PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2024,51(7):1711~1726

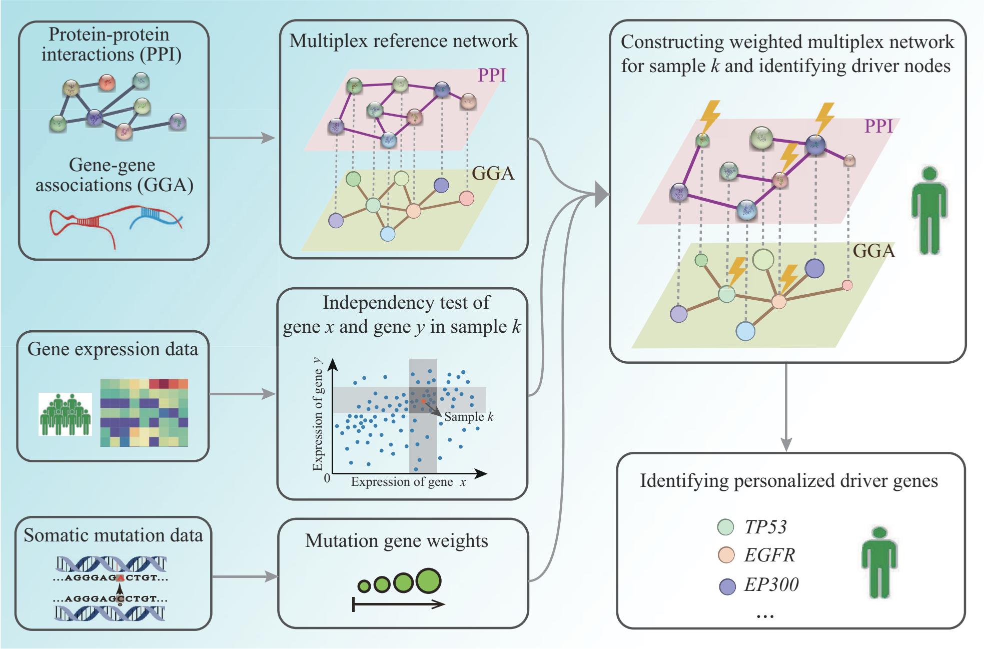
www.pibb.ac.cn

A Multiplex Network Control Method for Identifying Personalized Cancer Driver Genes*

ZHANG Tong[1,2], ZHANG Shao-Wu[1]**, LI Yan[1], XIE Ming-Yu[1]

([1]*Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation,*
Northwestern Polytechnical University, Xi'an 710072, China;
[2]*School of Electrical and Mechanical Engineering, Pingdingshan University, Pingdingshan* 467000, China)

Graphical abstract


Protein-protein interactions (PPI)
Gene-gene associations (GGA)
Multiplex reference network
PPI
GGA
Constructing weighted multiplex network for sample k and identifying driver nodes
PPI
GGA
Gene expression data
Independency test of gene x and gene y in sample k
Expression of gene y
Sample k
0
Expression of gene x
Somatic mutation data
...AGGGAGCCTGT...
...AGGGAGCCTGT...
Mutation gene weights
Identifying personalized driver genes
TP53
EGFR
EP300
...

Abstract **Objective** Inferring cancer driver genes, especially rare or sample-specific cancer driver genes, is crucial for precision oncology. Considering the high inter-tumor heterogeneity, a few recent methods attempt to reveal cancer driver genes at the individual level. However, most of these methods generally integrate multi-omics data into a single biomolecular network (*e.g.*, gene regulatory network or protein-protein interaction network) to identify cancer driver genes, which results in missing important interactions highlighted in different networks. Thus, the development of a multiplex network method is imperative in order to integrate the interactions of different biomolecular networks and facilitate the identification of cancer driver genes. **Methods** A multiplex network control method called Personalized cancer Driver Genes with Multiplex biomolecular Networks (PDGMN) was proposed. Firstly, the sample-specific multiplex network, which contains protein-protein interaction layer and gene-gene association layer, was constructed based on gene expression data. Subsequently, somatic mutation data was integrated to weight the nodes in the

* This work was supported by a grant from The National Natural Science Foundation of China (62173271).

** Corresponding author.

Tel: 86-29-88431308, E-mail: zhangsw@nwpu.edu.cn

Received: October 10, 2023 Accepted: March 6, 2024

sample-specific multiplex network. Finally, a weighted minimum vertex cover set identification algorithm was designed to find the optimal set of driver nodes, facilitating the identification of personalized cancer driver genes. **Results** The results derived from three TCGA cancer datasets indicate that PDGMN outperforms other existing methods in identifying personalized cancer driver genes, and it can effectively identify the rare driver genes in individual patients. Particularly, the experimental results indicate that PDGMN can capture the unique characteristics of different biomolecular networks to improve cancer driver gene identification. **Conclusion** PDGMN can effectively identify personalized cancer driver genes and broaden our understanding of cancer driver gene identification from a multiplex network perspective. The source code and datasets used in this work are available at https://github.com/NWPU-903PR/PDGMN.

Key words multiplex biomolecular networks, multiplex network control, personalized cancer driver genes, sample-specific multiplex network, minimum vertex cover set

DOI：10.16476/j.pibb.2023.0392

Cancer is characterized by the acquisition of genetic mutations, some of which, termed as driver mutations, confer a selective growth advantage to the mutated cells and lead to abnormal and uncontrolled cellular growth[1-2]. Identifying the cancer driver genes that harbor driver mutations has been one of the main goals in cancer research since the establishment of cancer genomics[3-4].

In recent years, considerable efforts have been dedicated to identify the cancer driver genes within a cohort of patients[5]. For example, the mutation frequency-based methods commonly identify the genes that harbor higher statistical significance than the background mutation rate as cancer driver genes[6-8]. The network-based methods incorporate knowledge of pathways, protein-protein or gene-gene interactions to elucidate tumor initiation and progression for identifying the cancer driver genes[9-12]. These methods mainly focus on identifying the cancer driver genes in specific populations of different types or subtypes of cancer. Although the cohort-level methods have successfully identified well-established and potential cancer driver genes[13-14], they are not well-suited for detecting rare or patient-specific driver genes that occur in small cohorts or even single patients due to low statistical power. Moreover, the high heterogeneity of cancer may result in different driver genes in individual patients[15-16]. Hence, the identification of personalized driver genes (PDGs) plays a crucial role in the development of precision oncology and cancer therapeutics.

Recently, some methods have been proposed to identify cancer driver genes at the individual resolution, *i.e.*, identifying the PDGs in cancers. On one hand, a few methods use machine learning algorithms to identify PDGs. For example, sysSVM[17] and sysSVM2[18] utilized the discriminate features of both somatic alterations and gene properties of well-established driver genes and adopted the support vector machine (SVM) to predict the driver genes that best resemble these features for individual patients. IMCDriver[19] introduced inductive matrix completion to identify the mutated genes that are most functionally similar to the well-established driver genes as PDGs in individual patients. On the other hand, some network-based methods attempt to identify PDGs by elucidating the tumor initiation and progression at the network- or system-level. These methods commonly integrate multi-omics data into a single biomolecular network, such as gene regulatory network or protein-protein interaction (PPI) network, to identify PDGs. For example, some of these methods such as DawnRank[20], OncoIMPACT[21], PRODIGY[22], and PersonaDrive[23] integrated the personalized transcriptomic and genomic data to measure deregulations in a biomolecular network for prioritizing or identifying PDGs. Other methods such as SCS[24], PNC[25] and pDriver[26] adopted the network structural controllability theory[27-28] to identify coding and non-coding driver genes for individual patients. Generally, these methods combined the gene expression and somatic mutation data with a gene regulatory network or PPI network to construct the sample-specific network and then identified driver nodes from the sample-specific network by network control methods. From the perspective of system control, driver nodes are considered as key components in the sample-specific network, which can significantly influence the overall state of the network system, such as driving the state of patient-specific gene networks from a normal attractor to a disease attractor. Thus, in the context of cancer biology, driver nodes often correspond to genes that play curial roles in cancer development and progression[29], providing us a valuable resource to identify personalized cancer driver genes.

Although the above methods can be used to identify rare or patient-specific driver genes and further facilitate the combinatorial drug discovery for individual patients[30], the common limitation of these methods is that they are designed for investigating only one single biomolecular network (*e. g.*, gene regulatory network or PPI network) to identify PDGs. However, cellular processes are not driven by a single type of biomolecule. In fact, one cellular process generally can span multiple biomolecular networks[31]. Thus, separately investigating different single biomolecular network naturally leads to a great discount in understanding of intricate interaction patterns during tumor initiation and progression, and eventually it will compromise the performance for identifying PDGs. Recently, the rapidly increasing of available genomic and proteomic data enables researchers to study the complex biological mechanisms on a systematic scale. For example, Liu *et al.*[32] designed a perturbation process to effectively identify essential genes and cancer genes that contribute most to the robustness of a multilayer network, which consists of a gene regulatory network, a PPI network and a metabolic network. Klosik *et al.*[33] applied a framework of interdependent networks of three biological molecule types (*i. e.*, genes, proteins and metabolites) to systematically investigate the model organism *Escherichia coli*. Valdeolivas *et al.*[34] extended the random walk with restart (RWR) algorithm to a multilayer network where different layers represent different types of interactions (*i. e.*, gene-gene or protein-protein associations) to explore the disease-related genes. These efforts generally integrate different biomolecular networks into a multilayer network to investigate the biological processes, which can deepen our understanding of the complexity of biological systems. Therefore, it is urgent to develop a computational method that can integrate different types of biomolecular networks to facilitate the identification of PDGs.

