Interdisciplinary Sciences: Computational Life Sciences (2022) 14:151–167
https://doi.org/10.1007/s12539-021-00475-y

ORIGINAL RESEARCH ARTICLE

Integrating Protein–Protein Interaction Networks and Somatic Mutation Data to Detect Driver Modules in Pan-Cancer

Hao Wu[1,2] · Zhongli Chen[1,3] · Yingfu Wu[1] · Hongming Zhang[1] · Quanzhong Liu[1]

Received: 11 May 2021 / Revised: 20 August 2021 / Accepted: 22 August 2021 / Published online: 7 September 2021

Abstract

With the constant update of large-scale sequencing data and the continuous improvement of cancer genomics data, such as International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), it gains increasing importance to detect the functional high-frequency mutation gene set in cells that causes cancer in the field of medicine. In this study, we propose a new recognition method of driver modules, named ECSWalk to solve the issue of mutated gene heterogeneity and improve the accuracy of driver modules detection, based on human protein–protein interaction networks and pan-cancer somatic mutation data. This study first utilizes high mutual exclusivity and high coverage between mutation genes and topological structure similarity of the nodes in complex networks to calculate interaction weights between genes. Second, the method of random walk with restart is utilized to construct a weighted directed network, and the strong connectivity principle of the directed graph is utilized to create the initial candidate modules with a certain number of genes. Finally, the large modules in the candidate modules are split using induced subgraph method, and the small modules are expanded using a greedy strategy to obtain the optimal driver modules. This method is applied to TCGA pan-cancer data and the experimental results show that ECSWalk can detect driver modules more effectively and accurately, and can identify new candidate gene sets with higher biological relevance and statistical significance than MEXCOWalk and HotNet2. Thus, ECSWalk is of theoretical implication and practical value for cancer diagnosis, treatment and drug targets.

Hao Wu and Zhongli Chen are co-first author of the paper.

✉ Hao Wu
haowu@sdu.edu.cn

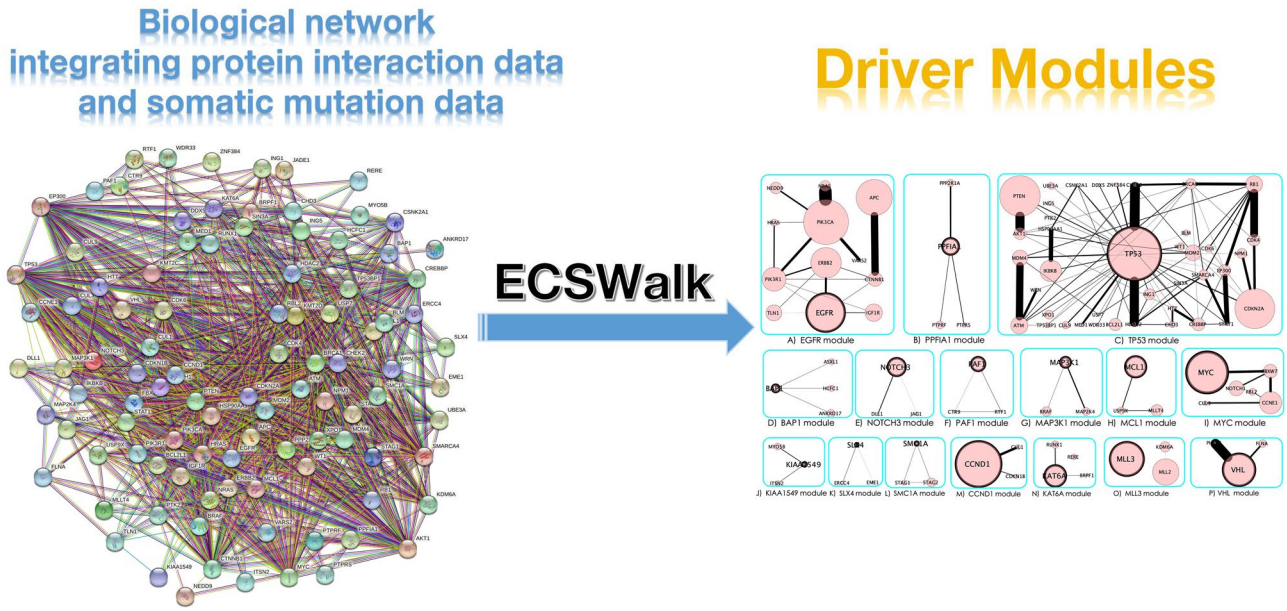
✉ Hongming Zhang
zhm@nwsuaf.edu.cn

[1] College of Information Engineering, Northwest A&F University, Yangling 712100, Shaanxi, China

[2] School of Software, Shandong University, Jinan 250100, Shandong, China

[3] Tibet Center for Disease Control and Prevention, Lhasa 850000, China

Graphic Abstract



Keywords Driver modules · Node similarity · Random walk with restart · Complex networks

Abbreviations

ICGC	International cancer genome consortium.
TCGA	The cancer genome atlas.
ECSWalk	A carcinogenic driver module detection method based on a network model.
MEXCOWalk	Mutual exclusion and coverage based random walk to detect cancer modules.
HotNet2	An algorithm for finding significantly altered subnetworks in a large gene interaction network.
KL	Kullback–Leibler, A method of describing the difference between two probability distributions.
JS	Jensen–Shannon, an improved method based on KL divergence.
PPI	Protein–protein interaction.
HINT+HI2012	A combination of high-quality protein–protein interactions from HINT and the recent HI2012 set of protein–protein interactions.
DAVID	The database for annotation, visualization and integrated discovery.
SCC	Strongly connected component of the directed graph.
TP	True positive
FP	False positive
TN	True negative
FN	False negative
GBM	Glioblastoma multiforme
BLCA	Bladder urothelial carcinoma
UCEC	Uterine corpus endometrial carcinoma
NSCLC	Non-small-cell lung cancer
PAAD	Pancreatic adenocarcinoma
CML	Chronic myelocytic leukemia
LAML	Acute myeloid leukemia
COAD	Colon adenocarcinoma

1 Introduction

In recent years, with the continuous advancement of cancer research at the bio-molecular level, targeted therapy for cancer-causing genes has become a major area in cancer research and treatment [1]. For the driver genes that have functional mutations in cancer, specific drugs can be used to control their expression and transcription levels, and effectively control the development and deterioration of cancer. However, all driver mutations in cancer cannot could be identified effectively by statistical analysis of a single mutation gene [2]. Therefore, compared with the screening of a single driver mutation gene, detecting mutated driver gene sets in cancer is of high biological relevance and statistical significance [3, 4]. In particular, the driver gene sets can be used not only to investigate the pathophysiology of cancer in more depth, but also to identify therapeutic targets for clinical cancer treatment by inferring the interaction of upstream

and downstream genes in the driver modules, which thus provides reliable theoretical basis and data support for precision medicine and personalized medicine [5].

Previous studies have found that screening for high-frequency mutations is conducive to identifying the group of mutated driver genes in cancer [6–8], such as EGFR [9], TP53 [10], PIK3CA [11] and other high-frequency mutation genes, which are screened as driver factors during tumorigenesis. However, it normally involves a large amount of time and efforts to identify high-frequency mutation gene sites in a large number of tumor samples for biological experimental verification, which is difficult to achieve with the current technical methods, experimental conditions, and research capabilities. In addition, the current methods tend to ignore the problem of mutation heterogeneity in complex diseases [12, 13]. To solve this problem, most of the previous studies utilized the high mutual exclusivity and high coverage widely existed in the genome map to explore the driver pathways that lead to the occurrence of cancer [6, 7, 14, 15].

Based on the characteristics of high mutual exclusivity and high coverage, Vandin et al. [6] proposed the Dendrix algorithm to identify driver pathways from somatic mutation data. This method introduces a penalty overlap and a reward coverage mechanism to solve the driver pathway identification maximum weight sub-matrix problem based on gene mutation data. The Dendrix algorithm not only improves coverage but also guarantees mutual exclusivity among the genes in a driver pathway. However, this iterative search method is prone to producing local optimal solutions, and in advance, it is necessary to specify the number of genes in a driver pathway. Leiserson et al. [7] proposed a Multi-Dendrix algorithm based on the Dendrix algorithm, which detects multiple driver pathways at the same time. The algorithm utilizes linear regression to simultaneously detect multiple gene sets that meet high coverage and high mutual exclusivity, and can generate a globally optimal solution. However, the Multi-Dendrix algorithm needs to pre-specify both the maximum number of genes in a driver pathway and the number of driver pathways. Thus, the two algorithms do not have good universality and robustness, and tend to have certain limitations when being applied to different data sets.

Pan-cancer data analysis provides new ideas and methods for clinical diagnosis and treatment across cancer types [16–19]. As one application, Leiserson et al. [17] proposed the HotNet2 algorithm by integrating differential expression genes, significant mutation genes and protein–protein interaction networks to detect combinations of rare somatic mutations. Under the principle of thermal diffusion and the random walk with restart model, the edge weight that becomes stable after the random walk is restarted as the weight of the directed edge. By removing the directed edge with a small weight, the method detects strongly connected components as driver modules in the directed graph. Although the algorithm reduces the output of false positive results and improves the accuracy of the prediction results, the conversion probability just considers the degree of the vertex during the random walk process, but ignores the mutual exclusivity among genes. Therefore, the problem of gene mutation heterogeneity cannot be effectively solved in the pan-cancer data of great different varieties.

Based on the HotNet2 algorithm, Rafsan et al. [19] proposed a MEXCOWalk algorithm to detect driver modules using split and expansion techniques. The algorithm utilizes mutual exclusivity between genes and coverage scores of gene set to reflect the edge weight of the network. Then the random walk with restart strategy is used to construct a weighted directed network, and the split and expansion techniques are utilized to identify driver modules. Although the set of driver genes identified by this algorithm has high mutual exclusivity and high coverage, the algorithm ignores the mutual exclusivity between expanded leaf nodes and seed modules in the expansion stage of small modules. Therefore, this algorithm reduces the accuracy of driver module identification to a certain extent.

