node i;
i: node of M_I;
j: node of module M_J;
M_{IJ}: edge weight between M_I and M_J

Fig. 1. Framework of mRank. (a) Phenotype-driven module detection. (b) Block-based module ranking.

[23] and weigh the regulations by *DMI* of two terminal genes from each edge for selecting more reliable and disease-specific networks. Starting from a given DEG, we check if there exists other DEGs interacts with the current DEG through a shortest path. If it exists, we choose those nodes included in this shortest path as our original base nodes of this module (OBNs) and those DEGs as our base source nodes (BSNs). Otherwise, we omit the current DEG. Then, we iteratively select the candidate node sets of this module (CNM) to join in the current module, which comprises the first-order neighbors of BSNs. Specifically, we evaluate the different effect of this module in the classification of tumor and normal samples between before and after adding the neighbor gene. The gene that maximizes the classification effect is selected to join in the module. Here, support vector machine algorithm with recursive feature elimination [31] (RFESVM) is employed to fulfill the classification and realize feature selection simultaneously. The area under receiver operating characteristics (ROC) curve (AUC) is used to guide the module amplification, as shown in Fig. 1 **(a)**.

Then, we build up a block-based ranking algorithm to generate the order of these modules by introducing the network topology, which includes ranks in the intra-module and in the inter-module [32]. In the intra-module ranking, the block-rank procedure employs a weighted PageRank (PR) algorithm [33]. In the inter-module ranking, a hypergraph network is constructed, in which each module is represented by a super node. Rank of each module is obtained by applying PR algorithm [34] in the hypergraph (Fig. 1 **(b)**). Finally, these modules are ranked according to their PR values and these top-ranked modules are recognized as the identified biomarkers for further validations.

2.3. Phenotype-driven module detection

We implement data preprocess and DEG analysis by following the guide of TCGA biolinks package [35]. After filtering, 12,892 genes are left and 2,850 genes among them are identified as DEGs.

Then, we map these corresponding molecular profiles of HCC RNA-seq data onto the GRN by calculating *DMI* for each edge as:

$$DMI = |MI(X_c, Y_c) - MI(X_d, Y_d)| \tag{1}$$

where *X* and *Y* stands for the expression vector of the two genes. *DMI* describes the absolute difference between the two mutual information (*MI*) values of control (*c*) and disease (*d*) states. Then, these *DMI* values are used for selecting reliable edges and further are normalized by the min–max method. To calculate *DMI*, a straightforward method proposed in [36] is employed, which estimates *MI* by firstly partitioning the variables into finite-size bins (*N*), then counts the number of points of *X* (or *Y* or *X* and *Y*) falling into these girds, and finally approximates *MI* as:

$$MI(X, Y) = \sum_{i,j} log \frac{p(i,j)}{p_x(i)p_y(j)} \tag{2}$$

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*, $n(i,j)$ is the number of points in their intersection. *N* is dependent on the variable size by drawing grids on a scatterplot of two variables.

Afterwards, we start our phenotype-driven module detection. Table 1 lists the detailed algorithm descriptions of supervised module detection. In this process, when the count of CNMs is great (e.g., greater than 50), RFESVM is implemented to select the most important 50 neighbors for the following module amplification, which calculates model performance by inputting each feature, then ranks all these models by their AUCs and selects the top-k-model corresponding features. In this paper, k is set to 50 based on the number of normal samples in TCGA HCC dataset. In order to solve data imbalance between normal and tumor samples, no-playback sampling technique is adapted.

To describe the relationship between detected modules and phenotypes (disease/normal), we define phenotype association score (*PAS*) to qualify it, which is defined as the *MI* between these gene nodes inside each module and sample phenotype:

$$PAS = \frac{1}{n} \sum_{i \in n} MI(E_i, P) \tag{3}$$

where *n* is the number of genes inside each module, E_i is the expression vector of gene *i*, *P* is the corresponding sign vector where '1' representing tumor state, '0' representing normal state. The higher value the *PAS* is, the more relevant the module is with phenotype.

Table 1
Algorithm of phenotype-driven module detection.

```
Algorithm: Phenotype-driven module detection
1. Input: DEGs, GRN, HCC RNA-seq data
2. Calculation:
   for DEG cur in DEGs
   Search shortest paths between DEG cur and other DEGs
   if exists shortest path
      Get current BNM
      Get BSN
      Get CNMs
      Calculate AUC of current OBN by leave-one out SVM
   if count of CNMs > 50
      Use RFE-SVM to select top-ranked 50 CNMs
   end
   Form candidate module list (CM) list by adding CNMs
   to BNM one by one
   Calculate AUC of members in CNMs by leave-one out SVM one by one
   Select those members with maximum AUC
   Form current module by selecting all nodes in selected CNM
   end
3. Output: Detected modules
```

2.4. Block-based module ranking

After obtaining modules, we rank them from the perspective of hypergraph. In this paper, inspired by exploiting the block structure of the web for computing PR in [32], we propose a block-based module ranking algorithm, which contains weighted PR [33] inside each module (denoted as wPRM) and PR [34] in a hypergraph (denoted as PRH).

In wPRM, we construct an adjacent matrix *B* for each module. According to network topology, if the *i*-th node interacts with *j*-th node inside each module, we set $B_{ij} = 1$. Otherwise, we set $B_{ij} = 0$. We weigh edges inside each module by their *DMI*s and the weighted adjacent matrix *BW* for each module is constructed. Then, the transition matrix *BT* of each module can be obtained:

$$(BT)_{i,j} = \begin{cases} (BW)_{ji} \times B_{i,j} / \sum_j B_{i,j}, & \text{if } \sum_j B_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Let S_0 be the initial probability vector inside each module, and R_i be the vector in which the *i*-th element holds the probability of finding the random walker at node *i* at the current step *t*. The probability vector at *t* + 1 step is given by:

$$R_{t+1} = (1 - \mathrm{d}) \times \mathrm{BT}^{\mathrm{T}} \times \mathrm{R_t} + \mathrm{d} \times \mathrm{S_0} \tag{5}$$

where *d* ($d \in (0, 1)$) is the restart probability, and it represents the chance of random walker going back to the seed nodes.

After iterating some steps, the probability will reach a steady state, which can be obtained by performing the iterations until the difference between R_t and R_{t+1} measured by the L_1 norm falls below a threshold, e.g., 10^{-10}. In this study, we set *d* to be 0.15 by empirical trails. Finally, each node inside each module has its own PR value.

In PRH, we firstly construct its adjacent matrix *M* by using all detected modules and the network topology of GRN, which is defined as:

$$M_{IJ} = \sum_{i \in I, j \in J} A_{ij} P_i \tag{6}$$

where *I*, *J* are two detected modules, *A* is the adjacent matrix of GRN, P_i means the PR value of node *i* inside module *I* getting from wPR part.