In this work, we propose a multiplex network control method namely Personalized cancer Driver Genes with Multiplex biomolecular Networks (PDGMN). The innovations of our PDGMN mainly lie in two aspects. (1) PDGMN is designed based on the control of multiplex networks, which integrate different types of biomolecular networks, such as the PPI network and gene-gene association (GGA) network. This integration ability of multiplex networks enable them to contain more informative interactions/associations than a single network, thus facilitating the PDG identification. (2) A weighted minimum vertex cover set identification algorithm is proposed, which not only resolves the difficulty of finding the optimal driver genes among multiple driver gene sets based on network control but also utilizes omics data to score the importance of network nodes, thereby enhancing the accuracy of PDG identification.

We evaluated the performance of our PDGMN on three cancer datasets from The Cancer Genome Consortium (TCGA) [35], and compared it with other PDG identification methods that only use a single biomolecular network. The experimental results show that PDGMN outperforms other existing methods in terms of Precision, Recall and F1-Score, and it also effectively infers the rare driver genes for individual patients. The results on different biomolecular networks show that even though different biomolecular networks may have their unique characteristics, PDGMN can effectively capture these characteristics to improve the identification of PDGs. In summary, PDGMN can identify sample-specific driver genes and broaden our understanding of driver gene identification from the perspective of multiplex networks.

1 Materials and methods

1.1 Datasets

The datasets used in this work contain two parts. The first part comprises the multi-omics data of three cancer types: breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), and prostate adenocarcinoma (PRAD). This omics data can be downloaded from the TCGA data portal[35] *via* the Xena platform[36]. The multi-omics data of each cancer dataset contains the somatic mutation (*i. e.*, SNP and INDEL) data, copy number variants (CNVs), and the corresponding gene expression data of tumor samples from different patients. Our study is restricted to those samples for which that somatic mutation, CNVs, and gene expression data are all available. As a result, we obtained 789 samples of BRCA, 509 samples of LUAD, and 494 samples of PRAD.

The second part is the prior-known biomolecular interaction data, which is used as the reference network to construct the sample-specific multiplex network. The prior-known biomolecular interaction data contains a PPI network and a GGA network. The PPI network, originally constructed by Cheng *et al.* [37], integrates multiple databases of protein-protein interactome with experimental evidence. To adapt this PPI network for our work, we mapped the protein Entrez IDs of the PPI network into human Gene Symbols using the tool of package clusterProfiler[38], and removed the duplicate edges. Consequently, the PPI network (named as PPI-cheng) consisting of 15 903 nodes (*i.e.*, proteins) and 213 809 interaction edges was obtained. The GGA network, initially constructed based on the repository of RegNetwork[39], collects and integrates various genetic regulation relationships from different databases. We removed the miRNAs and their interaction edges from the repository of RegNetwork, and filtered out the duplicate edges. As a result, an undirected GGA network (named as RegNet) consisting of 20 250 nodes (*i. e.*, genes) and 151 215 linking edges was constructed.

1.2 PDGMN

The overview of PDGMN is shown in Figure 1. PDGMN contains two main steps to identify PDGs: (1) constructing the weighted sample-specific multiplex network; (2) identifying driver nodes from the weighted sample-specific multiplex network based on multiplex network control, and then identifying PDGs from the driver nodes.

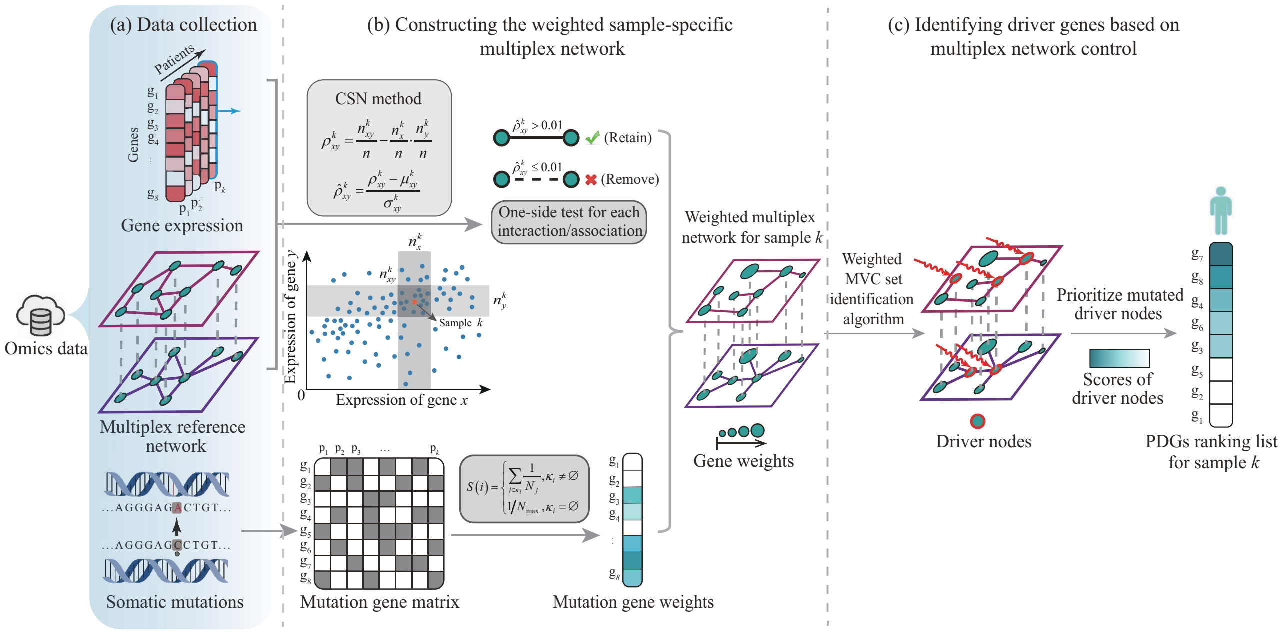

(a) Data collection
Patients
Genes
g_1 g_2 g_3 g_4 ⋮ g_8
p_1 p_2 p_k
Gene expression
Omics data
Multiplex reference network
...AGGGAGACTGT...
...AGGGAGCCTGT...
Somatic mutations
(b) Constructing the weighted sample-specific multiplex network
CSN method
$\rho_{xy}^k=\frac{n_{xy}^k}{n}-\frac{n_x^k}{n}\cdot\frac{n_y^k}{n}$
$\hat{\rho}_{xy}^k=\frac{\rho_{xy}^k-\mu_{xy}^k}{\sigma_{xy}^k}$
$\hat{\rho}_{xy}^k>0.01$ ✓ (Retain)
$\hat{\rho}_{xy}^k\le 0.01$ ✗ (Remove)
One-side test for each interaction/association
n_x^k n_{xy}^k n_y^k
Sample k
Expression of gene y
Expression of gene x
0
p_1 p_2 p_3 … p_k
g_1 … g_8
$S(i)=\begin{cases}\sum_{j\in\kappa_i}\frac{1}{N_j},\kappa_i\neq\varnothing\\ 1/N_{\max},\kappa_i=\varnothing\end{cases}$
Mutation gene matrix
Mutation gene weights
Weighted multiplex network for sample k
Gene weights
(c) Identifying driver genes based on multiplex network control
Weighted MVC set identification algorithm
Driver nodes
Prioritize mutated driver nodes
Scores of driver nodes
g_7 g_8 g_4 g_6 g_3 g_5 g_2 g_1
PDGs ranking list for sample k

Fig. 1 Overview of PDGMN

(a) Data collection. The gene expression and somatic mutation data of three cancers (BRCA, LUAD and PRAD) are collected from TCGA; the PPI network and GGA network are used to construct the multiplex reference network in which the gene nodes in GGA layer correspond one-to-one with the protein nodes in the PPI layer, but the edges in each layer of the multiplex network is different. (b) Constructing the weighted sample-specific multiplex network. The cell-specific network (CSN) method is adopted to generate the sample-specific multiplex network (SSMN) for each cancer patient based on the gene expression data of each patient. Subsequently, the somatic mutation data is used to measure the importance of nodes in SSMN. (c) Identifying personalized driver genes (PGDs). We identify driver nodes from the weighted SSMN using our multiplex network control method, and use these driver nodes to generate a personalized driver gene ranking list for each sample based on the node's importance score in network control.

1.2.1 Constructing the weighted sample-specific multiplex network

In this work, the gene expression data, somatic mutation data, and biomolecular networks (*i. e.*, the PPI network and the GGA network) are utilized to construct the weighted sample-specific multiplex networks. Firstly, the proteins and genes that are both available in the PPI network and the GGA network are selected. Then, the PPI and GGA are employed to construct a two-layer reference network M = $\left(V_1,V_2,E_1,E_2,E_3\right)$. In the reference network M, V_1 denotes the gene node set, V_2 denotes the protein node set, E_1 and E_2 denote the PPI edge set and GGA edge set, respectively, and E_3 denotes the edge set connecting gene nodes and protein nodes. The node number in V_1 is equal to that in V_2, and the gene node is linked to its unique corresponding protein node.