As mentioned above, although the previous methods can detect the gene sets with high mutual exclusivity and high coverage, they just focus on the mutual exclusivity and coverage between genes, instead of the topological structure of complex networks. To effectively solve the problem of mutated gene heterogeneity and improve the accuracy of driver modules, this study proposes a driver module detection algorithm (ECSWalk) based on gene mutation and human protein–protein interaction network. The algorithm takes into account aspects, such as high mutual exclusivity and high coverage between genes, and high similarity of topological structure. First, the complex network topology analysis method is used in human protein–protein interaction network data to calculate the topological similarity between network nodes, and then the two characteristics of high coverage and high mutual exclusivity of the mutated genes are combined to obtain the weight of the vertices and edges in the human protein–protein network. The weights of vertices in the human protein–protein network are obtained according to the coverage of mutated gene, and the random walk with restart strategy is utilized to calculate the weights of edges in the network by the three characteristics, namely, the coverage, the mutual exclusivity, and the similarity of the topological structure between the nodes. Second, based on the weighted network constructed in the previous step, the large modules are split into several candidate gene sets using the method of the induced subgraph. In addition, the greedy strategy is utilized to add the nodes in the leaf module to the seed module to achieve the optimal gene sets. These mutated gene sets with high mutual exclusivity, high coverage and high similarity of the topological structure are likely to work

as driver modules in cancer [20, 21]. This study not only applies the analysis method of complex network topology to the biological network, but also improves the method of determining module size based on split and expansion. The study clarifies the interactions between genes in the detected driver modules, which promote the study of cancer pathogenesis and drug targets.

2 Methods

2.1 Mutual Exclusivity and Coverage

To accurately detect and classify a large number of genes in the cancer genome map, and reduce the error in the actual biological sequencing experiment process, this study adopts the definition of mutual exclusivity and coverage [19] as shown below.

Let $G(V, E)$ denote the protein–protein interaction (PPI) network, each vertex $u_i \in V$ corresponds to a protein in PPI network, and each protein u_i corresponds to a mutated gene g_i. The undirected edge $(u_i, u_j) \in E$ in PPI network corresponds to the interaction between gene pair (g_i, g_j). Therefore, the node g_i represents both a gene and the corresponding protein in G. The sample with gene g_i mutated is represented by S_i, and $M \subseteq V$ is a subset of genes. For any pair of genes $g_i, g_j \in M$, $g_i \neq g_j$, if $S_i \cap S_j = \varnothing$, the genes in M are mutually exclusive.

The mutual exclusivity of gene subset M is represented as

$$ED(M) = \frac{\left|\cup_{\forall g_i \in M} S_i\right|}{\sum_{\forall g_i \in M} |S_i|} \tag{1}$$

if $ED(M) = 1$, then the genes in the subset M are mutually exclusive. That is, at most one gene within the subset M is mutated in each sample, and it is possible that all genes of some samples within the subset M are not mutated.

The coverage of gene subset M is represented as

$$CD(M) = \frac{\left|\cup_{\forall g_i \in M} S_i\right|}{\left|\cup_{\forall g_i \in V} S_i\right|} \tag{2}$$

If $CD(M) = 1$, then the gene subset M completely covers all patients, That is, at least one gene within the subset M is mutated in each sample, and it is possible that none gene of some samples within the subset M is mutated.

Figure 1 is a mutation matrix, in which each row represents a patient, each column represents a gene, the black rectangular box M_{ij} represents g_j is mutated in sample S_i, and the white rectangular box M_{ij} represents g_j is not mutated in sample S_i. The mutation matrix includes two gene sets (M_1, M_2), and the exclusive degree of each gene set is 1, that is, a gene set is considered mutually exclusive if at most one gene within the gene set is mutated in each sample.

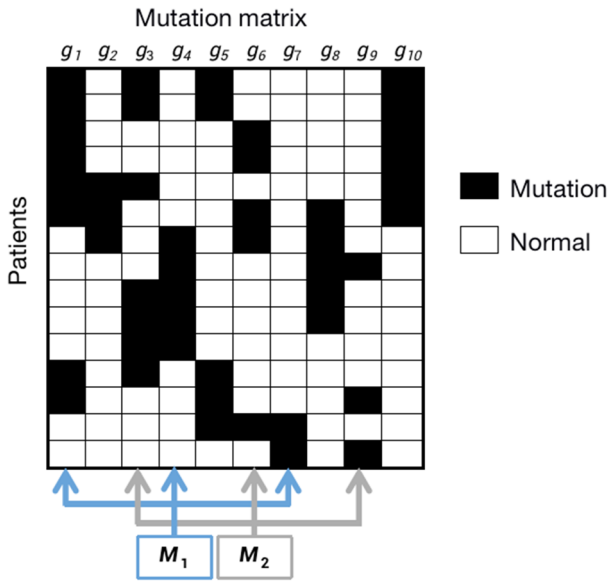


Fig. 1 Mutation matrix (M_1, M_2 are two driver modules with mutual exclusivity $ED(M) = 1$ and coverage $CD(M) = 1$)

Errors in real biological data will interfere with the calculation of mutual exclusivity and coverage, therefore, a gene set that is approximately mutually exclusive and high coverage is usually considered as a driver pathway.

2.2 Node Similarity

The abnormal local area network composed of nodes and their neighbor nodes may affect the performance of the entire network in complex networks, where neighbor nodes refer to the nodes within one step of a core node. Therefore, to measure the topological relationship of each mutated gene in the same driver module, we set the local area network with node g_i as the center and its direct neighbor node as the radius. Combined with the characteristics of the local area network structure, we propose node similarity based on Jensen–Shannon (JS) divergence and analyze the topological structure of the protein–protein interaction network [22–24]. The similarity is mainly defined by the discrete probability set, and the index construction process mainly includes the following two steps, namely, to construct the probability set [25], and to define the node similarity according to the JS divergence.

Construction of the probability set. Let d_i be the degree of the ith gene node, and d_{max} be the maximum node degree in the local area network. Suppose there are N nodes in the probability set of one node, $N = d_{max} + 1$. The sum of node degrees D_{g_i} of gene node g_i in its local area network is expressed as follows:

$$D_{g_i} = \sum_{j=1}^{n} d(j) \tag{3}$$

where N is the number of genes in the local area network, and $d(j)$ denotes the degree of the jth gene in the local area network of gene g_i.

In the local area network, the discrete probability of gene g_i is expressed as follows:

$$p(i) = \frac{d_i}{D_{g_i}} \tag{4}$$

Standardize the discrete probabilities of gene g_i, and sort the discrete probabilities of genes in the local area network of gene g_i from large to small, and obtain the set of discrete probabilities as $P(i)$:

$$P(i) = (p_i(1), p_i(2), ..., p_i(n), ..., p_i(N)) \tag{5}$$

where $p_i(n)$ represents the discrete probability value of the nth gene in the local area network with N gene nodes ($n \leqslant N$). Therefore, in the discrete probability set $P(i)$, the N elements in set $P(i)$ are the discrete probability values of each node in the local area network.

Jensen–Shannon (JS) divergence is an improved method based on KullbackLeibler (KL) divergence, which is known as relative entropy. In information theory, relative entropy is a measure of the difference between two probability distributions $P(x)$ and $Q(x)$. The difference between the Shannon entropies of the two probability distributions $P(x)$ and $Q(x)$ is the value of KL divergence, which is expressed as follows:

$$\begin{aligned} \mathrm{KL(P\|Q)} &= \sum_{\mathrm{x\in X}} \mathrm{P(x)} \log \frac{1}{\mathrm{Q(x)}} - \sum_{\mathrm{x\in X}} \mathrm{P(x)} \log \frac{1}{\mathrm{P(x)}} \\ &= \sum_{\mathrm{x\in X}} \mathrm{P(x)} \log \frac{\mathrm{P(x)}}{\mathrm{Q(x)}} \end{aligned} \tag{6}$$

The KL divergence is asymmetric, so the position exchange of the two probability distributions $P(x)$ and $Q(x)$ will get different results. JS divergence is a variant form of KL that solves asymmetric problems, which is expressed as follows:

$$\begin{aligned} JS(P||Q) =& \frac{1}{2} KL\left(P(x)\| \frac{P(x)+Q(x)}{2} \right) \\ &+ \frac{1}{2} KL\left(Q(x)\| \frac{P(x)+Q(x)}{2} \right) \end{aligned} \tag{7}$$

In this study, we suppose that two adjacent genes g_i and g_j in the constructed gene network correspond to two different probability sets $P(i)$ and $P(j)$, where $P(i)$ and $P(j)$ have the same number of genes in the local area network. According to the set of discrete probability constructed in formula (6), KL divergence value between genes g_i and g_j is expressed as

$$D_{KL}(P(i)||P(j)) = \sum_{k=1}^{N} p_i(k) log \frac{p_i(k)}{p_j(k)} \tag{8}$$

JS divergence value is obtained based on the KL divergence value, which is expressed as

$$\begin{aligned} D_{JS}(P(i)||P(j)) =& \frac{1}{2} KL(P(i)|| \frac{P(i)+P(j)}{2}) \\ &+ \frac{1}{2} KL\left(P(j)|| \frac{P(i)+P(j)}{2} \right) \end{aligned} \tag{9}$$

Therefore, we investigate the similarity of gene pairs using network topology characteristics. Node similarity is defined as follows:

$$SIM(g_i, g_j) = 1 - D_{JS}(P(i)||P(j)) \tag{10}$$

Obviously, the larger the $SIM(g_i, g_j)$ value, the more similar the two gene nodes in the network. It can be seen from formula (10) that the value range of $SIM(g_i, g_j)$ is [0, 1]. $SIM(g_i, g_j) = 1$ indicates that the two gene nodes in the network have the same topological structure.