Then, we construct the transition matrix *MT* for the hyper graph based on *M*, which is defined as:

$$(MT)_{I,J} = \begin{cases} (MT)_{IJ} / \sum_J M_{IJ}, \; if \; \sum_J M_{IJ} \neq 0 \\ 0, \text{othewise} \end{cases} \tag{7}$$

Supplementary Table S1.2 shows the detailed description of the block-based module ranking algorithm, which contains the construction of adjacent matrix, transition matrix and iterative PR process respectively.

3. Results and discussion

3.1. Top-ranked modules

Our phenotype-driven module detection totally results in 2,562 modules, with 3,073 genes and among them 2,579 are identified as DEGs, with AUC values varying from 0.946 to 0.995 and node size varying from 4 to 16. Here, AUC value is obtained by using SVM classifier with cross-validations on TCGA HCC RNA-seq data at the module detection step. After block-based module ranking, each module obtains its own rank by its PR value. We choose some modules of top-10 ranked modules for declaration. Fig. 2 gives some detailed structures of these identified modules with high

PAS values, i.e., M3, M5, M6, M7, M8 and M10. In Fig. 2, we find that some members in these modules are differentially expressed genes, with higher LogFC (logarithm of fold change) values which mean they are significantly differential expressed. While some genes are not differentially expressed. Besides, we also document cancer genes from GeneCards database [37], some important HCC genes are detected in these modules individually, such as 'CDK6', 'AR', 'ESR1' in M3, 'CDKN2C', 'CDKN2A' in M5, 'JUN', 'CDK4' in M6.

3.2. Classification performance and PAS

To verify the ability of the top-10 ranked modules in distinguishing disease samples from controls, we use SVM classifier to validate those top-ranked modules on TCGA HCC dataset, including 371 tumor samples and 50 controls. Fig. 3**(a)** shows the ROC curves and AUC values of classifications. We find these modules achieve high AUCs with a mean AUC of 0.995 in classifying HCC samples from controls. The detailed classification metrics are listed in Table 2, we find that in our training dataset, those top-ranked modules get high classification AUC values.

To check the correspondence between modules and phenotypes, we calculate the *PAS* values of those top-ranked modules. Fig. 3**(b)** shows the *PAS* values of top-10 ranked modules individually. For comparing with the corresponding baselines, the boxplots of *PAS* in the random-chosen same-size gene sets are also shown. These top-ranked modules have much higher *PAS* values than those randomly chosen gene sets, which indicate their strong relationships with phenotypes. Fig. 3**(c)** gives the regression relationship between AUCs of modules and their corresponding *PAS* values, and we find that *PAS* has a positive relationship with AUC, which gives further evidence for the indication that a high *PAS* means a strong relationship with phenotypic state.

Currently, our proposed method is not built for identifying biomarkers in the classification of different subclasses (e.g., grades, genotypes) of complex disease. Discovering the stratification biomarkers for HCC, i.e., in different stages, types and genotypes, is another invaluable research topic. For testing purpose, we also conduct the classification experiments for our identified top-10-ranked modules in these multi-class classifications. We compile two datasets with entry IDs of GSE6764 [38] and GSE89377 [39] from NIH GEO database [24] (as shown in Supplementary Tables 1.3 and 1.4). Specifically, we test the classification performances of these top-ranked- modules in the two multi-class datasets by using one-against-the-rest technique. We find few of our identified modules even have good classification performance in some kinds of subclass classification. However, most of the modules could not achieve high classification AUC values. The results indicate the identification of diagnostic biomarkers for two classes is much simpler than the identification of stratification biomarkers for HCC. The classification results are shown in Supplementary Tables 1.5 and 1.6.

3.3. Comparing with other methods

Currently, there are no similar methods with mRank for identifying network-based biomarkers by module detecting and ranking simultaneously. Some methods, such as GSEA, GSA, SAFE, CRank and Conductance are important methods for identifying important gene sets or network modules, and they need input predefined gene sets. It is clear that mRank requires no predefined gene sets, which embeds a module detection step. To compare with the existing methods, we firstly use a fast greedy cluster algorithm [40] to separate the GRN into small modules. Then, we input these modules to those aforementioned methods, and rank them based on their corresponding scores. The flowchart of how we compare mRank with other five methods is shown in Supplementary Fig. S1.1.

Fig. 4 shows the comparing results with other five methods on TCGA HCC dataset. We find that our method can detect modules with stronger phenotype relationships than the Modularity method [19] which only considers network topology as shown in Fig. 4**(a)**. In mRank, we build up a phenotype-driven module detection strategy to select nodes with stronger phenotype relationship than their neighbors as shown in Fig. 4**(b)**, which gives an example of module M2. Fig. 4**(c)** presents the mean ROC curves of top-ranked modules by different methods in classifying HCC samples and normal samples. Herein, AUC is achieved by using SVM with cross-validation on TCGA HCC RNA-seq data. We find these top-10 modules by mRank achieve higher mean AUC values than other five methods given with modularity-based modules. Besides, we also find that top-ranked modules obtain higher *PAS* values than those ranked by the other methods (Fig. 4**(d)**). The details about their comparison results are shown in Table 3.

What's more, we also analyze those genes of top-10 ranked modules by different methods, and we find mRank shares some common genes with other methods. Supplementary Fig. S1.2 lists the intersections between top-10 ranked modules from different methods. Furthermore, we validate top-10 ranked modules from different methods by those independent datasets, and Supplementary Fig. S1.3 shows the boxplots of AUCs of top-10 ranked modules identified by different methods in independent validating datasets. The AUC values are obtained by training on the TCGA HCC dataset and testing on different independent validation datasets. It is consistent that those top-10 ranked modules by mRank get higher mean AUC values. Supplementary Table S1.7 shows the classification performances of top-10 ranked modules identified by different methods in the validation datasets.

3.4. Functional enrichment analysis and justification

To identify the enriched functions of those detected modules, we implement network ontology analysis (NOA) for these top-10 ranked modules [41]. We find that these GO functions, such as cell death, cell proliferation, cell differentiation, signal transduction are enriched. The full list is shown in Supplementary Table S1.8.