Secondly, due to high heterogeneities among tumor samples, two proteins/genes in the reference network M may have interactions/associations in some samples while not in other samples, resulting in the differences of sample-specific multiplex networks (SSMNs) among different samples. Thus, SSMNs are constructed based on the prior-known interaction/association in M and the statistical independency of gene expression data in individual samples. Specifically, assuming gene x and gene y are two independent variables, there is $p\left(x,y\right) = p\left(x\right)p\left(y\right)$ based on probability theory, where $p\left(x\right)$ and $p\left(y\right)$ are the marginal probability distribution of x and y, respectively, and $p\left(x,y\right)$ is the joint probability distribution. Then, the degree of independence between gene x and gene y in sample k is measured by $\rho_{xy}^k = p^{(k)}\left(x_k,y_k\right) - p^{(k)}\left(x_k\right)p^{(k)}\left(y_k\right)$, following the idea of cell-specific network (CSN) method[40]. To estimate the values of $p^{(k)}\left(x_k\right)$, $p^{(k)}\left(y_k\right)$ and $p^{(k)}\left(x_k,y_k\right)$, a scatter diagram based on gene expression data of gene x and gene y is created, in which each dot represents a sample, as shown in Figure 1. Around sample k, four boundary lines are drawn, where two lines parallel to x axis and the other two lines parallel to y axis, to generate three boxes containing n_x^k, n_y^k and n_{xy}^k dots, respectively (Figure 1). Then, $p^{(k)}\left(x_k\right)$, $p^{(k)}\left(y_k\right)$ and $p^{(k)}\left(x_k,y_k\right)$ can be estimated by n_x^k/n, n_y^k/n and n_{xy}^k/n where n is the number of samples in a cancer dataset, and ρ_{xy}^k can be estimated by Equation (1).

$$\rho_{xy}^k = \frac{n_{xy}^k}{n} - \frac{n_x^k}{n}\cdot\frac{n_y^k}{n} \tag{1}$$

Dai *et al.*[40] demonstrated that if gene x and gene y are independent of each other, no matter which distributions the genes follow, the statistic ρ_{xy}^k approximates to a normal distribution, and its mean value μ_{xy}^k and standard deviation σ_{xy}^k are defined as follows:

$$\mu_{xy}^k = 0 \tag{2}$$

$$\sigma_{xy}^k = \sqrt{\frac{n_x^k n_y^k\left(n - n_x^k\right)\left(n - n_y^k\right)}{n^4\left(n - 1\right)}}$$

The normalization of ρ_{xy}^k is performed as follows:

$$\hat{\rho}_{xy}^k = \frac{\rho_{xy}^k - \mu_{xy}^k}{\sigma_{xy}^k} \tag{3}$$

If gene x and gene y are independent of each other, $\hat{\rho}_{xy}^k$ follows standard normal distribution. Then, the one-sided hypothesis test for each interaction/association edge in M is conducted. The null hypothesis (H_0) is that genes x and y are independent in sample k, while the alternative hypothesis (H_1) is that gene x and gene y are interacted with each other in sample k. If $\hat{\rho}_{xy}^k$ is larger than the α quantile of the standard normal distribution, where α is a significant level (*e. g.*, 0.01), H_0 will be rejected, and the edge between gene x and gene y in the SSMN of sample k is retained; otherwise, the edge is removed. The hypothesis test is repeated for each edge in E_1 and E_2, and the edges between independent pairs of genes are removed. The resulting E_1^k and E_2^k are then used to construct the SSMN of sample k, *i. e.*, $\mathrm{M}^k = \left(V_1, V_2, E_1^k, E_2^k, E_3\right)$. This process is repeated for every sample in each cancer dataset to build the SSMNs.

Thirdly, as driver genes tend to be mutated in a relatively large number of samples in a cancer dataset[41-42], a score function $S\left(i\right)$ is introduced to measure the importance of different proteins/genes in M^k based on somatic mutation data. Specifically, given a cancer dataset which consists of m genes and n samples, its somatic mutation data is stored in a binary matrix $M \in R^{m\times n}$. If gene i is mutated in sample j, we set $M_{ij} = 1$, otherwise $M_{ij} = 0$. Thus, the score function $S\left(i\right)$ for gene i is defined as:

$$S\left(i\right) = \begin{cases} \sum_{j\in\kappa_i}\frac{1}{N_j}, & \kappa_i \neq \varnothing \\ 1/N_{max}, & \kappa_i = \varnothing \end{cases} \tag{4}$$

where κ_i is the set of samples in which gene i is mutated, N_j is the number of mutated genes in sample j, and $N_{\max}$ is the maximum number of mutated genes across all the samples in a cancer dataset. As defined in Equation (4), proteins/genes with high mutation frequencies are assigned relatively large values by $S\left(i\right)$. Meanwhile, a background score $1/N_{max}$ is assigned to genes without mutations in any samples in a cancer dataset. In this way, $S\left(i\right)$ not only help to identify driver genes with high mutation frequencies but also make it possible to identify driver genes with low mutation rate that nonetheless play important roles (*e. g.*, high-degree hubs) in SSMNs. Here, to facilitate the computation of minimum optimization in Equation (5), the weight of each vertex in M^k is represented by $1/S(i)$. Finally, the weighted SSMN M^k is constructed for sample *k*.

1.2.2 Identifying driver genes based on feedback vertex set control method

To identify PDGs within the SSMN, the first step is to apply the network structural control method to find a minimum set of driver nodes from the SSMN. In the viewpoint of network control, driver nodes typically play crucial roles in governing the overall state of the network system, such as steering the network state from a normal state (initiation attractor) to a disease state (dynamical attractor). Thus, in the context of cancer biology, driver nodes often correspond to genes that play important roles in cancer development and progression[29], which provide us a valuable resource to assist cancer driver gene identification[43-44]. The second step is to identify PGDs from the driver nodes based on their importance in network control.

Before solving the problem of finding minimum driver node set of the multilayer network M^k, the algorithm to find the minimum driver node set of a single layer $G_l^k\left(V_l^k,E_l^k\right)$ in M^k is introduced. Based on the framework of feedback vertex set control (FC), which has been used to large biological networks with nonlinear dynamics to identify driver nodes, the problem of finding the minimum driver node set can be solved by finding the feedback vertex set (FVS) of G_l^k[45]. Under the framework of FC, the minimum driver node set of G_l^k is equal to the minimum FVS whose removal leaves the network without cycles. Here, PDGMN adopts the approach of nonlinear control of undirected network algorithm (NCUA) [25], which treats all undirected edges as bi-directional edges, allowing the state of nodes in G_l^k to be characterized by $\mathrm{d}x_i/\mathrm{d}t = F_i\left(x_i,I_i\right)$, where x_i is the state of node *i* at time *t*, I_i is the neighboring nodes of node *i*, and $F_i\left(x_i,I_i\right)$ represents the nonlinear response of node *i* to its neighboring nodes. By considering each bi-directional edge in G_l^k as a feedback loop, the minimum FVS is equal to the minimum node set covering all the edges in G_l^k, which is fundamentally the minimum vertex cover (MVC) problem. In this way, the minimum driver node set identification with FC is transformed into the MVC problem. Although NCUA uses a simple integer linear programming (ILP) formulation to get the optimal solution of this MVC problem, it ignores the weights of nodes. To incorporate the weights of vertices, a weighted MVC set identification algorithm is proposed to find the minimum weighted driver node set of G_l^k. Specifically, for one single layer $G_l^k\left(V_l^k,E_l^k\right)$ in M^k, the MVC set in G_l^k is determined by solving the following ILP:

$$\min f = \sum_{i \in V^k} x_i + \lambda \sum_{i \in V^k} w(i) x_i \qquad (5)$$

$$s.t.\ x_i + x_j \geqslant 1, \forall e\left(i,j\right) \in E_l^k \qquad (6)$$

$$x_i \in \left\{0,1\right\}, \forall i \in V_l^k \qquad (7)$$

where the binary variable x_i indicates whether the node *i* is selected ($x_i = 1$) or not ($x_i = 0$), $w(i)$ is weight of node *i* (*i. e.*, $1/S(i)$), and λ is a penalty parameter to adjust the importance of node weights. In this work, the value of parameter λ is set to 0.1, and the effect of λ on PDGMN is detailed in Document S1 and Figure S1. Equations (6) and (7) guarantee that each edge in G_l^k is covered by at least one selected node. The objective function of ILP is to obtain the minimal number of selected nodes as well as the sum of selected nodes' weights. Compared with NCUA, our algorithm incorporates the node weight $w(i)$ to facilitate the identification of the optimal driver node set from multiple MVC sets and the penalty parameter λ to balance the importance of node weights and the size of identified driver node set. After solving the optimization problem (Equations (5)-(7)) by adopting the LP-based classic branch and bound method[46], the optimal driver node set D_l^k for the one-layer network G_l^k is obtained.