As shown in Fig. 2, each gene in the network has a specific topological structure. We first choose three nodes g_2, g_4 and g_6 and calculate the discrete probability set of the three nodes according to formula (5), that is, $P(g_2) = [0.42, 0.28, 0.28, 0]$, $P(g_4) = [0.27, 0.27, 0.25, 0.25]$, $P(g_6) = [0.27, 0.27, 0.25, 0.25]$. Topological similarity of the two nodes can be calculated according to formula (10), that is, $SIM(g_2, g_4) = 0.95$, $SIM(g_2, g_6) = 0.95$, $SIM(g_4, g_6) = 1$. From the results of similarity, we can see that nodes g_4 and g_6 have the highest topological similarity. It can also be seen from Fig. 2 that nodes g_4 and g_6 have the same neighbors g_3 and g_5; however, nodes g_2 and g_4 have only one common

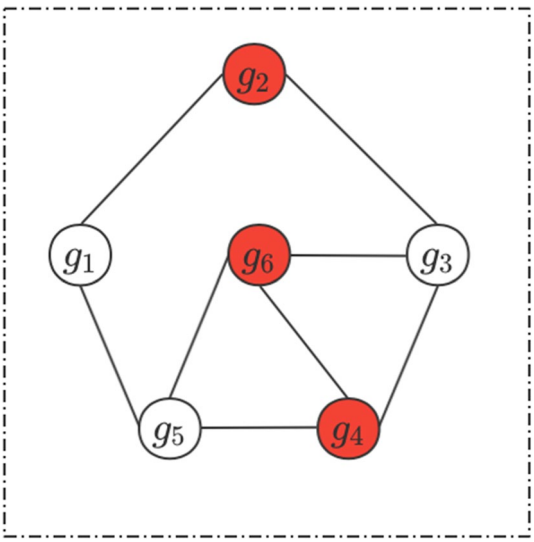


Fig. 2 Network Topology

neighbor g_3, and nodes g_2 and g_6 also have only one common neighbor g_3.

2.3 Construction of Edge-Weighted Networks

Given a PPI network $G(V, E)$, where node set $V = (u_1, u_2, u_3, ..., u_n)$ represents the set of mutated genes corresponding to the PPI network, edge set $E = \{e = (u_i, u_j)\}$ includes these edges satisfying the condition that if there are edges between proteins in the protein–protein network corresponding to the mutated genes, then there are also edges between these mutated genes.

Construction of a weighted undirected graph G_ω. For each vertex $g_i \in V$, the coverage of vertex g_i represents the weight of vertex g_i, that is, $\omega(g_i) = CD(g_i)$. Obviously, the more the mutated samples in gene g_i , the greater the weight of the vertex g_i.

Taking into account the chance of increasing the coexistence of a gene and its surrounding genes, where surrounding genes refer to the genes within one step of a core gene. This study defines the set of node g_i and its direct neighbor nodes as the local area network $Ne(g_i)$ as follows:

$$Ne(g_i) = \{g_i\} \bigcup (\cup_{\forall(g_i,g_j)\in E} g_j) \tag{11}$$

To balance the mutual exclusivity between genes and the opportunities for the coexistence between a gene and its surrounding genes, this study utilizes the average value of $ED(Ne(g_i))$ and $ED(Ne(g_j))$ as the mutual exclusivity $ED(g_i, g_j)$ of gene pairs in the network, as shown below:

$$ED(g_i, g_j) = \frac{ED(Ne(g_i)) + ED(Ne(g_j))}{2} \tag{12}$$

To reduce the chance of a single gene with large coverage added to the edge weight, the product of the coverage of two genes is used to represent the coverage $CD(g_i, g_j)$ between gene pairs, as shown below:

$$CD(g_i, g_j) = CD(\{g_i\}) \times CD(\{g_j\}) \tag{13}$$

The study integrates the three characteristics of mutual exclusivity, coverage, and similarity among gene pairs to calculate the edge weight of the weighted undirected graph as follows:

$$\omega(g_i, g_j) = \begin{cases} \frac{2\times SIM(g_i,g_j)}{\frac{1}{ED(g_i,g_j)}+\frac{1}{CD(g_i,g_j)}} & SIM(g_i, g_j) \neq 0 \\ & ED(g_i, g_j) \neq 0 \\ & CD(g_i, g_j) \neq 0 \\ 0 & otherwise \end{cases} \tag{14}$$

The principle of thermal diffusion is utilized to construct a weighted directed graph by performing a random walk with restart on G_ω [17]. The random walk with restart means that the source node gene g_i transfers to its neighboring nodes with a certain probability, and they utilize the restart probability to transfer to the source node again. This process is repeated until it reaches a stable state. The formula is expressed as follows:

$$F_{t+1} = (1 - \beta)PF_t + \beta F_0 \tag{15}$$

where F_0 is the initial state of the source node gene g_i, and $F_0 = CD(g_i).F_t$ is the probability distribution at time t; β represents the probability of returning from the current node to the initial node in the process of restarting the random walk. There are two options for restarting the random walk from the current node, returning to the initial node or going to the neighboring node. The probability of the two options are β and 1-β, respectively, where $0 \leqslant \beta \leqslant 1$, which is used to control the heat of the source node diffusion to the rest nodes of the network. It is necessary to choose a suitable β, where all source nodes retain most of the heat in their direct neighbor nodes [17]. According to [17, 19], the value of β in the study is set to be 0.4; E represents the transition probability matrix of the restart random walk process, which is positively correlated with the edge weight, as shown below:

$$P(g_i, g_j) = \begin{cases} \frac{\omega(g_i,g_j)}{\sum_k \omega(g_i,g_j)} & (g_i, g_j) \in E \\ 0 & otherwise \end{cases} \tag{16}$$

where $\sum_k \omega(g_i, g_j)$ represents the sum of the edge weights between the source node g_i and its direct neighbor nodes.

As the value of t increases, F_{t+1} gradually converges, and then the random walk restarts until it reaches a stable state [26]. The edge weight value F is calculated according to the following formula [17]:

$$F = \beta(I - (1 - \beta)(P(g_i, g_j))^{-1}F_0 \tag{17}$$

where I is the identity matrix. Restart the random walk to create a directed edge with weight F for each pair of gene pair g_i and g_j $(i \neq j)$ [17], and finally realize the construction of a weighted directed graph G_d. The algorithm is described as follows.

Algorithm 1 Construction of the weighted directed network

Input: a PPI network $G(V, E)$,gene mutation sample set S
Output: $G(V, E, F)$
1: $Initialization: j = |E|, i = 1, \beta = 0.4$
2: **for** i *to* j **do**
3: **compute** $ED(g_i, g_j), CD(g_i, g_j), SIM(g_i, g_j), \omega(g_i, g_j)$
4: **if** $ED(g_i, g_j) \neq 0$ and $CD(g_i, g_j) \neq 0$ and $SIM(g_i, g_j) \neq 0$ **then**
5: $genenetwork[g_i][g_j] = \omega(g_i, g_j)$
6: $i = i + 1$
7: **if** $(g_i, g_j) \in E$ **then**
8: $F = \beta(I - (1-\beta)(\frac{\omega(g_i,g_j)}{\sum_k \omega(g_i,g_j)}))^{-1} F_0$

2.4 Driver Modules Detection

In a directed graph, two vertices are strongly connected if there is at least one directed path between the two vertices. A graph is regarded as a strongly connected graph if any two vertices in the graph are strongly connected. The strongly connected component (SCC) refers to the maximal strongly connected subgraph in the directed graph. The strongly connected component (SCC) division method of the directed graph is utilized to generate the driver modules in this study [17]. Figure 3 is a directed non-strongly connected graph; however, $\{g_3, g_4, g_5, g_6\}$ is the strongly connected component (SCC) of the directed graph. The detection process of driver modules is composed of the following three steps.

The first step is to create a set of initial candidate modules. The SCC is first employed as an initial set of candidate modules. The minimum weight edge in G_d is iteratively deleted until strongly connected subgraphs are generated from G_d, and then the strongly connected subgraph is added to the initial module set P. Finally, all the modules in P whose gene number is less than *min_module_size* are removed. The above process is carried out iteratively until the number of genes in P decreases to *total_genes*. We finally obtain the initial module set $P = (M_1, M_2, ..., M_r)$.

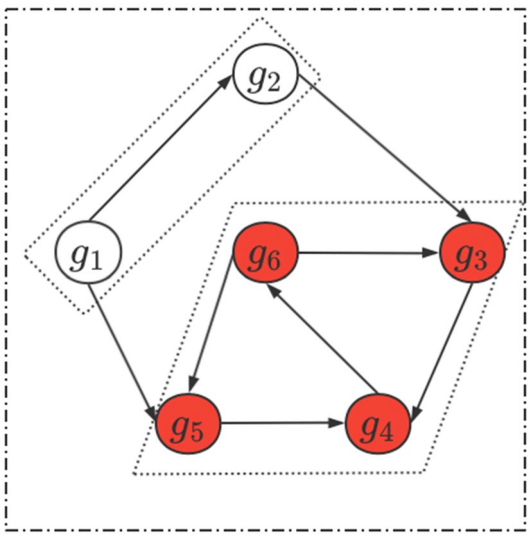


Fig. 3 Directed non-strongly connected graph. ($\{g_3, g_4, g_5, g_6\}$ is the strongly connected component (SCC) of the directed graph.)