Due to the important implications of GO functions listed above in cancer mechanism [42,43], we select 69 genes annotating the GO terms of cell communication (GO:0007154), cell death (GO:0008219), cell proliferation (GO:0008283), cell differentiation (GO:0030154), signal transduction (GO:0007165), cell surface receptor signaling pathway (GO:0007166), apoptotic process (GO:0006915), and programmed cell death (GO:0012501) as our final identified candidate HCC biomarker genes (shown in Supplementary Table S1.8). Fig. 5**(a)** gives the GO chard plots of these identified candidate biomarkers, and Fig. 5**(b)** shows the network topology of those selected candidate biomarker genes and the modules in which they are contained. The enriched functions of GO analysis on biomarker genes are listed in Supplementary Table S1.9, the pathways of cell cycle, hepatocellular carcinoma, p53 signaling pathway, and hepatitis B are significantly enriched in these identified biomarkers.

Then, we carefully search the identified HCC biomarkers in literatures and obtain the existing HCC biomarkers, as listed in Supplementary Table S1.10. We find that an important HCC biomarker is AFP, which is proved by Harry Abelev and his colleagues in 1962 by using blood of mice with experimental liver tumor [44]. And some biomarkers have also been found by similar time-consuming and expensive experiments. Currently, more and more high-throughput data such as transcriptome and interactome become available. It is an opportunity and also a challenge to discover new diagnostic HCC biomarkers from these resources. This is

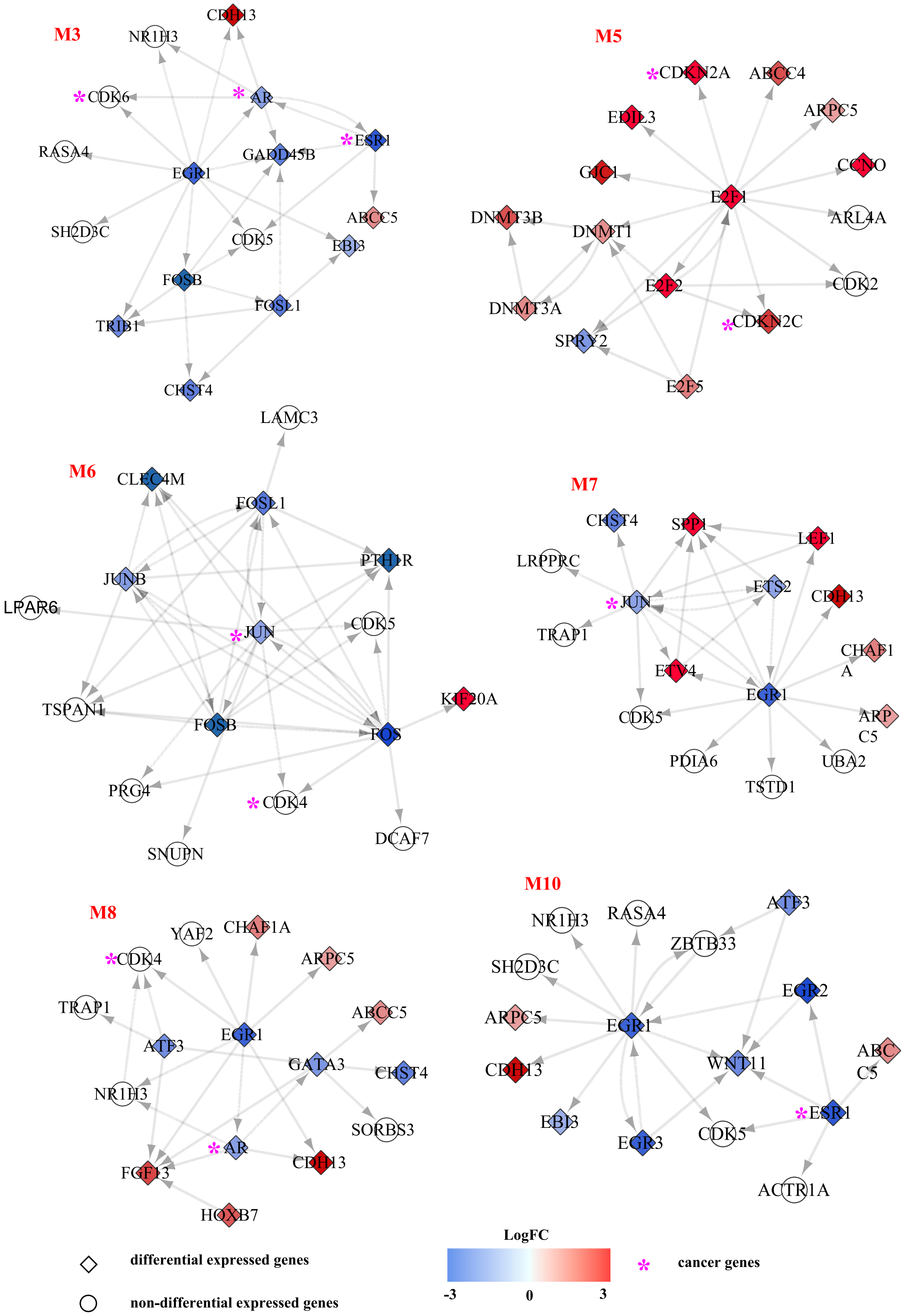


Fig. 2. Detailed structures of six modules with higher *PAS* values among top-10 ranked modules (M3, M5, M6, M7, M8, and M10). *P*-values are calculated based on tumor and normal samples from TCGA HCC dataset.

our motivation of proposing such a computational method mRank to discover biomarkers from omics data. Interestingly, we find some important HCC biomarkers such as AFP and SPP1 are really included in our identified HCC biomarkers. The overlaps with the reported HCC biomarkers provide more evidence for the effectiveness of our proposed method.

We further search HCC disease genes documented in KEGG [21] and MalaCards [22]. After double checking these genes individually, we obtain a list of 37 reliable HCC disease genes. The 37 HCC genes and their gene symbols are shown in Supplementary Table S1.11. There is only one overlapping gene between these genes and our identified HCC biomarkers, i.e., CDKN2A. Afterwards,