Considering the SSMN $\mathrm{M}^k = \left(V_1,V_2,E_1^k,E_2^k,E_3\right)$ as a multilayer networked system, where each layer is an undirected network with the nonlinear dynamical process, the objective of control network M^k is to find the minimum driver node set D^k of M^k. To ensure that D^k can control all layers of M^k, it is assumed that each node in a multiplex network is either a driver node in each layer or it is not a driver node in any layer[44, 47]. In other words, if a node belongs to the minimum driver node set of any single layer of M^k, it must belongs to D^k. A simple algorithm to get D^k is to merge the minimum driver node set of each single layer to get D^k, *i. e.*, $D^k = D_1^k \cup D_2^k$. However, this simple algorithm ignores the interlayer edges in the multiplex network, resulting in a suboptimal driver node set for M^k.

An improved algorithm is proposed herein to get the optimal D^k by the following steps. Step 1: for a two-layer multiplex network M^k, Equations (5) – (7) are applied to find the minimum driver node set D_1^k for the first layer G_1^k. Step 2: as nodes in each layer of M^k are linked by one-to-one interlayer edges, driver nodes in the first layer G_1^k can pass control signals to their equivalent nodes in the second layer G_2^k through interlayer edges. Thus, in the second layer G_2^k, the equivalent nodes of D_1^k and edges can also be controlled by D_1^k. After removing the equivalent nodes of D_1^k and their edges from G_2^k, the remaining nodes and edges in G_2^k denoted as $\hat{V}_2^k$ and $\hat{E}_2^k$ are still uncontrolled. Step 3: Equations (5)−(7) are applied to find the minimum driver node set $\hat{D}_2^k$ to control the remaining network $\hat{G}_2^k\left(\hat{V}_2^k,\hat{E}_2^k\right)$. Then, the union of D_1^k and $\hat{D}_2^k$, denoted as $D^k = D_1^k \cup \hat{D}_2^k$, is capable of controlling all nodes in M^k. This union set D^k is considered as the optimal driver node set for the two-layer multiplex network M^k.

Finally, the PGDs are identified from the driver nodes based on their importance in network control. Concretely, under the framework of FC, driver nodes with larger network degrees can send control signals to more neighboring nodes, and driver nodes with smaller weights (*i. e.*, $1/S(i)$) receive less penalties when solving the ILP (Equations (5) − (7)). Accordingly, to measure the importance of each driver node in network control, $Score\left(d_i\right)$ is defined as follows:

$$Score\left(d_i\right) = Norm\left(deg\left(d_i\right)\right) + Norm\left(S\left(d_i\right)\right) \tag{8}$$

where d_i is the *i*-th driver node in D^k, $deg\left(d_i\right)$ is the sum of network degree of driver node d_i in G_1^k and G_2^k, and $Norm(\cdot)$ represents the min-max normalization function. Then, mutated driver nodes are prioritized based on the values of $Score\left(d_i\right)$ to generate a PDGs ranking list of sample *k*. The mutated driver nodes represent driver nodes whose corresponding genes carry somatic mutations or CNVs.

2 Results

2.1 Performance comparison

In this section, we compared the performance of our PDGMN against four existing PDG identification methods, *i. e.*, DawnRank[20], PRODIGY[22], PersonaDrive[23], and SCS[24], and one baseline method $\mathrm{SSMN}_{\mathrm{degree}}$. These methods commonly integrate multi-omics data into a biomolecular network and employ various algorithms to identify PDGs. Bellow is the brief description of each method:

(1) DawnRank[20] evaluates the impact of each personalized mutated gene on downstream differentially expressed genes (DEGs) by applying the PageRank algorithm to a gene network, thereby identifying the most influential mutated genes as PDGs.

(2) PRODIGY[22] employs a prize-collecting Steiner tree model to evaluate the influence of personalized mutated genes on perturbed signaling pathways and identifies the most influential mutated genes as PDGs.

(3) PersonaDrive[23] constructs the bipartite network to model the relationship between personalized mutated genes and DEGs in similar patients. Then, it identifies PDGs based on the co-occurrence frequency of mutated genes and DEGs within the same signaling pathways.

(4) SCS[24] utilizes a network structural control method to measure the influence of a personalized mutated gene on downstream DEGs in a network and identifies the most influential mutated genes as PDGs.

(5) $\mathrm{SSMN}_{\mathrm{degree}}$, the baseline method, identifies high-degree hubs as PDGs by ranking personalized mutated genes based on their degrees in SSMNs.

Each method was applied to the same samples of BRCA, LUAD, and PRAD datasets obtained from TCGA data portal, which include somatic mutation data, gene expression data, and CNVs. In addition to the PPI-cheng network used in each method, PersonaDrive and PRODIGY need the pathway data to extract pathway information, and PDGMN requires the GGA network (RegNet) to construct SSMNs.

Here, the performances of different methods in identifying PDGs were evaluated at the individual level. Firstly, 711 well-established cancer driver genes from the Network of Cancer Genes (NCG 6.0)[13] were used as the ground truth, which includes 708 cancer driver genes in the Cancer Gene Census (CGC)[48] and 125 cancer driver genes in the manually curated list[49]. Secondly, following the assumption that a tumor generally needs 5 driver genes to fully explain its cancer development[18, 50-51], the top 5 genes from the PDGs ranking list of each sample were considered as predicted PDGs by each method for the following analysis. The genes in the intersection set of predicted PDGs and the ground truth were considered as true PDGs in a tumor sample. Finally, to measure the performance of each method, we calculated the average Precision, Recall, and F1-score with respect to true PDGs across the cancer-specific dataset (Document S1). Here, we do not compare with PNC[25] and pDriver[26], which are two network control-based methods to identify PDGs, since they output the unranking set of drivers for each sample, while our performance measures rely on the personalized ranking list of driver genes.

The F1-scores on cancer BRCA, LUAD, and PRAD are shown in Figure 2, and the corresponding Precision and Recall are shown in Figure S2. As shown in Figure 2 and Figure S2, we can see that PDGMN performs better than other five methods on three cancer datasets in terms of F1-score, Precision, and Recall. These results show that PDGMN owns the advantages over other methods in identifying PDGs. It is worth noting that the results of SSMN$_{degree}$ are competitive with those of DawnRank on PRAD dataset and better than those of SCS and PRODIGY on BRCA, LUAD, and PRAD datasets. SSMN$_{degree}$ simply ranks personalized mutated genes by their network degrees of the SSMN, and identifies high-degrees hubs as PDGs for each sample. The results indicate that driver genes are closely related to high-degree hubs in biomolecular networks, and the high-degree property of driver genes is helpful to identify cancer driver genes, which are consistent with the previous research findings[22, 52]. PDGMN combines the high-degree property with the mutation score to assist finding driver nodes in the SSMNs, resulting in its better results in identifying PDGs.

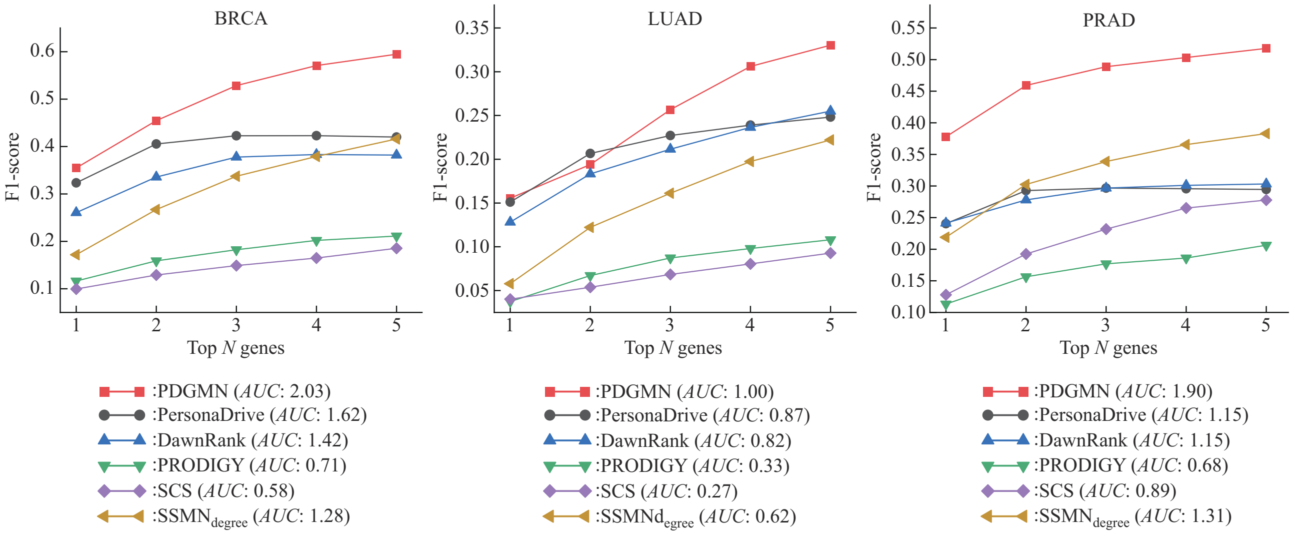


Fig. 2　F1-scores (calculated according to the top 5 predicted PDGs) of PDGMN and other 5 methods for BRCA, LUAD and PRAD cancers benchmarked on NCG 6.0

In addition to the driver genes from NCG 6.0, two other lists of well-known driver genes from the OncoKB[53] and CancerMine[54] databases were also used as ground truth driver genes to evaluate PDGMN and other methods. The results are shown in Figure S3, S4. We can see that PDGMN consistently outperforms other methods across three cancer datasets, regardless of using ground truth driver genes from OncoKB or CancerMine databases. The consistent performance across different sources of ground truth demonstrates the robustness of PDGMN and confirms that its effectiveness is not biased toward to a specific cancer gene database.