The second step is to split the large and medium-sized modules into module set P [19]. For the weighted directed graph G_d and a module M_q, $G_d(M_q)$ denotes the gene set of derived subgraphs in the directed graph G_d (the derived subgraph is different from the strongly connected subgraph), which corresponds to the genes in M_q. L denotes the set of derived subgraphs, as shown below:

$$L = \{G_d(M_q)\} \tag{18}$$

Let *split_size* be the degree of the node with the largest value in the subgraph derived by module M_q, the modules with more nodes than *split_size* will be split as large modules. In the splitting process, $G_c \in L$ is a subgraph derived from directed graph M_q, v' is the node with the largest out-degree value in G_c, and $IN(v')$ represents the local area network of v' in G_c. If the number of nodes in $IN(v')$ is above *min_module_size*, then they will be classified as seed modules, otherwise they will be classified as leaf modules, where the leaf module is a small module with the number of nodes less than *min_module_size*. All the strongly connected subgraphs that meet the conditions in the directed graph G_c are classified in the same way.

The third step is to add leaf modules to the seed module. The leaf node g_m connected to any node in the seed module is selected to extend the seed module by utilizing the greedy strategy, and the extension function is defined as follows:

$$G(g_m) = \overline{G^{in}(g_m)} - \overline{G^{out}(g_m)} \tag{19}$$

where $\overline{G^{in}(g_m)}$ represents the average weight of the edge between node g_m in the leaf module and the node in the seed module, $\overline{G^{out}(g_m)}$ represents the average weight of the edge between node g_m in the leaf module and the rest of the nodes in the leaf module. If $\overline{G^{in}(g_m)}$ is higher $\overline{G^{out}(g_m)}$, then the node is added to the seed module.

Algorithm 2 Driver modules detection

Input: $G_d(V, E, F), total\_genes, min\_module\_size$
Output: Driver module set P

repeat
  $P = SCC(G_d)\ and\ M_q \in P$
  $P = P - M_q\ with\ |M_q| < min\_module\_size, E = E - e\ with\ e = E_{min}$
until $(\left|\cup_{\forall M_q \in P} M_q\right| == total\_genes)$
$split\_size = outdeg(G_d(M_q))_{max}$
for $M_q \in P\ and\ |M_q| > split\_size$ **do**
  $L = \{G_d(M_q)\}\ and\ G_c \in L$
  while $L = \varnothing$ **do**
    $IN(v^{'}) \subseteq G_c\ with\ v^{'} = outdeg(G_c)_{max}$
    $L = L - G_c\ and\ G_c = G_c - IN(v^{'})$
    $seed_q = seed_q \cup IN(v^{'})\ or\ leaf_q = leaf_q \cup IN(v^{'})$
    for $M_j \in SCC(G_c)$ **do**
      $seed_q = seed_q \cup M_j\ or\ leaf_q = leaf_q \cup M_j\ or\ L = L \cup M_j$
  for $M_i \in leaf_q$ **do**
    $seed_q = seed_q \cup M_i\ with\ G(g_m) \geqslant 0$
  $P = seed_q$

3 Experimental Results and Analysis

3.1 Data Preprocessing

This study utilizes the somatic mutation data and the combined human PPI network data from HINT+HI2012 [17]. Somatic mutation data comes from the TCGA pan-cancer data set containing 12 cancer types, which are composed of 3281 samples with 20472 SNVs and 4334 samples with 720 CNAs. According to the data preprocessing method of [17], the samples with hyper-mutation and the genes that are low-expression in all tumor types are first screened out. And then, MutSigCV tool is applied to filtering the genes without obvious mutations in SNVs, after that 218 genes are deleted, 5 samples without SNV and 1973 samples with only CNA are deleted, and 7894 genes with <3 RNA-seq reads in> 30% of tumors of each cancer type are deleted. Finally, we obtain the data set containing 3110 samples and a total of 11565 somatic mutation genes. (2) HINT+HI2012 as the PPI network data includes high-quality interaction database (HINT) and human interaction database (HI2012). The data preprocessing is as follows, a merge operation is first performed based on the interaction relationship in the HINT and HI2012 database. Then, we delete closed loops and duplicate edges in the new PPI network. Finally, we obtain a protein–protein interaction network composed of 9858 proteins and 40704 interactions.

3.2 Parameter Setting

If the number of genes in a driver module is less than 3, the gene set is not usually considered as a driver module

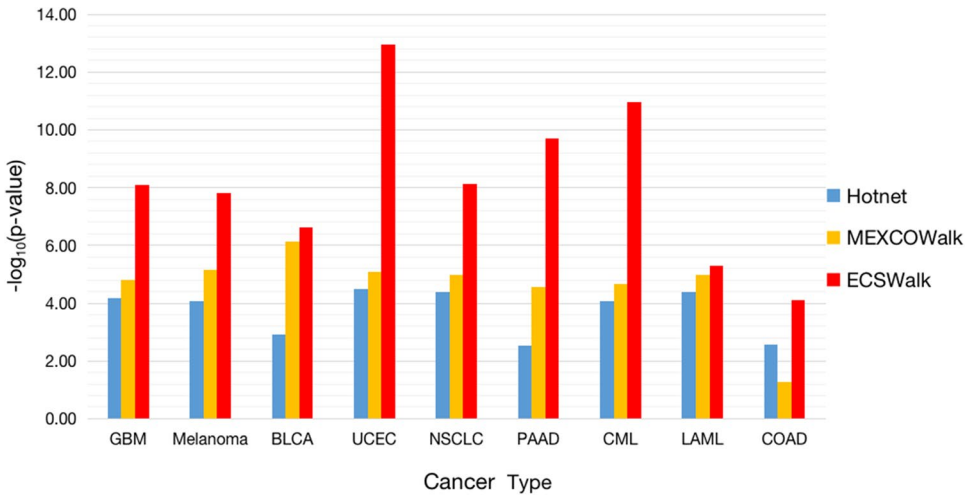


Fig. 4 Comparison of enrichment effect

[13, 17]. Therefore, the minimum module size, namely, *min_module_size*, is set to be 3.

3.3 Enrichment Analysis

To verify the enrichment effect of the modules identified by ECSWalk, we utilize functional annotation tools (DAVID) to analyze the enrichment of the driver modules detected by ECSWalk, HotNet2, and MEXCOwalk algorithm. The results in nine types of cancer are shown in Fig. 4. The value is set to be 100 to obtain the driver modules in the ECSWalk and MEXCOWalk, and HotNet2 obtains 14 consensus driver modules.

As shown in Fig. 4, the enrichment effect of the driver modules detected by ECSWalk is better than that detected by HotNet2 and MEXCOWalk algorithms among the nine types of cancer, especially in Glioblastoma Multiforme (GBM), Melanoma, Uterine Corpus Endometrial Carcinoma (UCEC), Non-Small-Cell Lung Cancer (NSCLC), Pancreatic Adenocarcinoma (PAAD) and Chronic Myelocytic Leukemia (CML). Therefore, the driver modules identified by ECSWalk show extremely high statistical significance and biological relevance.

3.4 Comparison of Module Accuracy

To evaluate the accuracy of the driver modules detected by ECSWalk, this study uses the Accuracy and F-measure evaluation indices to measure the accuracy of the driver modules based on the known pathways [27]. The higher the Accuracy value, the better the classification effect. The higher the F-measure value, the more driver modules can be enriched in the known biological pathways, indicating that the method is more accurate in mining driver modules. The calculation formulas of Accuracy and F-measure are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{20}$$

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{23}$$

where True Negative (*TN*) indicates the number of modules in which negative classes are predicted to be negative classes; True Positive (*TP*) indicates the number of modules in which positive classes are predicted to be positive classes; False Negative (*FN*) indicates the number of modules in which positive classes are predicted to be negative classes; False Positive (*FP*) indicates the number of modules in which negative classes are predicted to be positive classes.

The following formula is utilized to calculate the enrichment of the driver modules in one known biological pathway.

$$p\ value = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{24}$$

where N represents the total number of genes, K represents the number of genes in a known biological pathway, n represents the number of genes in a driver module, and k represents the number of overlapping genes between a known biological pathway overlaps and a driver module. The driver modules with p value <0.01 are set to be the positive type, and the driver modules with p value $\geqslant 0.01$ are set to be the negative type. The Benjamin–Hochberg method is used to correct all p values.

Figure 5 shows the enrichment performance of the driver modules obtained by ECSWalk, MEXCOWalk and HotNet2 based on the known biological pathways obtained using the DAVID tool for enrichment analysis. It can be seen from Fig. 5a that the ACC value of the ECSWalk is 1 in the nine types of cancer, which shows that ECSWalk has extremely high accuracy in detecting driver modules. Specifically, the ACC value of the ECSWalk is 100%, 200% and 50% higher than that of the HotNet2, respectively, in the three cancers of GBM, BLCA and PAAD, and the ACC value of the ECSWalk is 50%, 200%, 50%, 33.3%, 100% and 100% higher than MEXCOWalk, respectively, in GBM, Bladder Urothelial Carcinoma (BLCA), UCEC, NSCLC, PAAD and Acute Myeloid Leukemia (LAML). In addition, of note, the ACC value of the ECSWalk is 1, but the ACC values of the MEXCOWalk and HotNet2 are 0 in Colon Adenocarcinoma (COAD). Therefore, ECSWalk has better performance than MEXCOWalk and HotNet2 in detecting driver modules in cancer.

As can be seen in Fig. 5b, the F-measure values of the ECSWalk algorithm are 1 in the nine types of cancer, which means that the accuracy and recall of the ECSWalk algorithm are both 1 in these nine cancers. This shows all the predicted driver modules are positive. It is interesting to note that the F-measure values of gene sets detected by MEXCOWalk and HotNet2 algorithms are both 0 in COAD. Therefore, ECSWalk has a better capability in detecting driver modules in nine types of cancers.