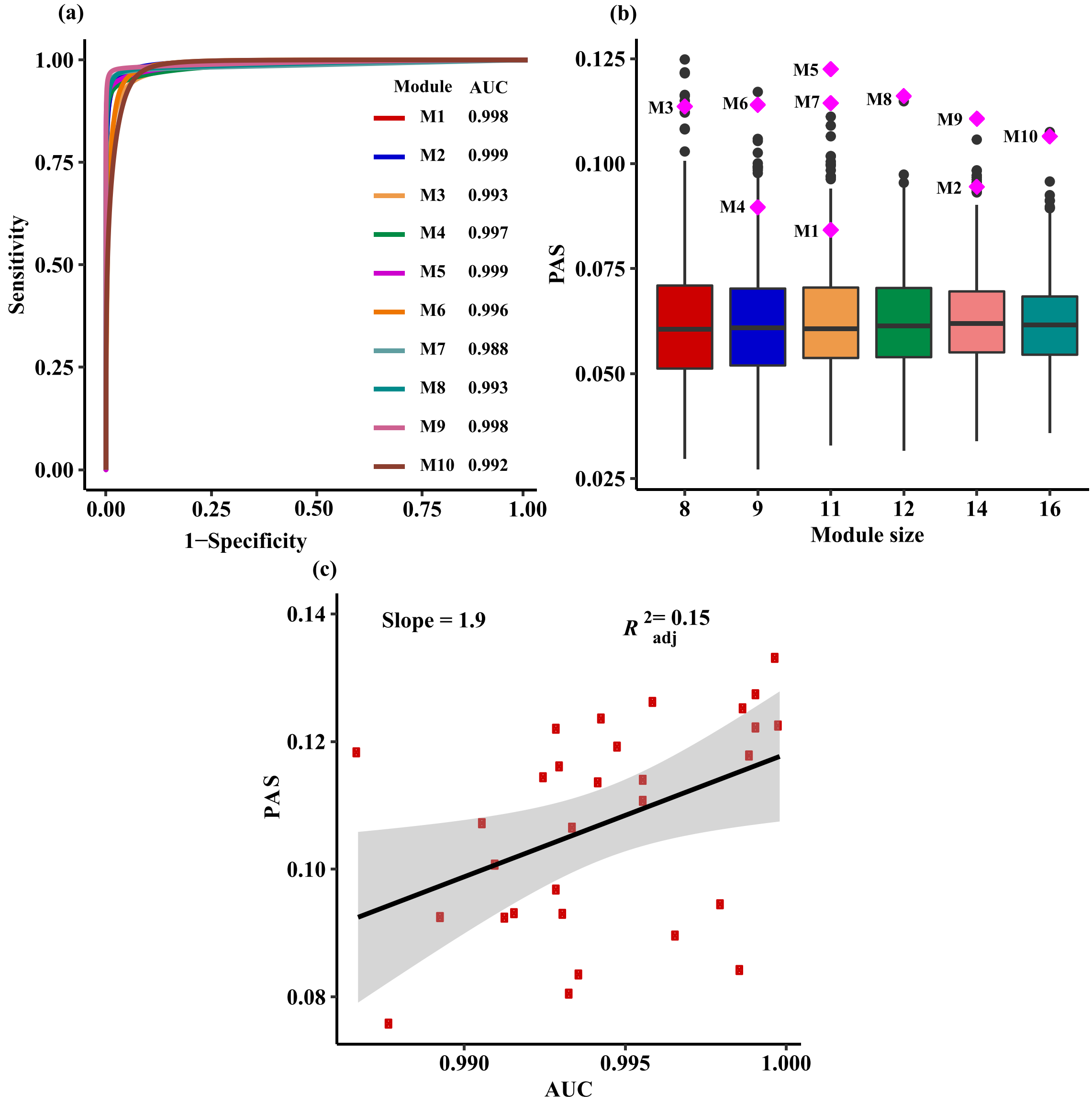


Fig. 3. Classification performance, *PAS* and the relationship between AUC and *PAS* on TCGA HCC dataset. (a) The ROC curves of top-10 ranked modules. (b) The *PAS* values of top-10 ranked modules and the boxplots of *PAS* values of randomly-chosen same-size gene sets. (c) Regression between *PAS* and AUC in these top-ranked modules. The grey area means 95% confidence interval.

Table 2
Classification results of top-10 ranked modules in TCGA HCC dataset.

Module	SE	SP	ACC	F1	AUC
M1	0.980	0.980	0.980	0.980	0.998
M2	0.980	1.000	0.990	0.989	0.998
M3	0.960	0.980	0.970	0.969	0.993
M4	0.980	0.960	0.970	0.970	0.996
M5	0.980	1.000	0.990	0.989	0.999
M6	1.000	0.960	0.980	0.980	0.995
M7	0.980	1.000	0.990	0.980	0.988
M8	0.980	1.000	0.990	0.989	0.993
M9	1.000	0.960	0.980	0.980	0.998
M10	0.980	0.960	0.970	0.970	0.992
Mean ± SD	0.982 ± 0.011	0.980 ± 0.019	0.981 ± 0.009	0.979 ± 0.008	0.995 ± 0.003

we want to check if there exists some undirected relationship between the two gene sets. Thus, we build up a GRN of these genes from the prior human gene regulatory network collected in RegNetwork [23]. For integrity, we also extract their neighbor genes. Interestingly, we find that 45 genes in our identified HCC biomarkers, such as AFP, AR, JUN, are regulated by some of the 37 known HCC genes, such as STAT3, TP53, CTNNB and RB1. Supplementary Fig. S1.4 demonstrates the regulatory relationships between the known HCC gene and our identified biomarker genes. The results demonstrate that our identified biomarker genes are