2.2　Ablation experiments of two strategies in PDGMN

In this section, the influence of two strategies in PDGMN was evaluated. The first strategy is the integration of the PPI and GGA networks into a multiplex network, which contains more informative interactions/associations than a single network. The second strategy is the application of the weighted MVC set identification algorithm, which considers both network topological properties of genes (*e. g*, high degrees) and biological omics data of genes (*e.g*, somatic mutation frequency) for the identification of driver node set. The following ablation study was carried out to investigate the effectiveness of these two strategies in improving the performance of PDGMN. For the first strategy, the performance of PDGMN was tested without the multiplex network integration. PDGMN-PPI and PDGMN-GGA were employed to denote that PDGMN was tested solely on the PPI network and the GGA network, respectively. For the second strategy, each node was assigned an equal weight of 1, and the item (*i.e.*, $\lambda \sum_{i \in V^k} w(i) x_i$) was omitted from Equation (5) when running PDGMN (referred to as Unweighted-PDGMN). We performed PDGMN, Unweighted-PDGMN, PDGMN-PPI, and PDGMN-GGA on 1 792 samples from BRCA, LUAD, and PRAD datasets. Following the above section, the top 5 genes from the PDGs ranking list of each sample were considered as predicted PDGs, and the genes in the intersection set of predicted PDGs and the ground truth were considered as true PDGs. The results of PDGMN, Unweighted-PDGMN, PDGMN-PPI, and PDGMN-GGA in term of average Precision, Recall, and F1-score across all samples are shown in Figure 3.

The result in Figure 3 indicates that PDGMN performs better than Unweighted-PDGMN, PDGMN-PPI, and PDGMN-GGA in terms of average Precision, Recall, and F1-score, with statistical significance (P<0.05). Firstly, the average Precision, Recall, and F1-score of PDGMN are 0.639 0, 0.413 9, and 0.421 3, which are 0.032 8, 0.022 4, and 0.024 1 higher than that of PDGMN-PPI, respectively, and 0.024 4, 0.017 9, and 0.020 3 higher than that of PDGMN-GGA, respectively. Compared with PDGMN-PPI and PDGMN-GGA, PDGMN identifies PDGs using the two-layer SSMN that incorporates more informative interactions/associations than the one-layer PPI or GGA network. The result suggests that the first strategy of using the multiplex network to integrate the PPI and GGA networks is effective to improve the identification of PDGs. Secondly, the result in Figure 3 reveals that Unweighted-PDGMN generates much worse results than other methods. Concretely, the average Precision, Recall, and F1-score of Unweighted-PDGMN are 0.244 1, 0.115 1, and 0.139 7 lower than that of PDGMN, respectively. Compared with PDGMN, Unweighted-PDGMN ignores the node weight and solely relies on the network topology properties of nodes to identify driver nodes, resulting in its poor performance. The results demonstrate that using the weighted MVC set identification algorithm to integrate both network topology properties and biological omics data of genes is an effective way to improve the identification of PDGs.

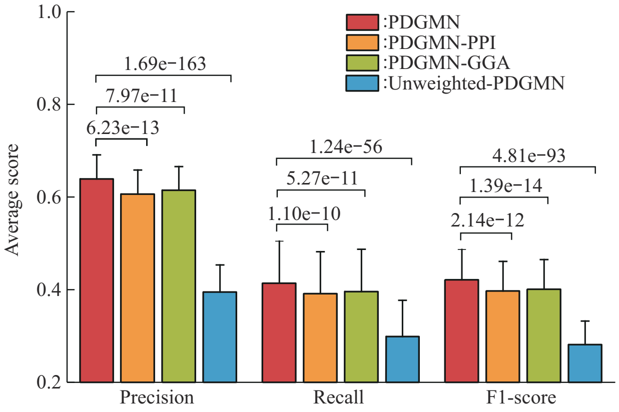


Fig. 3 Results of PDGMN (red), PDGMN−PPI (orange), PDGMN−GGA (green) and Unweighted−PDGMN (blue) in terms of average F1−score, Recall and Precision across 1 792 samples from BRCA, LUAD and PRAD datasets

PDGMN and Unweighted-PDGMN were performed on the two-layer SSMN. PDGMN-PPI and PDGMN-GGA were performed on the PPI layer and the GGA layer of SSMN, respectively. The numbers above the brackets represent the *P*-values of Wilcoxon signed-rank test between pairs of methods.

2.3 Analysis of personalized driver genes identified by PDGMN

Here, we conducted the following analysis of PDGs predicted by PDGMN. Firstly, the ability of PDGMN in identifying rare driver genes (*i. e.*, driver genes mutated in only few patients) was assessed. The predicted PDGs of each sample were divided into three groups according to the frequency that a gene is mutated among a cancer-specific cohort: rare-frequency PDGs (mutated in less than 1% patients, denoted as mf < 1%), medium-frequency PDGs (1% ⩽ mf < 10%), and high-frequency PDGs (mf ⩾ 10%). The average proportion of each group was calculated for BRCA, LUAD, and PRAD cancers, respectively, as illustrated in Figure 4a. The predicted PDGs represent the top 5 genes in the personaliezed ranking list generated by PDGMN, as described in **3.1**. As shown in Figure 4a, the proportions of rare-frequency PDGs in BRCA, LUAD, and PRAD datasets are 52%, 39%, 67%, which are higher than the proportions of medium-frequency and high-frequency PDGs. Although the rare-frequency PDGs have small values of somatic mutation score $S(i)$ (as defined in Equation (4)), they commonly have relatively high network degrees in the PPI or GGA network (Table S1). Thus, PDGMN can identify these rare-frequency PDGs through their high network degrees, as high-degree nodes play important roles in identifying the minimum driver node set. The lists of rare-frequency PDGs and their network degrees for BRCA, LUAD, and PRAD datasets are provided in Table S1.

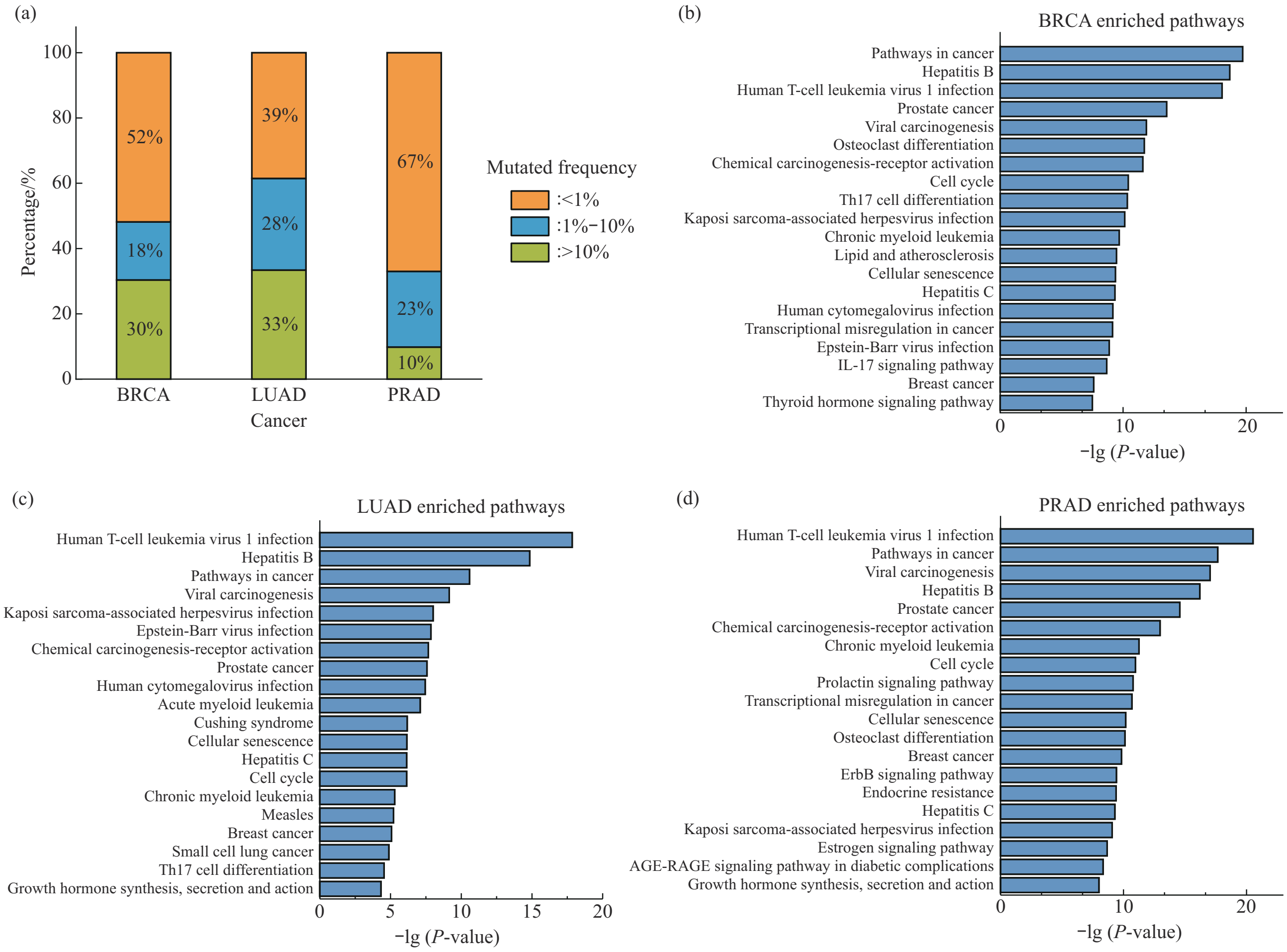


Fig. 4　Analysis of PDGs identified by PDGMN

(a) Average proportions of three groups of PDGs across samples in BRCA, LUAD, and PRAD datasets. Rare-frequency PDGs (orange), medium-frequency PDGs (blue), and high-frequency PDGs (green). (b−d) Top 20 enriched KEGG pathways of rare-frequency PDGs from all the samples in BRCA (b), LUAD (c), and PRAD (d), respectively.