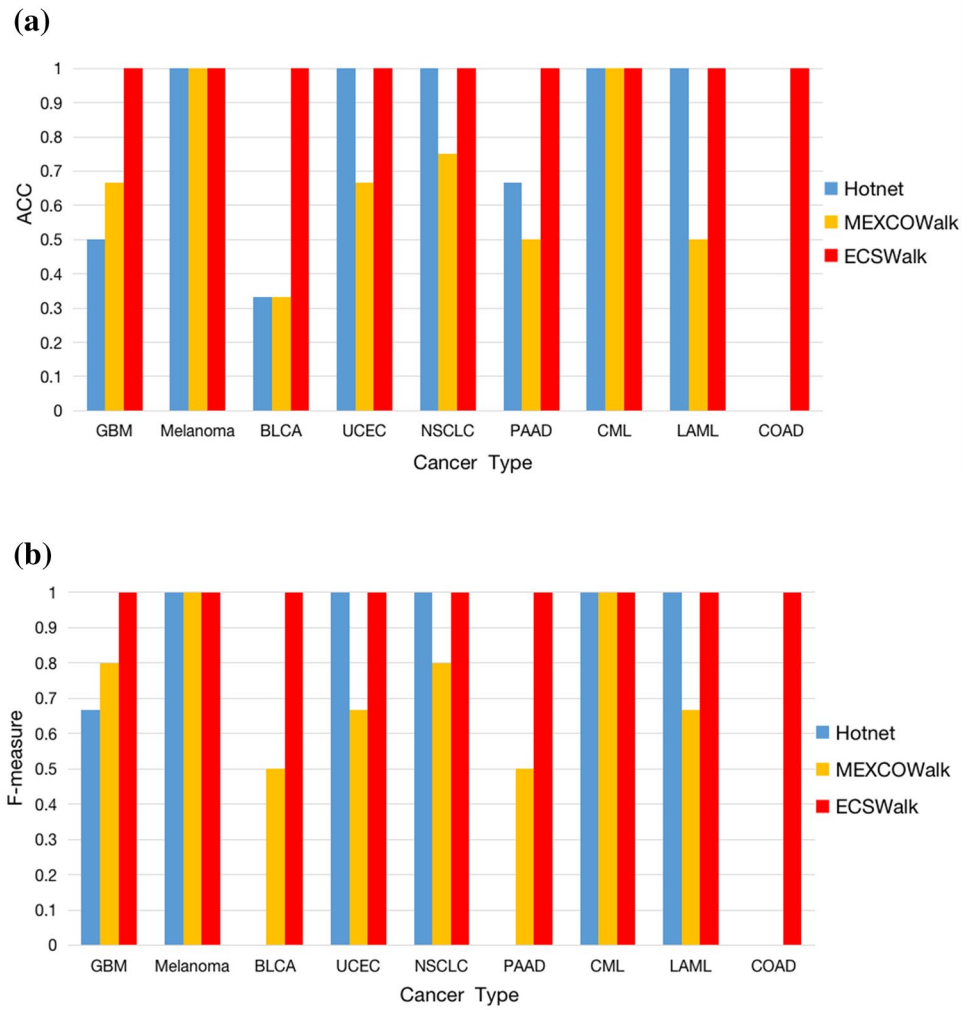


Fig. 5 Comparison and evaluation of three algorithms. (**a**) ACC values, (**b**) F-measure values

3.5 Comparison of the Optimal Modules

3.5.1 EGFR Module

The weighted diagram of the EGFR module detected by ECSWalk is shown in Fig. 6A. The EGFR module detected by ECSWalk, MEXCOWalk and HotNet2 is shown in Table 1. The genes of the EGFR module identified by each method together with the overlapping genes and non-overlapping genes are shown in Table S1 (*SI Appendix*, *Table S*1).

It can be seen from Table 1 that the *p* value of the EGFR module detected by ECSWalk is 1.09E-13, and the *p* values of the EGFR module detected by MEXCOWalk and HotNet2 are 8.1E-07 and 6.9E-06, respectively. Although the number of genes in the EGFR module detected by ECSWalk is less than that detected by MEXCOWalk, the EGFR module detected by ECSWalk has higher coverage than those of the other two algorithms. Therefore, the gene set detected by ECSWalk has higher biological relevance and statistical significance than those detected by the other two algorithms.

The EGFR module detected by ECSWalk mutates in 62.77% (1952/3110) samples, in which PIK3CA, EGFR, APC, ERBB2 and PIK3R1 have high mutation frequency, and mutate in 19.36%, 8.39%, 7.52%, 6.37% and 4.98% samples, respectively. This module is enriched in a variety of cancers, especially when the p value of this module in UCEC is 1.09E-13 according to the DAVID enrichment analysis. It can be also seen from Fig. 6A that PIK3CA and NRAS, PIK3R1 have large edge weights, and PIK3R1 and HRAS have a large edge weight, which indicates that the four mutated genes have strong interaction relationships. In the ErbB signaling pathway (Fig. 7a), EGFR and IGF1R promote the expression of PIK3CA and PIK3R1, and phosphorylation promotes the expression of HRAS and NRAS. At the same time, HRAS and NRAS promote the expression of PIK3CA and PIK3R1. Besides, CTNNB1 and APC have large weight values, which indicates that the two mutated

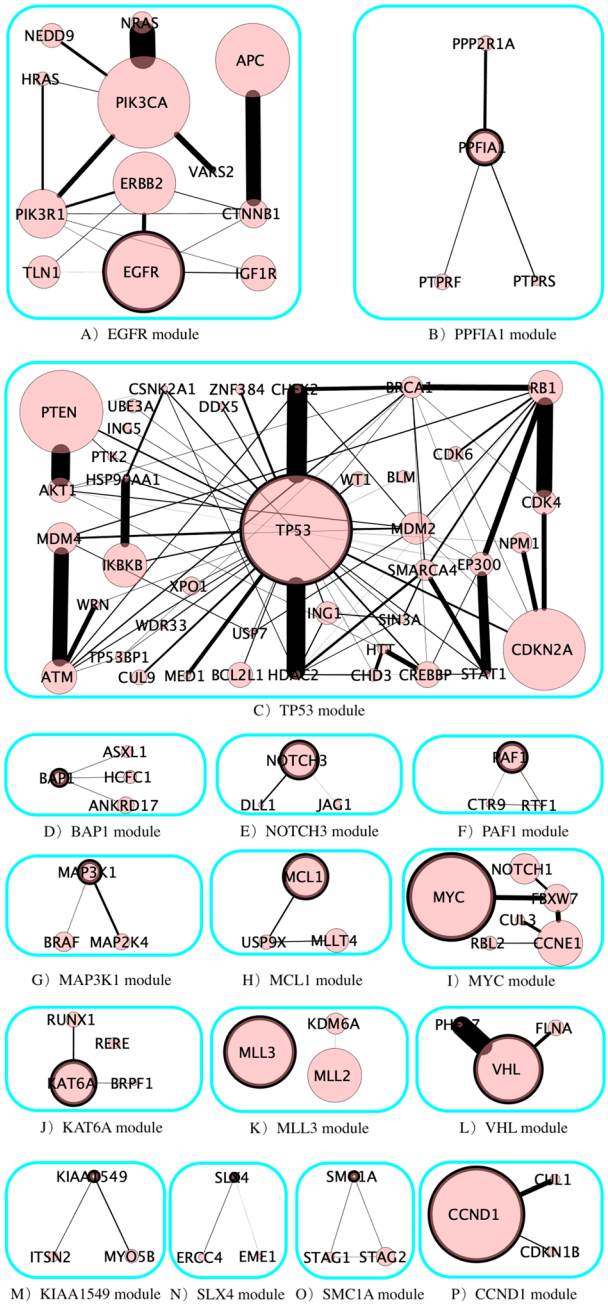


Fig. 6 Driver modules detected by ECSWalk. (The genes in the rough edge represent the names of the dysregulated mo-dules. The size of the circle is proportional to the mutation frequency, and the thickness of the line segment is proportional to the weight between no-des.)

genes have a strong interaction relationship. In the Wnt signaling pathway (Fig. 7b), CTNNB1 promotes the expression of itself and APC. This shows that the mutations of CTNNB1 and APC constitute two key genes in the Wnt/β-Catenin pathway, and thus can be used as predictors of cervical cancer susceptibility [28].

3.5.2 PPFIA1 Module

The weighted diagram of the PPFIA1 module detected by ECSWalk is shown in Fig. 6B. The PPFIA1 module detected by ECSWalk, MEXCOWalk and HotNet2 is shown in Table 2. The genes of the EGFR module identified by each method together with the overlapping genes and non-overlapping genes are shown in Table S2 (*SI Appendix*, *Table S2*). HotNet2 does not mine any genes contained in the module, and the gene set detected by ECSWalk contains one more gene PPFIA1 than that detected by MEXCOwalk. This module has a coverage of 7.72% and mutual exclusivity of 96.77%. Although the PPFIA1 modules detected by the three algorithms do not show enrichment information in the DAVID enrichment analysis, studies have shown that there is an interaction between genes within the module [29, 30]. The accuracy of the module can be verified in the previous research [30] in which PTPRF, a surface tyrosine phosphatase receptor, and its adapter, PPFIA1, govern fibronectin fibrillogenesis and vascular morphogenesis by driving active $\alpha 5\beta 1$ integrin recycling. Besides, the proteins encoded by PTPRF and PTPRS are important members of the PTP family of protein tyrosine phosphatases. PTPS is a signal molecule that regulates cell growth cycle, differentiation process, mitosis, gene mutations and other cell life processes. The lack of PTPRS and PTPRF affects the proliferation of mandibular cells, and results in craniofacial deformities. In cells lacking PTPRS and PTPRF, the WNT and BMP signaling pathways are dysregulated [29]. Therefore, PPFIA1, PTPRS and PTPRF may be mutated in the same pathway and cause cancer. In summary, the PPFIA1 module detected by ECSWalk is of more biological relevance compared with the modules detected by the other two algorithms.