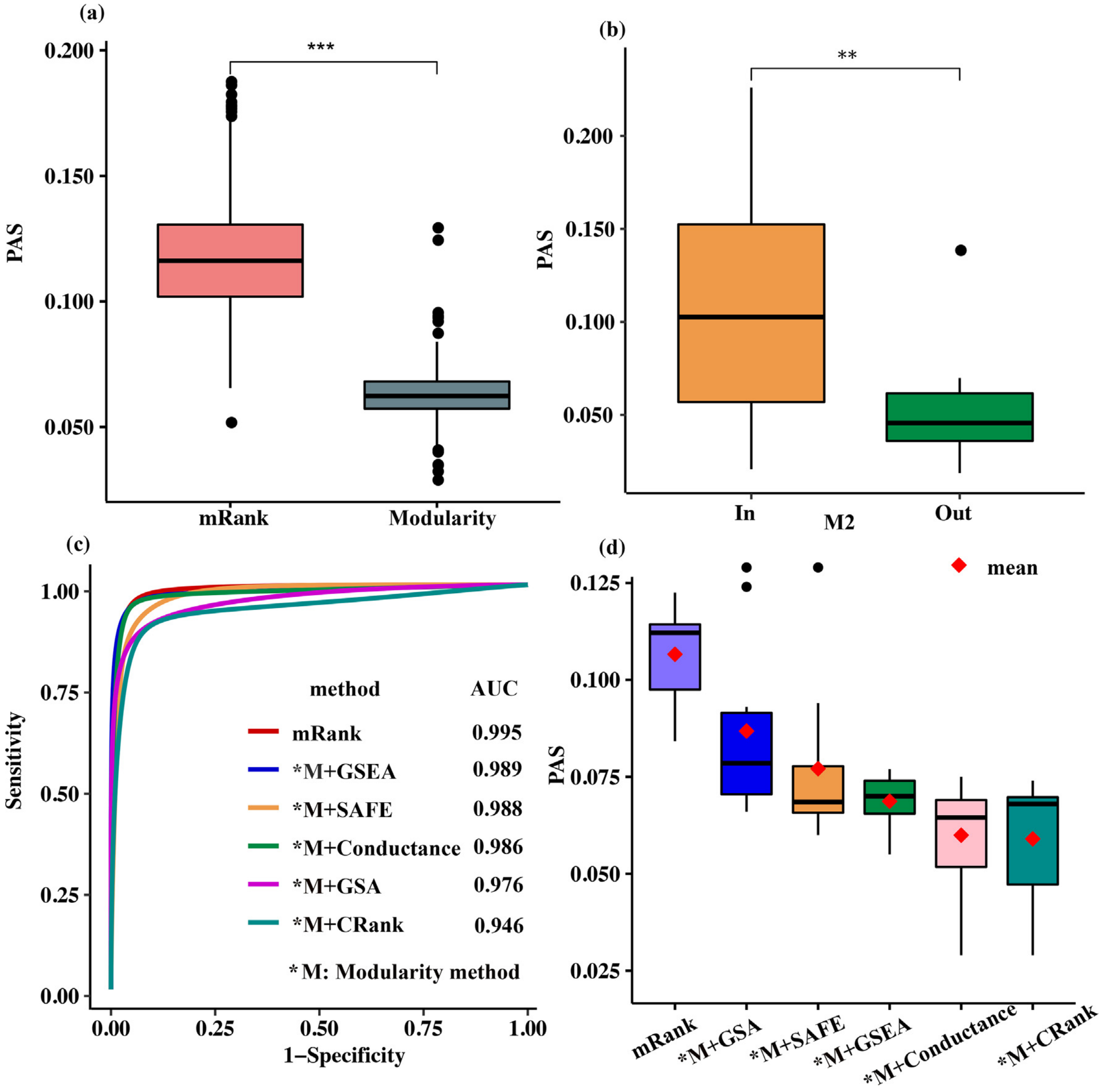


Fig. 4. Comparison results with other methods based TCGA HCC dataset. (a) Boxplots of *PAS* values between mRank-detected and Modularity-detected modules. (b) Boxplots of *PAS* values between nodes inside a module and outside a module (M2 of mRank). (c) Mean ROC curves of top-10 ranked modules ranked by different methods. (d) Boxplots of *PAS* values of top-10 ranked modules by different methods. In (b) and (c), *** means P value < 1e − 10, ** means P value < 1e − 2, by two-sided Wilcoxon test.

Table 3
Comparisons of top-10 ranked modules with other methods in TCGA HCC dataset (Mean ± SD). *M refers to input the modules identified by the modularity-based method.

Methods	AUC	*PAS*
mRank	0.995 ± 0.003	0.117 ± 0.020
*M + GSEA	0.989 ± 0.006	0.069 ± 0.007
*M + SAFE	0.988 ± 0.006	0.077 ± 0.021
*M + Conductance	0.985 ± 0.012	0.060 ± 0.012
*M + GSA	0.976 ± 0.007	0.087 ± 0.023
*M + CRank	0.946 ± 0.045	0.059 ± 0.012

closely related with the known HCC disease genes although they are not these genes.

In addition, we also check our identified HCC biomarkers with the known HCC biomarkers documented in PubMed. Supplementary Table S1.12 lists the searching results of our identified 69 HCC biomarkers. 58 genes in the 69 biomarker genes have been studied in HCC. The 11 new biomarker genes are ACBB4, ARL4A, CCNO, CD3EAP, GJC1, MRGPRF, NES, PRRT2, PTH1R, RASA4 and SH2D3C. The overlaps imply that most of identified biomarkers have been found with their relationships with HCC. In this paper, we identify them simultaneously from a gene regulatory network by proposing a module detection and ranking method.

Here, we mainly aim to propose a computational module-based biomarker discovery method. For demonstrating the effective of our proposed pipeline, we perform our study on detecting HCC biomarkers. Our identified candidate biomarker genes can be used for targeting genuine biomarkers of HCC used in clinics. The next steps include many laborious wet experiments and clinical trials, such the experiments in cell lines, animal experiments and further multi-stage clinical trials. Due to that mRank is a fully computational biotechnology, the data-driven candidate biomarker genes need to combine with prior knowledge for the following step-by-step experiments in the HCC biomarker development.

3.5. Validations in external independent datasets

For justification, we verify our identified 69 candidate HCC biomarkers in the other independent datasets, i.e., GSE14520,