Moreover, to investigate the association between the rare-frequency PDGs with the cancer development, the functional enrichment analysis for KEGG pathway was performed on the rare-frequency PDGs identified by PDGMN. The top 20 enriched KEGG pathways of rare-frequency PDGs from all the patients in BRCA, LUAD, and PRAD datasets are shown in Figure 4b−d. We can see that these rare-frequency PDGs are significantly enriched in some cancer pathways, such as Pathways in cancer, viral carcinogenesis, chemical carcinogenesis-receptor activation, breast cancer, prostate cancer, transcriptional misregulation in cancer, and so on.

In addition, to assess whether the PDGs predicted by PDGMN can provide helpful information to precision ontology for each single patient, we collected the actionable genes from TARGET database[55], which refer to the genes that are directly linked to a clinical action, and the druggable genes from the Drug-Gene Interaction Database (DGIdb)[56]. It was observed that the PDGs of each patient across three cancer datasets harbored at least one druggable target. Moreover, 1 568 out of 1 792 (87.5%) patients across three cancer datasets carried at least one actionable gene within their PDGs (Figure 5). These results indicate that the PDGs identified by PDGMN can provide assistance for developing therapeutic plans for individual patients.

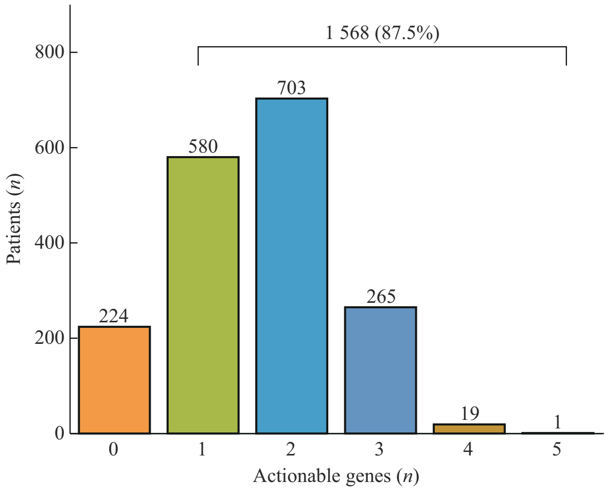


Fig. 5 Distribution of actionable genes per sample across all samples in BRCA, LUAD, and PRAD datasets

The *x*-axis represents the number of actionable genes in PDGs per sample, and the *y*-axis represents the number of samples harboring different numbers of actionable genes in their PDGs.

2.4 The influence of different biomolecular networks

To study the influence of different biomolecular networks on PDGMN, the following experiments were conducted on samples of BRCA cancer. Firstly, different biomolecular networks were collected from various sources, including three PPI networks of PPI-cheng, STRING and BioGRID, and two GGA networks of RegNet and GGI-TS. The details of these five networks are provided in Table 1, and they are used as reference networks in the following experiments.

We selected one PPI network from three PPI netowrks and one GGA network from two GGA networks, respectively, to construct a two-layer reference network. As a result, six two-layer reference networks were established: PPI-cheng+RegNet, PPI-cheng+GGA-TS, STRING+RegNet, STRING+GGA-TS, BioGRID+RegNet, and BioGRID+GGA-TS networks. Then, PDGMN was applied on six reference networks to identify their minimum driver node sets. Additionally, for each two-layer reference network, PDGMN was utilized separately to find the minimum driver node set of its PPI layer and GGA layer. The 711 well-known driver genes from NCG 6.0[13] were used as ground truth, and these well-known driver genes included in the minimum driver node set were treated as true positives. The metrics in terms of F1-score, Recall, and Precision of PDGMN are shown in Figure 6 and Figure S5. It is can be seen that while the choice of networks affects the F1-score of PDGMN, the PDGMN with the two-layer network produces a higher F1-score than that with the one-layer PPI or GGA network. The results indicate that the driver node set derived from the two-layer multiplex network can provide more informative candidate nodes for identifying PDGs than the one-layer network, either the PPI or GGA network. Thus, we conclude that the integration of PPI and GGA networks into a multiplex network is an effective strategy for improving the identification of PDGs.

Table 1 Details of five reference networks

Type	Name	Number of nodes/edges	Brief description	Reference
PPI	PPI-cheng	15 903/213 809	Integration of multiple databases of protein-protein interactome with experimental evidence	[37]
	STRING	12 385/311 434	The most confident 5% of interactions from STRING v11	[57]
	BioGRID	23 814/499 363	The human proteins and their interactions from BioGRID (version 4.3.195)	[58]
GGA	RegNet	20 250/151 215	Obtained by removing the miRNAs and their interactions from the repository of RegNetwork, where the data is inferred based on the transcription factor binding sites. Transcription factor and microRNA which are the key regulators in gene regulations	[39]
	GGA-TS	15 440/276 092	Reconstruction of the tissue-specific gene regulatory networks by considering the cell types used to derive these tissue-specific gene regulatory networks (Document S1, Table S2)	[59]

All networks in this work are accessable at https://github.com/NWPU-903PR/PDGMN.

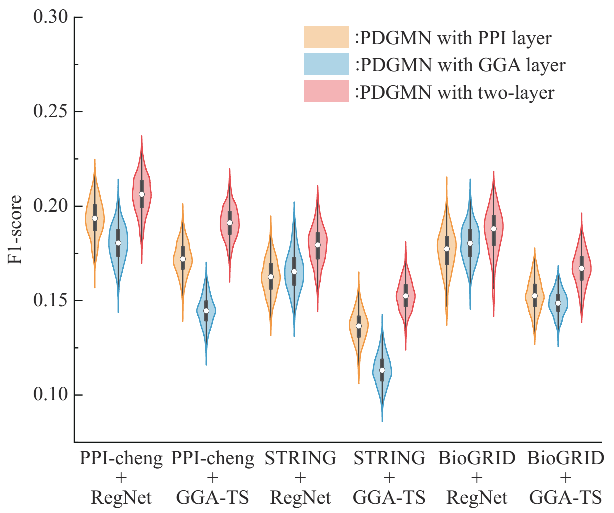


Fig. 6　F1-scores of PDGMN with different multiplex reference networks for BRCA dataset

The *x*-axis indicates six two-layer multiplex reference networks constructed by different biomolecular networks. The *y*-axis indicates the F1-score of true positive driver genes in the driver node set identified by PDGMN with different networks, *i. e.*, PPI single-layer network (orange), GGA single-layer network (blue), and two-layer network (red).

To further investigate the relationships between different biomolecular networks, the edge intersections among six two-layer reference networks were illustrated in Figure 7a. On one hand, it can be seen that for two PPI networks, PPI-cheng are greatly intersected with STRING and BioGRID, which is consistent with the fact that PPI-cheng is derived by integrating multiple databases including STRING and BioGRID. In addition, when constructing PPI-cheng network, Cheng *et al.*[37] also excluded the unreliable interactions which were inferred from omics data (*e.g.* metabolic associations, evolutionary analysis and gene expression data). That is, the interactions of PPI-cheng are not only comprehensive but also reliable. It elucidates to some extent that PDGMN performs better when combining PPI-cheng network with a GGA network into a two-layer network. For two GGA networks, the major associations in GGA-TS are also owned by RegNet, indicating that RegNet contains more comprehensive associations than GGA-TS. Thus, PDGMN performs better when combining RegNet with a PPI network into a two-layer reference network. On the other hand, the results show that the same type of biomolecular networks are greatly overlapped with each other but weakly intersected with the other type of biomolecular networks. For example, for GGA networks, the associations of GGA-TS are greatly overlapped with RegNet but weakly overlapped with other PPI networks. For PPI networks, the majority of interactions in PPI-cheng, STRING and BioGRID networks are intersected with each other but only a small part of interactions are overlapped with other GGA networks. These results suggest that different types of biomolecular networks such as PPI and GGA network may carry complementary information for each other.