3.5.3 TP53 Module

The weighted diagram of the TP53 module detected by ECSWalk is shown in Fig. 6C. The TP53 module detected by ECSWalk, MEXCOWalk and HotNet2 is shown in Table 3. The genes of the EGFR module identified by each method together with the overlapping genes and non-overlapping genes are shown in Table S3 (*SI Appendix*, *Table S3*). The TP53 module detected by ECSWalk has higher coverage than those detected by MEXCOWalk and HotNet2. The p value of the TP53 module detected by ECSWalk is 3.25E-20 according to the DAVID enrichment analysis, and the p values detected by MEXCOWalk and HotNet2 are 5.6E-08 and 3.84E-06, respectively. Therefore, the gene set detected by ECSWalk has higher biological relevance and statistical significance than those detected by the other two algorithms.

The TP53 module detected by ECSWalk mutates in 70.09% (2180/3110) samples. This module has significant enrichment in CML cancer, and the p value of this module in CML cancer

Table 1 EGFR module

Algorithm	Gene set	Number of modules	Coverage	*P* value
HotNet2	EGFR ERBB2 AREG ELF3 ERBB4 LRIG1 OSMR	7	17.94%	6.9E-06
MEXCOWalk	ERBB2 CDKN2A FLNA ERBB4 ATM MCM2 IGF1R MDM4 PHF17 VHL NPM1 CDH1 STK11 MDM2 EGFR TLN1	16	45.98%	8.1E-07
ECSWalk	PIK3CA ERBB2 EGFR TLN1 PIK3R1 CTNNB1 IGF1R NEDD9 APC VARS2 HRAS NRAS	12	47.85%	1.09E-13

Table 2 PPFIA1 module

Algorithm	Gene set	Number of modules	Coverage	*P*-value
HotNet2	N/A	N/A	N/A	N/A
MEXCOWalk	PPFIA1 PPP2R1A PTPRF	3	6.82%	N/A
ECSWalk	PTPRS PPFIA1 PPP2R1A PTPRF	4	7.72%	N/A

is 1.1E-11 according to the DAVID enrichment analysis, which indicates that the module has high biological relevance. It can be seen from Fig. 6C that TP53, ATM, CHEK2, MDM2, MDM4, PTEN, CDKN2A, CDK4 and CDK6 have large edge weights, which indicates that the nine mutated genes have strong relationships. In the P53 signaling pathway (Fig. 7d), phosphorylation of ATM promotes the expression of CHEK2, and phosphorylation of CHEK2 promotes the expression of TP53; TP53 inhibits the expression of CDK4 and CDK6 by promoting the expression of CDKN1A, and TP53 promotes the expression of upstream MDM2 and the expression of downstream MDM2 and PTEN; MDM4 promotes the expression of MDM2, but inhibits the expression of TP53; CDKN2A inhibits the expression of MDM2, and MDM2 inhibits the expression of MDM4. This indicates that regulating the p53 signaling pathway can interfere with CML, thereby regulating the occurrence and development of CML [31].

3.6 Analysis of the Remaining Modules

3.6.1 BAP1 Module

The weighted diagram of the BAP1 module detected by ECSWalk is shown in Fig. 6D, and the mutual exclusivity of the BAP1 module is 96.95%. The BAP1 module is a PR-DUB protein complex composed of BAP1 and ASXL1, which activates its downstream tumor suppressor genes, such as SOCS1/2, VHL and TXNIP by combining with transcription factors [32]; In blood tumor cells, the BAP1 mutation makes the C-terminal truncated ASXL1 mutation protein completely lose its ability to bind to transcription factors, causes the ASXL1 mutation protein to significantly weaken the transcription and regulation functions of the BAP1-ASXL1-FOXK1/K2 complex through a dominant negative mutation effect, and reduces the expression of tumor suppressor genes, thereby regulating glucose metabolism, JAK-STAT, hypoxia perception and other tumor-related signaling pathways. This thus contributes to promoting the proliferation and self-renewal of leukemia cells, and further inhibiting cell apoptosis under hypoxia [32]. It can be seen that ASXL1 and BAP1 have high biological relevance, and may exist in the same driver pathway and work together to promote the occurrence of cancer.

3.6.2 NOTCH3 Module

The weighted diagram of the NOTCH3 module is shown in Fig. 6E. The NOTCH3 gene has high coverage and is altered in 129 samples. The mutual exclusivity of this module is 98.82%, and it has been reported that the NOTCH signaling pathway plays a vital role in the tumor microenvironment, and the ligand genes DLL1 and JAG1 of the NOTCH signaling pathway have mutations in a variety of cancers, such as DLL1, JAG1 and NOTCH3 existing high probability of mutations in patients with colon cancer [33]. Besides, DLL1 and JAG1 can control the rate of neural development in the NOTCH3 gene expression domain, and JAG1 can activate the NOTCH signal transduction in the V1 and DL6 domains. DLL1 can also send signals to nerve cells outside the NOTCH gene expression domain [34]. Therefore, DLL1, JAG1 and NOTCH3 genes may exist in the same driver pathway.

Table 3 TP53 module

Algorithm	Gene set	Number of modules	Coverage	*P* value
HotNet2	CCND1 CDKN2A CUL9 NPM1 PTEN TP53 AGBL2 ALS2CR8 AMOTL1 ANKRD12 BRPF1 CACHD1 CARNS1 CDKN2AIP CELSR3 CHD8 EPHA3 HECW2 IFT140 IWS1 MAGI2 MAST3 MDM4 MLL5 PLEKHA8 PRDM2 PRKRIR SCAPER SETD2 SMG1 SMG5 SMG7 SNRK SPTBN2 STK11IP STRADA SWAP70 TEP1 TRIP12 TTLL5 UACA ZFP91 ZMIZ1 ZNF227 ZNF668	45	68.39%	3.84E-6
MEXCOWalk	E4F1 PTGS2 BMP1 ELL PPP1R13L WDR33 SMYD2 HINFP HIPK2 EGR1 KAT8 STK4 NOC2L EHMT1 CUL9 SNRPN CABLES1 WT1 WWOX CCT5 ARID3A HSPA9 ZNF384 RFWD2 TOP1 PLK3 RNF20 ERCC6 TOPORS TP53 RB1CC1 CDKN1B CUL1 KDM5A BRCA1 CTNNB1 CDK4 CCND1 RB1 CDK6	40	49.58%	5.6E-8
ECSWalk	BLM CSNK2A1 PTK2 WDR33 ING1 USP7 ING5 BRCA1 STAT1 UBE3A EP300 MDM2 MDM4 MED1 ATM CHD3 CREBBP WRN CUL9 IKBKB WT1 HSP90AA1 CHEK2 BCL2L1 CDK4 CDK6 ZNF384 CDKN2A XPO1 SIN3A HDAC2 HTT DDX5 TP53 AKT1 NPM1 RB1 TP53BP1 PTEN SMARCA4	40	70.09%	3.25E-20

3.6.3 PAF1 Module

The weighted diagram of the PAF1 module detected by ECSWalk is shown in Fig. 6F. The PAF1 module detected by ECSWalk mutates in 4.34% (135/3110) samples, and the PAF1 gene mutates in 104 samples. The mutual exclusivity between genes in the module is 100%. This module has a significant enrichment relationship in CDC73/PAF1 complex, and the *p* value of this module in CDC73/PAF1 complex is 2.9E-9 according to the DAVID enrichment analysis, which indicates that the genes in this module have high biological relevance. Besides, CTR9, as the main component of the RTF1 complex, participates in the assembly of the PAF1 complex by the TRP domain, and CDC73/PAF1 is a polyprotein complex associated with general RNA polymerase II and RNA polymerase II transcription factor complexes [35], and it may be involved in the beginning and extension of transcription. Furthermore, the mutated genes PAF1 and CTR9 in the CDC73/PAF1 complex can cause a variety of diseases including malignant tumors. The PAF1 module also has a significant enrichment relationship in transcription elongation from RNA polymeraseII promoter pathways. The p value of this module is 2.7E-8 according to the DAVID enrichment analysis. It has been reported in [36] that RTF1, LEO1 and CTR9 are the main members of the PAF1/RNA polymerase II complex, PAF1, RTF1 and LEO1 have frequent interactions with PAF1, CDC73 and PolII, and the deletion of PAF1 or CTR9 will lead to similar severe pleiotropic phenotypes. Therefore, the genes in this module have high biological relevance and may exist in the same driver pathway to cause cancer.