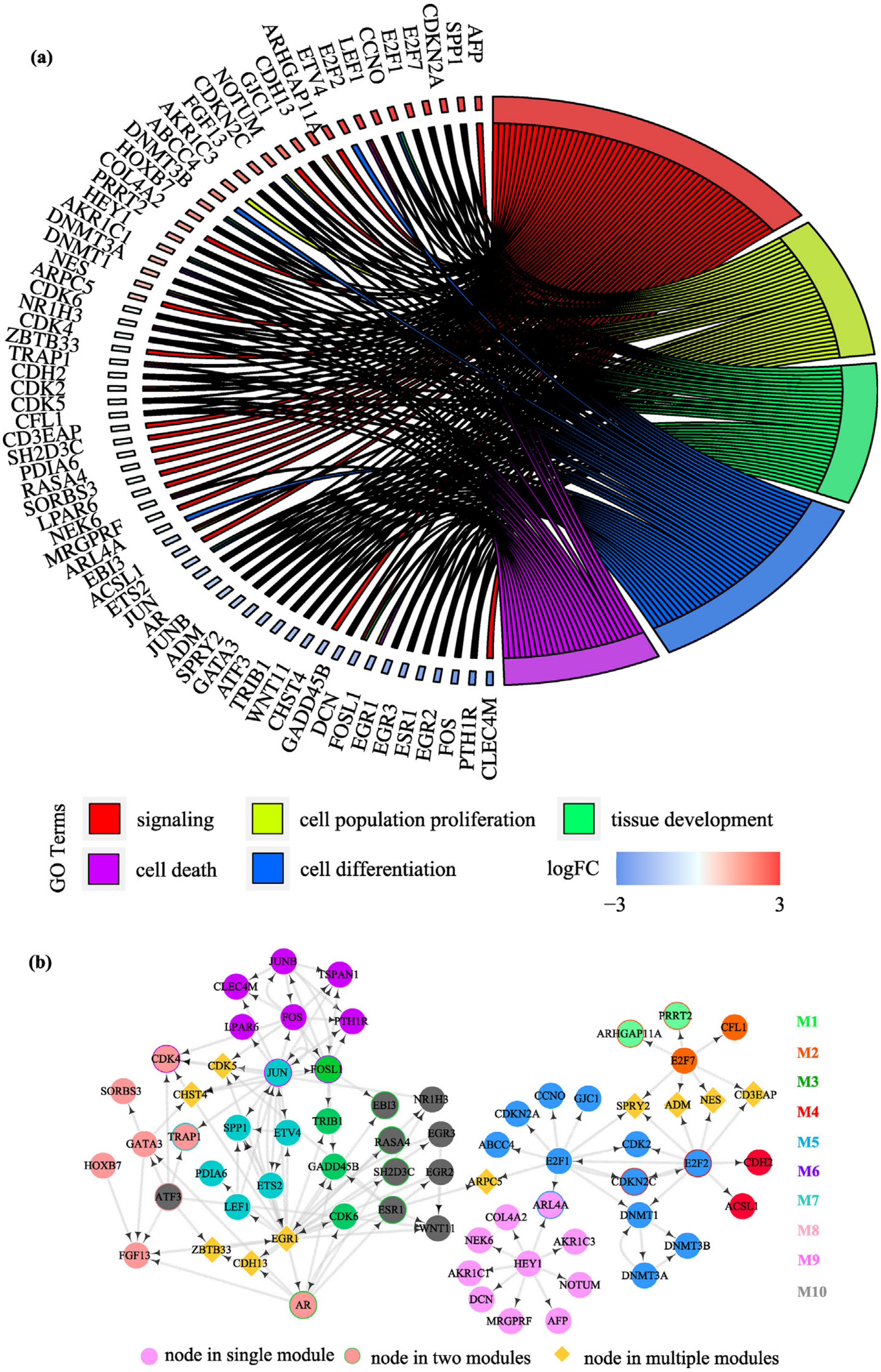


Fig. 5. The identified HCC biomarkers. (a) GO chard plot of identified 69 HCC biomarker genes. LogFC are calculated based on TCGA HCC dataset. (b) Network structure of identified 69 HCC biomarkers. Different colors indicate different modules (M1, M2, . . . , M10). Node filled with single color belongs to one module, node filled with single color and with edge filled with another color belongs to two different modules, nodes in yellow diamond belong to more than two modules. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

GSE25097, GSE45436, GSE63898, GSE22058 and GSE64041. We use TCGA HCC dataset as the training set and the other independent datasets individually as the testing sets. Fig. 6**(a)** shows the ROC curves of identified 69 HCC biomarkers tested in the six validation datasets respectively. The details of classification results can be found in Table 4. From Fig. 6**(a)**, we find our biomarkers get high AUC values in some datasets, i.e., GSE22058 with AUC of 0.935, GSE45436 with AUC of 0.928, GSE25097 with AUC of 0.887, GSE14520 with AUC of 0.875, GSE64041 with AUC of 0.745, GSE63898 with AUC of 0.705. Altogether, our identified biomarkers achieve a mean AUC value of 0.846 with low standard deviation of 0.097 in 976 tumor samples and 827 controls. Fig. 6**(b)** shows the numbers of biomarker genes which are simultaneously contained in these independent validation datasets.

Besides, we also compare our identified HCC biomarkers with known HCC biomarkers and dysregulated gene sets. We firstly test the classification results of these reported HCC biomarkers, e.g., AFP, DCP, GPC3, SPP1, CD44, and DKK1, by using TCGA HCC dataset

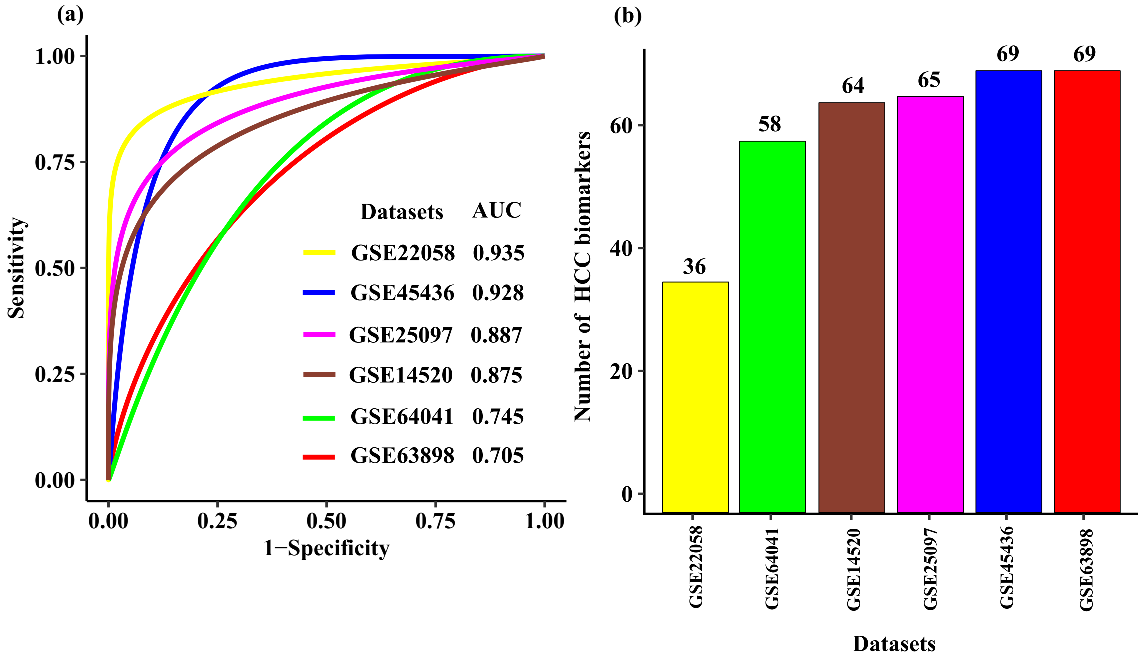


Fig. 6. Validations of identified 69 HCC biomarkers in independent datasets. (a) ROC curves of classification in numerous independent datasets. (b) Numbers of HCC biomarkers contained in the independent validation datasets.

Table 4
Classification results of identified 69 HCC biomarkers in the independent datasets.

Datasets	SE	SP	ACC	F1	AUC
GSE22058	0.980	0.887	0.933	0.936	0.935
GSE45436	0.768	0.974	0.871	0.857	0.923
GSE25097	0.787	0.864	0.826	0.819	0.887
GSE14520	0.827	0.809	0.818	0.820	0.875
GSE64041	0.850	0.583	0.717	0.750	0.745
GSE63898	0.658	0.673	0.665	0.663	0.705
Mean ± SD	0.812 ± 0.106	0.798 ± 0.145	0.805 ± 0.099	0.807 ± 0.093	0.846 ± 0.097

Table 5
Classification comparisons between our identified 69 candidate HCC biomarkers and known HCC biomarkers as well as 8 dysregulated gene sets from MSigDB in the independent validation datasets.