To study the topological characteristics of different biomolecular networks, the frequency of

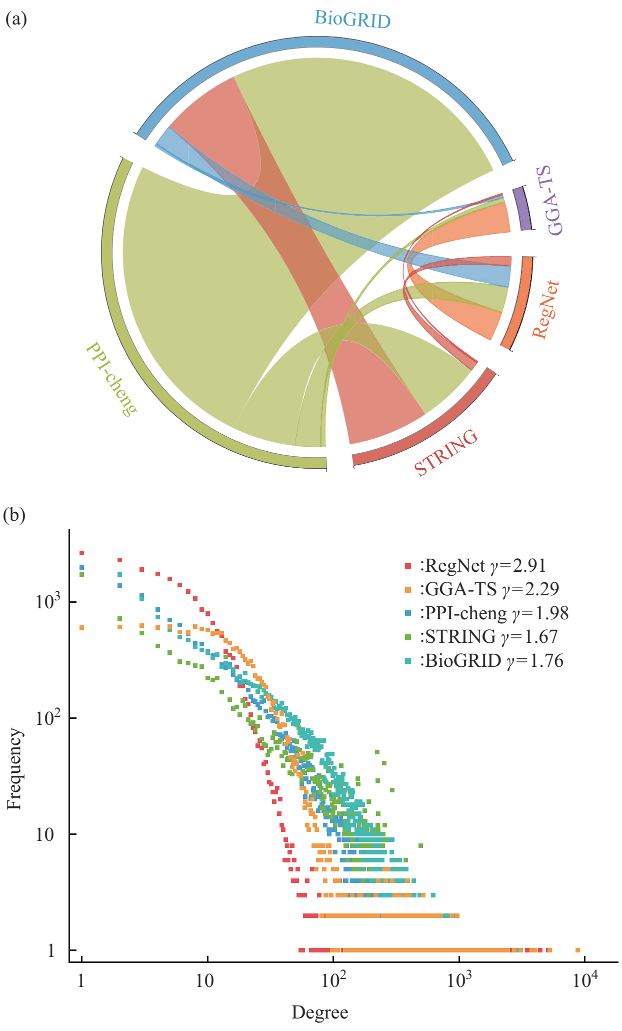


Fig. 7　Statistical results of different biomolecular networks

(a) Chord diagram for different biomolecular networks. Arcs are coloured by different networks, and thickness is proportional to the number of interactions shared by two networks. (b) The frequency of the vertex degree for different networks. Dots are coloured by different networks, *i. e.*, red for RegNet, orange for GGA-TS, blue for PPI-cheng, green for STRING, and cyan for BioGRID. The power law exponent γ of each network is also showed in the legend.

vertices with different degrees in each biomolecular network was quantified, as shown in Figure 7b. The distributions of scatters for GGA networks (*i. e.*, RegNet and GGA-TS) are distinct from that for PPI networks (*i. e.*, PPI-cheng, STRING and BioGRID). The power law exponent γ of each network was calculated using the Powerlaw package tool[60] (as indicated in the legend of Figure 7b). As shown in Figure 7b, the power law exponent γ of GGA networks are larger than that of PPI networks, indicating that edges in GGA networks are more concentrated on a few nodes with biological meanings than that in PPI networks. Thus, the minimum driver node set in PPI networks requires more nodes than in GGA networks, thus leading to PDGMN high Recall scores in PPI networks and high Precision scores in GGA networks (Figure S5). The results indicate that the PPI networks and GGA networks have their unique topological characteristics. In our PDGMN, the PPI and GGA networks were integrated into the two-layer multiplex network, which captures both the complementary information and unique topological characteristics of the PPI and GGA networks, thereby improving the identification of PDGs.

3 Discussion

Identifying and prioritizing driver genes for individual cancer patients is of great value in deciphering the complex mechanisms of cancer, helping to plan the personalized treatment. In this work, we develop a novel multiplex network control method (namely PDGMN) to identify the cancer driver genes for individual patients by integrating PPI network and GGA network into a multiplex network. While existing methods are all designed for identifying PDGs from one single biomolecular network, PDGMN uses the multiplex network to integrate different types of biomolecular networks to facilitate PDGs identification. Compared with some network control-based methods, such as SCS and PNC which require the matched omics data to construct sample-specific networks, PDGMN does not require pairwise gene expression data of tumor and normal samples from the same patient. Hence, PDGMN can be used for a more extensive range of cancer patients. Moreover, PDGMN utilizes somatic mutation data to generate biologically meaningful weights for genes/proteins in SSMNs. Compared with other network control-based methods that assign an identical value to each node in SSMNs, the weighted SSMNs can overcome the difficulty of finding the optimal driver gene set among multiple driver sets. Meanwhile, the weighted minimum vertex cover set identification algorithm considers both important topological properties of genes (*e. g.*, high degrees) and biological omics data of genes (*e. g.*, somatic mutation frequency) to improve the identification of PDGs.

We evaluated the performance of PDGMN on three cancer datasets (*i.e.*, BRCA, LUAD and PRAD) from TCGA. The comparison of PDGMN with other existing methods, which only use one single biomolecular network, revealed that PDGMN outperformed other methods in identifying PDGs. Meanwhile, PDGMN can effectively identify rare driver genes, highlighting its potential value in precision oncology. We also analyzed five networks curated by different projects or constructed by different researchers (Table 1). The results suggest that PPI networks and GGA networks may carry complementary information and possess different topological characteristics. Moreover, PDGMN can effectively capture the complementary information and unique topological characteristics of multiple biomolecular networks to improve the identification of PDGs.

Despite these advantages, there are still some limitations in the current PDGMN. Firstly, PDGMN overlooks the weight information of edges between genes or proteins, whereas the weighted GGA or PPI networks provide a more precise description of the correlations or interactions between genes or proteins. Therefore, the development of control methods for edge weighted multiplex networks may contribute to improve the PDG identification. Secondly, the sample-specific multiplex networks constructed by PDGMN simplify the interlayer connections by one-to-one edges, thus underestimating the complex interactions between different types of biomoleculars. For instance, a single transcription factor might regulate multiple genes, and a single gene may be regulated by multiple transcription factors. Therefore, the development of multilayer network methods that can handle the complex interlayer connections is another important direction for future research.

4 Conclusion

In summary, this work introduces a novel multiplex network control method, named as PDGMN, for the identification of personalized cancer driver genes. PDGMN integrates the PPI network with the GGA network into a multiplex network and utilizes gene expression and somatic mutation data to construct sample-specific multiplex networks. Then, a weighted minimum vertex cover set identification algorithm is developed to find the optimal driver node set in the sample-specific multiplex network, facilitating the identification of personalized cancer driver genes. PDGMN not only outperforms existing methods but also effectively identifies rare cancer driver genes in individual patients. Furthermore, experimental results suggest that PDGMN leverages the unique properties of different biomolecular networks through the multiplex network, thereby improving the identification of personalized cancer driver genes.

Supplementary Available online (http://www. pibb. ac.cn or http://www.cnki.net):
PIBB_20230392_Document_S1.pdf
PIBB_20230392_Figure_S1.pdf
PIBB_20230392_Figure_S2.pdf
PIBB_20230392_Figure_S3.pdf
PIBB_20230392_Figure_S4.pdf
PIBB_20230392_Figure_S5.pdf
PIBB_20230392_Table_S1.xlsx
PIBB_20230392_Table_S2.xlsx

References

[1] Stratton M R, Campbell P J, Futreal P A. The cancer genome. Nature, 2009, **458**(7239): 719-724

[2] Stratton M R. Exploring the genomes of cancer cells: progress and promise. Science, 2011, **331**(6024): 1553-1558

[3] Bailey M H, Tokheim C, Porta-Pardo E, *et al*. Comprehensive characterization of cancer driver genes and mutations. Cell, 2018, **173**(2): 371-385

[4] Hudson T J, Anderson W, Aretz A, *et al*. International network of cancer genome projects. Nature, 2010, **464**(7291): 993-998

[5] Cheng F, Zhao J, Zhao Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. Brief Bioinform, 2015, **17**(4): 642-656

[6] Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics, 2013, **29**(18): 2238-2244

[7] Lawrence M S, Stojanov P, Polak P, *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 2013, **499**(7457): 214-218

[8] Tokheim C J, Papadopoulos N, Kinzler K W, *et al*. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci USA, 2016, **113**(50): 14330-14335

[9] Bashashati A, Haffari G, Ding J, *et al*. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol, 2012, **13**(12): R124

[10] Leiserson M D M, Vandin F, Wu H T, *et al*. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet, 2015, **47**(2): 106-114

[11] Jia P, Zhao Z. VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. PLoS Comput Biol, 2014, **10**(2): e1003460

[12] Pham V V H, Liu L, Bracken C P, *et al*. CBNA: a control theory based method for identifying coding and non-coding cancer drivers. PLoS Comput Biol, 2019, **15**(12): e1007538

[13] Repana D, Nulsen J, Dressler L, *et al*. The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol, 2019, **20**(1): 1

[14] Sondka Z, Bamford S, Cole C G, *et al*. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. Nat Rev Cancer, 2018, **18**(11): 696-705

[15] Pe'er D, Hacohen N. Principles and strategies for developing network models in cancer. Cell, 2011, **144**(6): 864-873

[16] Stratton M R. Journeys into the genome of cancer cells. EMBO Mol Med, 2013, **5**(2): 169-172

[17] Mourikis T P, Benedetti L, Foxall E, *et al*. Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. Nat Commun, 2019, **10**(1): 3101

[18] Nulsen J, Misetic H, Yau C, *et al*. Pan-cancer detection of driver genes at the single-patient resolution. Genome Med, 2021, **13**(1): 12