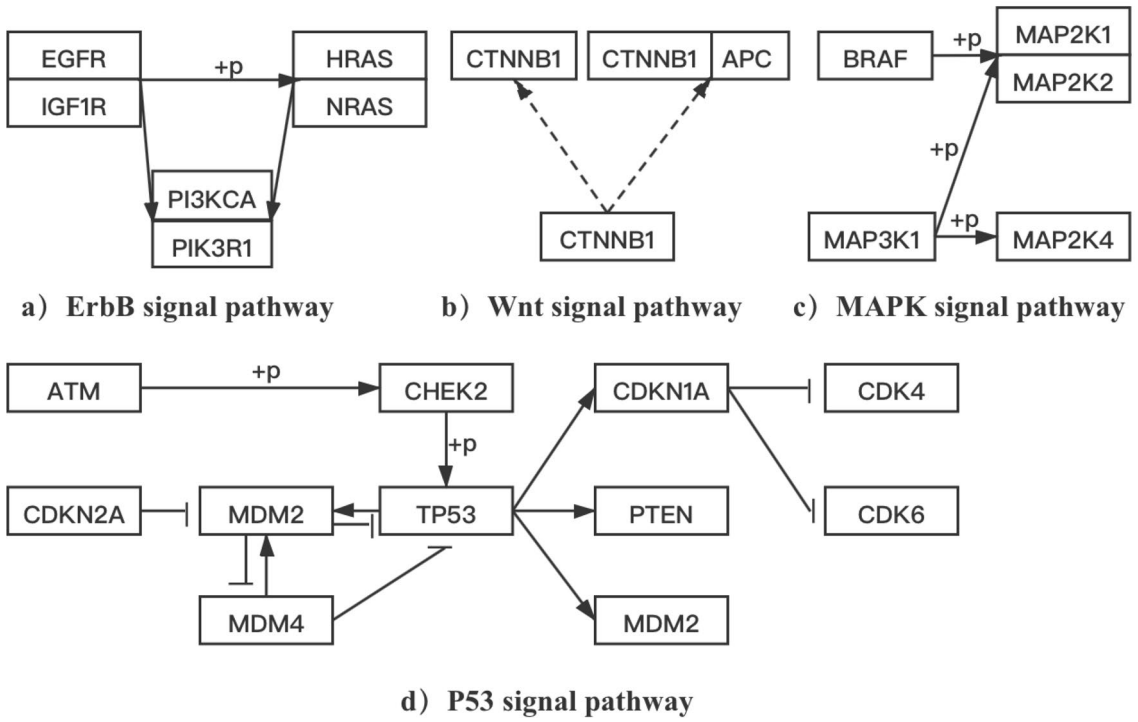


Fig. 7 Gene interaction diagram

3.6.4 MAP3K1 Module

The weighted diagram of the MAP3K1 module detected by ECSWalk is shown in Fig. 6G. The MAPK signaling pathway is significantly enriched in this module, and its *p* value is 1.3E-3 according to the DAVID enrichment analysis. The phosphorylation of MAP3K1 in this module activates MAP2K4, MAP2K1 and MAP2K2, and the phosphorylation of BRAF activates MAP2K1 and MAP2K2 (Fig. 7c). It has been reported in [37] that inhibiting the MAPK pathway can lead to lung cancer cell apoptosis. Therefore, the genes in this module have high biological relevance and may exist in the same driver pathway to cause cancer.

3.6.5 MCL1 Module

The weighted diagram of the MCL1 module detected by ECSWalk is shown in Fig. 6H. It has been reported in [38] that USP9X and MCL1 in this module have high biological relevance. For example, USP9X stabilizes the expression of MCL1 and promotes cell survival, and high expression of USP9X leads to an increase in the amount of MCL1 protein in human follicular lymphoma and diffuse large B-cell lymphoma, and vice versa. Therefore, USP9X is usually used as a target for clinical treatment by maintaining the stability of the amount of MCL1 and other proteins in human malignant tumors [38]. It can be seen that MCL1 and USP9X have high biological relevance and may exist in the same driver pathway.

3.6.6 MYC Module

The weighted diagram of the MYC module detected by ECSWalk is shown in Fig. 6I. The module has more significant enrichment in the nucleoplasm, and the *p* value of this module in nucleoplasm is 8.3E-5 according to the DAVID enrichment analysis, which indicates that this module has high biological relevance. It can be seen from Fig. 6I that MYC has the largest coverage in the module, and MYC is altered in 9.10% (283/3110) samples. It has been reported in [39] that MYC-encoded protein is a multifunctional nucleolar phosphate protein, which regulates target genes as a transcription factor during the cell life cycle and cell transformation process. Overexpression, mutation, translocation and rearrangement of MYC are closely related to the occurrence and development of a variety of cancers. It can be seen from Fig. 6I that FBXW7 and MYC have a large edge weight value, so these two genes may likely have biological relevance. It has been reported in [40] that the decrease of FBXW7 function is linked to a lower overall survival rate in muscle invasive bladder cancer and is thus linked to MYC accumulation.

4 Discussion

The HotNet2, MEXCOWalk, and ECSWalk models utilize the restart random walk algorithm to construct directed weighted biological networks. HotNet2 uses only gene coverage information to construct the biological networks, which ignores the mutual exclusivity among genes, so it cannot effectively solve the problem of gene mutation heterogeneity. MEXCOWalk improves HotNet2 by adding mutually exclusive information among mutated genes to construct biological networks. However, errors in the actual biological sequencing data processing will interfere the calculation of mutual exclusivity and coverage. To improve the accuracy of driver module detection effectively, ECSWalk adds the topological similarity of nodes in biological networks. The biological network constructed by ECSWalk not only contains the biological correlation between genes, but also reflects the topological correlation between genes. Driver genes with the same or similar biological attributes will be divided into the same module, which can more accurately reflect the true functional relevance. , a greedy strategy is utilized to expand the detected leaf modules, which can add effective genes with topological significance and screen out the detected unrelated genes in the seed module. The EGFR module just verify this fact. The EGFR module detected by ECSWalk includes fewer mutated genes compared with that by MEXCOWalk, but the EGFR module detected by ECSWalk has higher coverage and smaller p value. It shows that it is reasonable and necessary to use greedy strategy to improve the accuracy of driver module detection.

ECSWalk identifies more new candidate gene sets for biological verification, and the findings indicate that the identified candidate gene sets have high biological relevance and statistical significance. The enrichment effect of the driver modules detected by ECSWalk is better than that detected by HotNet2 and MEXCOWalk among the nine types of cancer, especially in GBM, Melanoma, UCEC, NSCLC, and CML. The EGFR, PPFIA1 and TP53 modules detected by ECSWalk have higher coverage than those detected by the other two algorithms. In particular, the PPFIA1 module detected by ECSWalk contains more PTPRS genes compared with the other two algorithms, which makes the PPFIA1 module higher biological important and statistical significant. Besides, we found that the EGFR module has extremely high enrichment in UCEC (p value: 1.09E-13), and the TP53 module has extremely high enrichment in CML (p value: 1.1E-11). The results are of theoretical significance and practical value for cancer diagnosis, treatment and drug targets. There are also limitations in this study, that is, although the greedy strategy can accurately identify the driver modules, this also leads to a decrease in the operating efficiency of the algorithm to some extent.

5 Conclusion

This study proposes a carcinogenic driver module detection algorithm (ECSWalk) by integrating somatic mutation data and protein–protein interaction network. This study first calculates the similarity between connected nodes in the protein–protein interaction network. Then, a weighted network is created by calculating mutual exclusivity, coverage and topological structure similarity between mutation genes, and a restart random walk clustering method is utilized to detect the carcinogenic driver modules. Finally, an induced subgraph method is utilized to split the large modules, and a greedy strategy is utilized to expand the small modules to generate a set of driver modules. The experimental results show that the ECSWalk can detect more accurate driver modules than the other two algorithms. The driver modules detected by ECSWalk have a lower p value by the DAVID enrichment analysis, which indicates that the results of the algorithm have higher biological relevance and statistical significance. It also shows that ECSWalk can achieve good results by combining biological characteristics and complex network characteristics in detecting carcinogenic driver modules. Therefore, the application of complex network topology to biological networks is conducive to the study of the pathogenesis of cancer based on the inherent properties of the data itself, and it is also useful for researchers and medical practitioners when carrying out research from the aspect of complex network topology. Besides, ECSWalk can identify the target gene set in some common cancers accurately, and analyze the biological relevance and statistical significance of the driver modules, which thus helps to enrich our understanding of the pathogenesis of cancer.

Supplementary Information The online version supplementary material available at https://doi.org/10.1007/s12539-021-00475-y.

Acknowledgements We thank Jihua Dong for her careful proofreading, and also thank Bing Zhou, Zhaoheng Ai, Mengdi Liu, Pengyu Zhang and Haoru Zhou for their helpful advice and discussions.

Author Contributions Conceive and design the experiments: HW ZC. Perform the experiments: HW ZC. Analyze the data: ZC YW. Contribute reagents/materials/analysis tools: ZC YW QL. Write the paper: HW ZC. Consult on the final version of the paper and edit the paper: HW ZC HZ. The authors read and approve the final version of the manuscript.

Funding The work was supported by the National Natural Science Foundation of China (Grant No.61972322), the Natural Science Foundation of Shaanxi Province (Grant No. 2021JM-110), the Humanities and Social Science Fund of Ministry of Education of China (Grant No.18YJCZH190) and the Fundamental Research Funds of Shandong University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability The data and codes we used can be download from https://github.com/HaoWuLab-Bioinformatics/wu-group.

Declarations

Conflicts of Interest The authors declare no conflict of interest.