Type	SE	SP	ACC	F1	AUC
Ours	0.812 ± 0.106	0.798 ± 0.145	0.805 ± 0.099	0.807 ± 0.093	0.846 ± 0.097
Known	0.569 ± 0.148	0.825 ± 0.142	0.697 ± 0.122	0.650 ± 0.136	0.666 ± 0.137
Gene list 3	0.717 ± 0.197	0.842 ± 0.111	0.779 ± 0.108	0.754 ± 0.143	0.805 ± 0.114
Gene list 1	0.563 ± 0.145	0.885 ± 0.047	0.724 ± 0.087	0.665 ± 0.123	0.715 ± 0.123
Gene list 5	0.856 ± 0.213	0.344 ± 0.399	0.600 ± 0.110	0.679 ± 0.057	0.609 ± 0.132
Gene list 6	0.447 ± 0.115	0.774 ± 0.137	0.610 ± 0.048	0.527 ± 0.083	0.603 ± 0.093
Gene list 8	0.897 ± 0.253	0.150 ± 0.367	0.524 ± 0.058	0.642 ± 0.062	0.533 ± 0.082
Gene list 7	0.993 ± 0.016	0.053 ± 0.132	0.523 ± 0.057	0.677 ± 0.024	0.522 ± 0.053
Gene list 2	1.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	0.667 ± 0.000	0.500 ± 0.000
Gene list 4	1.000 ± 0.000	0.000 ± 0.000	0.500 ± 0.000	0.667 ± 0.000	0.500 ± 0.000

for training and other HCC datasets from GEO database for testing, the results are as shown in Table 5 (See further details in Supplementary Table 1.13). From Table 5, we find that our identified 69 candidate HCC biomarkers achieve better classification performance than those known HCC biomarkers.

Then, we document known HCC dysregulated gene sets from MSigDB [13], and 8 HCC dysregulated gene sets are compiled (See details in Supplementary Table S1.14). Similarly, we use TCGA HCC dataset for training and the other datasets for testing datasets, the results are as shown in Table 5. (See further details in Supplementary Tables S1.15–S1.22). From Table 5, we find that our identified 69 candidate HCC biomarkers obtain better results than those 8 dysregulated HCC gene sets, which further illustrates the advantage of mRank.

Additionally, we find some of the known HCC biomarkers, e.g., AFP, FGF13, SPP1 and CDKN2A, are included in our discovered biomarkers, so we use them to construct a reliable biomarker gene set. And JUN is regulated by STAT3, MYC, RB1, TCF7, and NFE2L2, so we also include it to the reliable biomarker gene set. Then, we clarify their classification results in several HCC datasets from GEO database by using TCGA HCC data for training. As shown in Table 6, we find that they perform better classification results in four

Table 6
Classification results of 5 more reliable HCC biomarkers in the independent datasets.

Datasets	#of genes	SP	SE	ACC	F1	AUC
GSE25097	5	0.836	0.930	0.883	0.877	0.945
GSE14520	5	0.813	0.955	0.884	0.875	0.927
GSE45436	5	0.832	0.923	0.877	0.871	0.926
GSE64041	5	0.798	0.875	0.837	0.830	0.907
GSE63898	5	0.733	0.767	0.750	0.746	0.769
GSE22058	5	0.620	0.464	0.542	0.575	0.476
Mean ± SD	5	0.772 ± 0.083	0.819 ± 0.186	0.795 ± 0.134	0.796 ± 0.119	0.808 ± 0.199

datasets, such as they reach the AUC of 0.945 in GSE25097, 0.927 in GSE14520, 0.926 in GSE45436, and 0.907 in GSE64041, but in two datasets GSE63898 and GSE22058, they obtain the AUC of 0.769 in GSE63898 and even worse in GSE22058 with AUC of 0.476.

Moreover, we also test if our selected biomarker genes are significantly differential expressed in all these datasets. After calculating significance between normal and tumor samples of each dataset, we find that not all biomarker genes are significant in those datasets. Some genes are significant in TCGA HCC dataset, while they are not significant in the other validation datasets. The details about the significance differences of biomarker genes among all datasets are available in Supplementary Table S1.23.

Furthermore, we integrate HCC data from CCLE database [45] with 25 liver cancer samples and GTEx database [46] with 177 normal samples. We compare normal and liver cancer samples of the two datasets and get the significance of each gene. We find 55 genes of our selected 69 biomarkers are significantly differentially expressed with *P*-value < 0.05 (as shown in Supplementary Fig. S1.5 and S1.6). Then, we use hypergeometric test to check if our selection is significant. The *P*-value is 6.69e^{-3}, which demonstrates the significance of our identified biomarkers in the independent experiments. Supplementary Fig. S1.7 shows the heatmap of Pearson's correlation coefficient (PCC) of identified 69 HCC biomarkers from dataset GTEx and CCLE dataset. It is obvious that the correlations among these genes are totally different between samples in normal (GTEx) and disease (CCLE) datasets, which further indicates the differential information of our identified 69 HCC biomarkers in disease and control samples.

4. Conclusions

In this paper, a phenotype-driven module detection and block-based module ranking method, called mRank, has been proposed for discovering cancer biomarkers from transcriptome and interactome. In module detection, we considered the phenotype information and network topology simultaneously, which can provide guidance for searching more effective modules in classification. In module ranking, we proposed a ranking method based on BlockRank algorithm to gain prioritization value from both the intra-module and inter-module, which can provide more comprehensive network topology and cross-talking information for precision ranking. Compared to randomly-chosen gene sets, the *PAS* values of top-ranked modules illustrate their stronger relationships with phenotypes. The comparisons with the other existing gene-set-based methods with additional module inputs also delineate the advantages of our proposed mRank. The enriched functions of top-ranked modules and relations between our identified candidate HCC biomarkers with known HCC related genes indicate the effectiveness of mRank, which can provide more evidence for the efficacy of identified HCC biomarkers. Furthermore, the validations of selected biomarkers and comparisons with other known HCC biomarkers/dysregulated gene sets in the other independent datasets imply the general effectiveness of our findings.