[19] Zhang T, Zhang S W, Li Y. Identifying driver genes for individual patients through inductive matrix completion. Bioinformatics, 2021, **37**(23): 4477-4484

[20] Hou J P, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med, 2014, **6**(7): 56

[21] Bertrand D, Chng K R, Sherbaf F G, *et al*. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Res, 2015, **43**(7): e44

[22] Dinstag G, Shamir R. PRODIGY: personalized prioritization of driver genes. Bioinformatics, 2020, **36**(6): 1831-1839

[23] Erten C, Houdjedj A, Kazan H, *et al*. PersonaDrive: a method for the identification and prioritization of personalized cancer drivers. Bioinformatics, 2022, **38**(13): 3407-3414

[24] Guo W F, Zhang S W, Liu L L, *et al*. Discovering personalized driver mutation profiles of single samples in cancer by network

control strategy. Bioinformatics, 2018, **34**(11): 1893-1903

[25] Guo W F, Zhang S W, Zeng T, *et al*. A novel network control model for identifying personalized driver genes in cancer. PLoS Comput Biol, 2019, **15**(11): e1007520

[26] Pham V V H, Liu L, Bracken C P, *et al*. pDriver : a novel method for unravelling personalised coding and miRNA cancer drivers. Bioinformatics, 2021, **37**(19): 3285-3292

[27] Liu Y Y, Slotine J J, Barabási A L. Controllability of complex networks. Nature, 2011, **473**(7346): 167-173

[28] Fiedler B, Mochizuki A, Kurosawa G, *et al*. Dynamics and control at feedback vertex sets. I: informative and determining nodes in regulatory networks. J Dyn Differ Equ, 2013, **25**(3): 563-604

[29] Li M, Gao H, Wang J, *et al*. Control principles for complex biological networks. Brief Bioinform, 2019, **20**(6): 2253-2266

[30] Guo W F, Zhang S W, Feng Y H, *et al*. Network controllability-based algorithm to target personalized driver genes for discovering combinatorial drugs of individual patients. Nucleic Acids Res, 2021, **49**(7): e37

[31] Gosak M, Markovič R, Dolenšek J, *et al*. Network science of biological systems at different scales: a review. Phys Life Rev, 2018, **24**: 118-135

[32] Liu X, Maiorino E, Halu A, *et al*. Robustness and lethality in multilayer biological molecular networks. Nat Commun, 2020, **11**(1): 6043

[33] Klosik D F, Grimbs A, Bornholdt S, *et al*. The interdependent network of gene regulation and metabolism is robust where it needs to be. Nat Commun, 2017, **8**(1): 534

[34] Valdeolivas A, Tichit L, Navarro C, *et al*. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics, 2018, **35**(3): 497-505

[35] Weinstein J N, Collisson E A, Mills G B, *et al*. The cancer genome atlas pan-cancer analysis project. Nat Genet, 2013, **45**(10): 1113-1120

[36] Goldman M J, Craft B, Hastie M, *et al*. Visualizing and interpreting cancer genomics data *via* the Xena platform. Nat Biotechnol, 2020, **38**(6): 675-678

[37] Cheng F, Desai R J, Handy D E, *et al*. Network-based approach to prediction and population-based validation of *in silico* drug repurposing. Nat Commun, 2018, **9**(1): 2691

[38] Yu G, Wang L G, Han Y, *et al*. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics, 2012, **16**(5): 284-287

[39] Liu Z P, Wu C, Miao H, *et al*. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database, 2015, **2015**: bav095

[40] Dai H, Li L, Zeng T, *et al*. Cell-specific network constructed by single-cell RNA sequencing data. Nucleic Acids Res, 2019, **47**(11): e62

[41] Hou Y N, Gao B, Li G J, *et al*. MaxMIF: a new method for identifying cancer driver genes through effective data integration. Adv Sci, 2018, **5**(9): 1800640

[42] Vandin F, Upfal E, Raphael B J. *De novo* discovery of mutated driver pathways in cancer. Genome Res, 2012, **22**(2): 375-385

[43] Guo W F, Yu X, Shi Q Q, *et al*. Performance assessment of sample-specific network control methods for bulk and single-cell biological data analysis. PLoS Comput Biol, 2021, **17**(5): e1008962

[44] Zheng W, Wang D, Zou X. Control of multilayer biological networks and applied to target identification of complex diseases. BMC Bioinformatics, 2019, **20**(1): 271

[45] Zanudo J G T, Yang G, Albert R. Structure-based control of complex networks with nonlinear dynamics. Proc Natl Acad Sci USA, 2017, **114**(28): 7234-7239

[46] Lenstra H W. Integer programming with a fixed number of variables. Math Oper Res, 1983, **8**(4): 538-548

[47] Menichetti G, Dall'asta L, Bianconi G. Control of multilayer networks. Sci Rep, 2016, **6**: 20706

[48] Futreal P A, Coin L, Marshall M, *et al*. A census of human cancer genes. Nat Rev Cancer, 2004, **4**(3): 177-183

[49] Vogelstein B, Papadopoulos N, Velculescu V E, *et al*. Cancer genome landscapes. Science, 2013, **339**(6127): 1546-1558

[50] Martincorena I, Raine K M, Gerstung M, *et al*. Universal patterns of selection in cancer and somatic tissues. Cell, 2017, **171**(5): 1029-1041

[51] Campbell P J, Getz G, Korbel J O, *et al*. Pan-cancer analysis of whole genomes. Nature, 2020, **578**(7793): 82-93

[52] Shrestha R, Hodzic E, Sauerwald T, *et al*. HIT'nDRIVE: patient-specific multidriver gene prioritization for precision oncology. Genome Res, 2017, **27**(9): 1573-1588

[53] Chakravarty D, Gao J J, Phillips S, *et al*. OncoKB: a precision oncology knowledge base. JCO Precis Oncol, 2017, **2017**: PO.17.00011

[54] Lever J, Zhao E Y, Grewal J, *et al*. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Meth, 2019, **16**(6): 505-507

[55] Van Allen E M, Wagle N, Stojanov P, *et al*. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med, 2014, **20**(6): 682-688

[56] Griffith M, Griffith O L, Coffman A C, *et al*. DGIdb: mining the druggable genome. Nat Meth, 2013, **10**(12): 1209-1210

[57] Szklarczyk D, Gable A L, Lyon D, *et al*. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res, 2019, **47**(D1): D607-D613

[58] Oughtred R, Stark C, Breitkreutz B J, *et al*. The BioGRID interaction database: 2019 update. Nucleic Acids Res, 2019, **47**(D1): D529-D541

[59] Marbach D, Lamparter D, Quon G, *et al*. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat Meth, 2016, **13**(4): 366-370

[60] Alstott J, Bullmore E T, Plenz D. powerlaw: a Python package for analysis of heavy-tailed distributions. PLoS One, 2014, **9**(1): e85777

基于多层网络控制的个体化癌症驱动基因识别方法*

张　桐[1,2]　张绍武[1)**]　李　岩[1)]　谢明宇[1)]

（[1)] 西北工业大学自动化学院，信息融合技术教育部重点实验室，西安 710072；
[2)] 平顶山学院电气与机械工程学院，平顶山 467000）

摘要　目的　识别癌症驱动基因，特别是罕见或个体特异性癌症驱动基因，对精准肿瘤学至关重要。考虑到肿瘤间的高度异质性，最近有一些方法尝试在个体水平上识别癌症驱动基因。然而，这些方法大多是将多组学数据整合到单一生物分子网络（如基因调控网络或蛋白质相互作用网络）中来识别癌症驱动基因，容易忽略不同网络所特有的重要相互作用信息。为了整合不同生物分子网络的相互作用数据，促进癌症驱动基因识别，迫切需要发展一种多层网络方法。**方法**　本文提出了一种多层网络控制方法（PDGMN），利用多层网络识别个体化癌症驱动基因。首先，利用基因表达数据构建针对个体病人的个体化多层网络，其中包括蛋白质相互作用层和基因相互关联层。然后，整合突变数据，对个体化多层网络中的节点进行加权。最后，设计了一种加权最小顶点覆盖集识别算法，找到个体化多层网络中的最优驱动节点集，以提高个体化癌症驱动基因的识别效果。**结果**　在三个TCGA癌症数据集上的实验结果表明，PDGMN在个体化驱动基因识别方面优于其他现有方法，并能有效识别个体病人的罕见癌症驱动基因。特别是，在不同生物分子网络上的实验结果表明，PDGMN能够捕获不同生物分子网络的独有特征，从而改进癌症驱动基因的识别结果。**结论**　PDGMN能有效识别个体化癌症驱动基因，并从多层网络的视角，加深我们对癌症驱动基因识别的理解。本文所用的源代码和数据集可以从https://github.com/NWPU-903PR/PDGMN获取。

关键词　多层生物分子网络，多层网络控制，个体化癌症驱动基因，个体化多层网络，最小节点覆盖集

中图分类号　TP301　　**DOI**：10.16476/j.pibb.2023.0392

* 国家自然科学基金（62173271）资助项目。

** 通讯联系人。

Tel：029-88431308，E-mail：zhangsw@nwpu.edu.cn

收稿日期：2023-10-10，接受日期：2024-03-06