References

1. Spaans VM, Trietsch MD, Crobach S, Stelloo E, Kremer D, Osse EM, Haar NT, van Eijk R, Muller S, van Wezel T, Trimbos JB, Bosse T, Smit VT, Fleuren GJ (2014) Designing a high-throughput somatic mutation profiling panel specifically for gynaecological cancers. Plos One 9:e93451. https://doi.org/10.1371/journal.pone.0093451
2. Yu XT, Zeng T, Li GJ (2015) Integrative enrichment analysis: a new computational method to detect dysregulated pathways in heterogeneous samples. BMC Genomics 16:918. https://doi.org/10.1186/s12864-015-2188-7
3. Zhang JH, Wu LY, Zhang SX, Zhang SH (2014) Discovery of co-occurring driver pathways in cancer. BMC Bioinformatics 15:1–14. https://doi.org/10.1186/1471-2105-15-271
4. Zhao JF, Zhang SH, Wu LY, Zhang XS (2012) Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics 28:2940. https://doi.org/10.1093/bioinformatics/bts564
5. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Gand Bignell, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K (2007) Patterns of somatic mutation in human cancer genomes. Nature 446:153–158. https://doi.org/10.1038/nature05610
6. Vandin F, Upfal E, Raphael BJ (2012) De novo discovery of mutated driver pathways in cancer. Genome Res 22:175–181. https://doi.org/10.1007/978-3-642-20036-6_44
7. Leiserson MDM, Blokh D, Sharan RJ, Raphael B (2013) Simultaneous identification of multiple driver pathways in cancer. PLOS Computational Biology 9:e1003054. https://doi.org/10.1371/journal.pcbi.1003054
8. Hou JP, Ma JB (2014) Dawnrank: discovering personalized driver genes in cancer. Genome Med 6:5. https://doi.org/10.1186/s13073-014-0056-8
9. Udager AM, Rolland DCM, McHugh JB, Betz BL, Murga-Zamalloa C, Carey TE, Marentette LJ, Hermsen MA, DuRoss KE, Lim MS, Elenitoba-Johnson KSJ, Brown NA (2015) High-frequency targetable EGFR mutations in sinonasal squamous cell carcinomas arising from inverted sinonasal papilloma. Cancer Res 75:2600–2606. https://doi.org/10.1158/0008-5472.can-15-0340
10. Gonzalez D, Martinez P, Wade R, Hockley S, Oscier D, Matutes E, Dearden CE, Richards SM, Catovsky D, Morgan GJ (2016) Mutational status of the tp53 gene as a predictor of response and survival in patients with chronic lymphocytic leukemia: results from the lrf cll4 trial. J Clin Oncol 29:2223–2229. https://doi.org/10.1200/JCO.2010.32.0838
11. Zhong SY, Zhou SL, Li AQ, Lv H, Li M, Tang SX, Xu XL, Shui RH, Yang WT (2021) High frequency of PIK3CA and TERT promoter mutations in fibromatosis-like spindle cell carcinomas. J Clin Pathol https://doi.org/10.1136/JCLINPATH-2020-207071
12. Srihari S, Ragan MA (2013) Systematic tracking of dysregulated modules identifies novel genes in cancer. Bioinformatics 29:1553–1561. https://doi.org/10.1093/bioinformatics/btt191
13. Wu H, Gao L, Dong JH, Yang XF (2014) Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks. Plos One 9:e91856. https://doi.org/10.1371/journal.pone.0091856
14. Wu H (2018) Algorithm for Detecting Driver Pathways in Cancer Based on Mutated Gene Networks. Chinese J Comput 41(1400–1414), https://doi.org/10.11897/SP.J.1016.2018.01400
15. Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A (2011) Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Med Genomics 4:34. https://doi.org/10.1186/1755-8794-4-34
16. Kim YA, Cho DY, Dao P, Przytycka TM (2015) Memcover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. Bioinformatics 31:284–292. https://doi.org/10.1093/bioinformatics/btv247
17. Leiserson MDM, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim YH, Niu BF, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng YW, Ryslik GA, Lopez-Bigas N, Getz G, Ding L, Raphael BJ (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 47:106–114. https://doi.org/10.1038/ng.3168
18. Reyna MA, Leiserson MDM, Raphael BJ (2018) Hierarchical hotnet: identifying hierarchies of altered subnetworks. Bioinformatics 34:972–980. https://doi.org/10.1093/bioinformatics/bty613
19. Rafsan A, Ilyes B, Cesim E, Evis H, Hilal K (2019) Mexcowalk: mutual exclusion and coverage based random walk to identify cancer modules. Bioinformatics 36:872–879. https://doi.org/10.1101/547653
20. Wu H, Gao L, Li F, Yang XF, Kasabov N (2015) Identifying overlapping mutated driver pathways by constructing gene networks in cancer. Bioinformatics 16:S3. https://doi.org/10.1186/1471-2105-16-S5-S3
21. Nepusz T, Yu H, Paccanaro A (2012) Detecting overlapping protein complexes in protein–protein interaction networks. Nat Methods 9:471–472. https://doi.org/10.1038/nmeth.1938
22. Guo MZ, Wang SM, Liu XY, Tian Z (2017) Algorithm for predicting the associations between MiRNAs and diseases. J Softw 28:3094–3102. https://doi.org/10.13328/j.cnki.jos.005351
23. Tang DM, Zhu QX, Yang F, Chen K (2011) Efficient cluster analysis method for protein sequences. J Softw 22:1827–1837. https://doi.org/10.3724/sp.j.1001.2011.03848
24. Hou YX, Duan L, Li L, Lu L, Tang CJ (2018) Search of genes with similar phenotype based on disease information network. J Softw 29(721–733):10.13328/j.cnki.jos.005445
25. Zhang Q, Li M, Deng Y (2016) A new structure entropy of complex networks based on tsallis nonextensive statistical mechanics. Int J Modern Phys C 27:440–450. https://doi.org/10.1142/S0129183116501187
26. Hofree M, Shen JP, Carte H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations. Nat Methods 10:1108–1115. https://doi.org/10.1038/nmeth.2651
27. Li F, Gao L, Wang B (2020) Detection of driver modules with rarely mutated genes in cancers. IEEE/ACM Trans Comput Biol Bioinform 17:390–401. https://doi.org/10.1109/TCBB.2018.2846262
28. Wang BQ, Wang M, Li XP, Yang M, Liu L (2020) Variations in the Wnt/β-Catenin Pathway Key Genes as Predictors of Cervical Cancer Susceptibility. Pharmacogenom Personalized Med 13:157–165. https://doi.org/10.2147/PGPM.S248548
29. Katherine S, Noriko U, Wiljan H, Michel LT, Bouchard M (2013) Inactivation of lar family phosphatase genes ptprs and ptprf causes craniofacial malformations resembling pierre-robin sequence. Development 140:3413–3422. https://doi.org/10.1242/dev.094532
30. Mana G, Clapero F, Panieri E, Panero V, Böttcher R, Tseng HY, Saltarin F, Astanina E, Wolanska K, Morgan M, Humphries M, Santoro M, Serini G, Valdembri D (2016) PFIA1 drives active $\alpha5\beta1$ integrin recycling and controls fibronectin fibrillogenesis

and vascular morphogenesis. Nat Commun 7:13546. https://doi.org/10.1038/ncomms13546
31. Li H, Liu L, Liu C, Zhuang J, Zhou C, Yang J, Gao C, Liu G, Lv Q, Sun C (2018) Deciphering Key Pharmacological Pathways of Qingdai Acting on Chronic Myeloid Leukemia Using a Network Pharmacology-Based Strategy. Med Sci Monit 24(5668–5688):10.12659/MSM.908756
32. Xia YK, Zeng YR, Zhang ML, Liu P, Liu F, Zhang H, He CX, Sun YP, Zhang JY, Zhang C, Song L, Ding C, Tang YJ, Yang Z, Yang C, Wang P, Guan KL, Xiong Y, Ye D (2020) Tumor-derived neomorphic mutations in asxl1 impairs the bap1-asxl1-foxk1/k2 transcription network. Protein & Cell. https://doi.org/10.1007/s13238-020-00754-2
33. Wang XW, Xi XQ, Wu J, Wan YY, Hui HX, Cao XF (2015) MicroRNA-206 attenuates tumor proliferation and migration involvingthe downregulation of NOTCH3 in colorectal cancer. Oncol Rep 33:1402–1410. https://doi.org/10.3892/or.2015.3731
34. Catarina R, Susana R, Claudia G, Domingos H (2010) Two notch ligands, dll1 and jag1, are differently restricted in their range of action to control neurogenesis in the mammalian spinal cord. Plos One 5:e15515. https://doi.org/10.1371/journal.pone.0015515
35. Amrich CG, Davis CP, Rogal WP, Shirra MK, Heroux A, Gardner RG, Arnd KM, VanDemark AP (2012) Cdc73 subunit of Paf1 complex contains C-terminal Ras-like domain that promotes association of Paf1 complex with chromatin. J Biol Chem 287:10863–75. https://doi.org/10.1074/jbc.M111.325647
36. Mueller CL, Jaehning JA (2002) Ctr9, rtf1, and leo1 are components of the paf1/rna polymerase ii complex. Mol Cel Biol 22:1971–1980. https://doi.org/10.1128/MCB.22.7.1971-1980.2002
37. Tsujino I, Nakanish Y, Shimizu T, Obana Y, Ohni S, Takahashi N, Nemoto N, Hashimoto S (2012) 999 Correlation Between Differences in the Increase in MAPK (ERK1/2) Activity Due to Driver Mutations and Prognosis in Non-small-cell Lung Cancer. Euro J Cancer 48:S241–S241. https://doi.org/10.1016/S0959-8049(12)71617-9
38. Schwickart M, Huang XD, Lill JR, Liu JF, Ferrando R, French DM, Maecker H, O'Rourke K, Bazan F, Eastham-Anderson J, Yue P, Dornan D, Huang DCS, Dixit VM (2010) Deubiquitinase usp9x stabilizes mcl1 and promotes tumour cell survival. Nature 463:103–107. https://doi.org/10.1038/nature08646
39. Sabò A, Kress TR, Pelizzola M, De PS, Gorski MM, Tesi A, Morelli MJ, Bora P, Doni M, Verrecchia A, Tonelli C, Fagà G, Bianchi V, Ronchi A, Low D, Müller H, Guccione E, Campaner S, Amati B (2014) Selective transcriptional regulation by myc in cellular growth control and lymphomagenesis. Nature 511:488–492. https://doi.org/10.1038/nature13537
40. Matumoto T, Chen Y, Contreras-Sanz A, Ikeda K, Schulz G, Gao J, Oo HZ, Roberts M, Costa JBD, Nykopp TK (2010) FBXW7 loss of function contributes to worse overall survival and is associated with accumulation of MYC in muscle invasive bladder cancer. Urol Oncol 38:904–905. https://doi.org/10.1016/j.urolonc.2020.10.048