In current version of mRank, we have not directly considered the mutations in the biomarker identification. We weighed the context-specific gene regulatory network only by RNA-seq data. However in the gene regulatory network, we included the known mutated genes documented in the disease gene database Malacards. They are detected mainly by genome-wide association study (GWAS), and most of the mutations are used to identify HCC disease genes, as well as copy number variants (CNV) and DNA methylation data. In the module detection procedure of mRank, we set up the differentially expressed genes according to RNA-seq data as the source nodes. We found some of them are the known HCC disease genes. Thus, our method has partially included the mutation information in the proposed biomarker discovery strategy.

For fully consideration of mutation in the biomarker discovery, we think we need to modify the weight strategy beyond the RNA-seq gene expression profiles. When there are significant mutations, we can treat these mutated genes as the source nodes in mRank. It is also expected to construct a more reasonable weighted network in prioritizing module biomarkers by integrating more available information for HCC, such as mutation, DNA methylation, and other important genetic and epigenetic cancer-causing measurements in biomarker discovery.

Funding

This work was partially supported by National Key Research and Development Program of China (No. 2020YFA0712402); National Natural Science Foundation of China 61973190 and 61572287; Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project 2019JZZY010423); Natural Science Foundation of Shandong Province of China (ZR2020ZD25); the Innovation Method Fund of China (Ministry of Science and Technology of China, 2018IM020200); Tang Scholar and the Program of Qilu Young Scholar of Shandong University.

Data availability

All data and source code used in this study are available at: https://github.com/zpliulab/mRank.

CRediT authorship contribution statement

HS: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft. **ZL:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administraion, Resources, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Thanks are due to anonymous reviewers for their constructive suggestions and comments. We also thank the 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.csbj.2021.12.005.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA: A Cancer J Clinic 2019;69(1):7–34.
[2] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA: A Cancer J Clin 2020;70(1):7–30.
[3] Lichtenstein P 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 2000;343(2):78–85.
[4] Liu Z-P. Identifying network-based biomarkers of complex diseases from high-throughput data. Biomarkers Med 2016;10(6):633–50.
[5] Piñero F, Dirchwolf M, Pessôa MG. Biomarkers in Hepatocellular Carcinoma: Diagnosis, Prognosis and Treatment Response Assessment. Cells 2020;9(6):1370.
[6] Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med 2016;375(8):717–29.
[7] van Lint FHM, Mook ORF, Alders M, Bikker H, Lekanne dit Deprez RH, Christiaans I. Large next-generation sequencing gene panels in genetic heart disease: yield of pathogenic variants and variants of unknown significance. Netherl Heart J 2019;27(6):304–9.
[8] Gao Q, Zhu H, Dong L, Shi W, Chen R, Song Z, et al. Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. Cell 2019;179(2):561–577.e22.
[9] Ally A et al. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell 2017;169(7):1327–1341.e23.
[10] The Cancer Genome Atlas Research. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20.
[11] The International Cancer Genome C et al. International network of cancer genome projects. Nature 2010;464(7291):993–8.
[12] Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5(2):101–13.
[13] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A . 2005;102(43):15545–50.
[14] Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat 2007;1(1):107–29.
[15] Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005;21(9):1943–9.
[16] Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-S, et al. A novel signaling pathway impact analysis. Bioinformatics 2009;25(1):75–82.
[17] Zitnik M, Sosič R, Leskovec J. Prioritizing network communities. Nat Commun 2018;9(1). https://doi.org/10.1038/s41467-018-04948-5.
[18] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network Motifs: Simple Building Blocks of Complex Networks. Science 2002;298(5594):824–7.
[19] Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A . 2006;103(23):8577–82.
[20] Schaeffer SE. Graph clustering. Comput Sci Rev 2007;1(1):27–64.
[21] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45(D1):D353–61.
[22] Rappaport N et al. MalaCards: an integrated compendium for diseases and their annotation. Database(Oxford) 2013;2013:bat018.
[23] Liu Z-P, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database 2015;2015:bav095. https://doi.org/10.1093/database/bav095.
[24] Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 2006;411:352–69.
[25] Roessler S, Jia H-L, Budhu A, Forgues M, Ye Q-H, Lee J-S, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Res 2010;70(24):10202–12.
[26] Sung W-K et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet 2012;44(7):765–9.
[27] Wang H-W et al. Forfeited hepatogenesis program and increased embryonic stem cell traits in young hepatocellular carcinoma (HCC) comparing to elderly HCC. BMC Genomics 2013;14:736.
[28] Makowska Z et al. Gene expression analysis of biopsy samples reveals critical limitations of transcriptome-based molecular classifications of hepatocellular carcinoma. J Pathol Clin Res 2016;2(2):80–92.
[29] Villanueva A, Portela A, Sayols S, Battiston C, Hoshida Y, Méndez-González J, et al. DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma. Hepatology 2015;61(6):1945–56.
[30] Burchard J et al. microRNA-122 as a regulator of mitochondrial metabolic gene network in hepatocellular carcinoma. Mol Syst Biol 2010;6:402.
[31] Guyon I, Weston J, Barnhill S, et al. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 2002;46:389–422. https://doi.org/10.1023/A:1012487302797.
[32] Kamvar SD et al. Exploiting the block structure of the web for computing PageRank. Tech Rep 2003.
[33] Xing W, Ghorbani A. Weighted PageRank algorithm. Proceedings. Second Annual Conference on Communication Networks and Services Research 2004;2004:305–14.
[34] Page L, Brin S. The PageRank citation ranking: bringing order to the web. Tech Rep 1999.
[35] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucl Acids Res 2016;44(8):e71.
[36] Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics 2008;9:461.
[37] Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database(Oxford) 2010;2010:baq020.
[38] Wurmbach E, Chen Y-B, Khitrov G, Zhang W, Roayaie S, Schwartz M, et al. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. Hepatology 2007;45(4):938–47.
[39] Shen Q, Eun JW, Lee K, Kim HS, Yang HD, Kim SY, et al. Barrier to autointegration factor 1, procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3, and splicing factor 3b subunit 4 as early-stage cancer decision markers and drivers of hepatocellular carcinoma. Hepatology 2018;67(4):1360–77.
[40] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E 2004;70(6):066111. https://doi.org/10.1103/PhysRevE.70.066111.
[41] Wang J, Huang Q, Liu Z-P, Wang Y, Wu L-Y, Chen L, et al. NOA: a novel network ontology analysis method. Nucleic Acids Res 2011;39(13):e87.
[42] Hanahan D, Weinberg R. Hallmarks of cancer: the next generation. Cell 2011;144(5):646–74.
[43] Hanahan D, Weinberg RA. The Hallmarks of cancer. Cell 2000;100(1):57–70.
[44] Fau AG, Perova Sd, et al. Production of embryonal alpha-globulin by transplantable mouse hepatomas. Transplantation 1963;1:174–80.
[45] Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature 2019;569(7757):503–8.
[46] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45(6):580–5.