PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2022,49(2):381~394
www.pibb.ac.cn

Identification of Gene Signatures Associated With Lung Adenocarcinoma Diagnosis and Prognosis Based on WGCNA and SVM-RFE Algorithm*

WANG Mei, WANG Ke-Xin, TAN Jian-Jun**, WANG Jing-Jing

(*Department of Biomedical Engineering, Faculty of Environment and Life, Beijing University of Technology, Beijing International Science and Technology Cooperation Base for Intelligent Physiological Measurement and Clinical Transformation, Beijing* 100124, China)

Abstract **Objective** Lung cancer is one of the most common cancers in the world. Lung adenocarcinoma (LUAD) has the highest annual mortality rate among lung cancer patients. It has been reported that changes in gene spectrum were associated with the process of tumorigenesis and its development. The purpose of this study is to identify the gene signatures associated with LUAD and to further analyze their prognostic significance. **Methods** Weighted gene co-expression network analysis (WGCNA), differential gene analysis, cox regression analysis, and protein-protein interaction (PPI) network analysis were used to screen the hub genes highly related to LUAD based on The Cancer Genome Atlas (TCGA) database. The RNA-seq data sets from TCGA and GTEx (Genotype Tissue Expression) database were combined and divided into a training set and a validation set, which were used to construct the diagnostic model by support vector machine recursive feature elimination feature (SVM-RFE) algorithm. GSE32863 and GSE31210 were used to verify the diagnostic accuracy of the model and the prognostic value of our obtained gene signatures, respectively. **Results** The results demonstrated that the model of 5 gene signatures (*anln*, *cenpa*, *plk1*, *tpx2*, *cdca3*) obtained by the SVM-RFE algorithm had an outstanding performance in the classification of LUAD patients. Functional enrichment analysis showed that these 5 gene signatures were highly related to the biological process of tumor initiation and progression. What's more, LUAD patients with high expression of these 5 genes also exerted a poor outcome in survival status. **Conclusion** Therefore, we could conclude that our study obtained useful models with 5 gene signatures for the diagnosis and prognosis of LUAD, which were essential for the development of novel targets applied in precision therapy.

Key words lung adenocarcinoma, gene signature, WGCNA, SVM-RFE
DOI: 10.16476/j.pibb.2021.0010

Lung cancer is the leading cause of death globally, and non-small cell lung cancer (NSCLC) accounts for 85% of all lung cancer[1]. Lung adenocarcinoma (LUAD) is the most common type in NSCLC[2]. The 5-year survival rate after the diagnosis of lung cancer is less than 20%[3]. Although there are recent advances in surgical methods, immunotherapy, and neoadjuvant therapy, the mortality of NSCLC remains high[4].

With the development of high-throughput sequence technology, bioinformatics has become increasingly popular in genomic analysis to investigate the pathological mechanism of tumor and discover tumor-specific biomarkers[5]. Bioinformatics has made it possible to identify gene expression changes during tumorigenesis, contributing to determining the prognosis and treatment of lung cancer[6]. What's more, after the HGP (Human Genome Project) finished, lots of publicly available databases such as The Cancer Gene Atlas (TCGA, https://tcga-data. nci. nih. gov/), Gene Expression Omnibus (GEO, http://www. ncbi. nlm. nih. gov/geo/) and Genotype Tissue Expression (GTEx, https://

* This work was supported by grants from Beijing Natural Science Foundation (2202002) and The National Natural Science Foundation of China (21173014).

** Corresponding author.
Tel: 86-10-67392001, E-mail: tanjianjun@bjut.edu.cn

Received: January 13, 2021; Accepted: May 14, 2021

commonfund.nih.gov/GTEx/) showed cancer genome sequencing data. A careful and thorough analysis of these data can identify gene signatures and signal pathways about tumor, which will help to explore the mechanism of tumor formation and development.

At present, there have been numerous studies on gene signatures, which are helpful for the selection of lung cancer treatment methods and the prediction of survival rate after lung cancer surgery. For example, Dama *et al*.[7] suggested that 10 gene signatures may be important prognostic indicators of patients with stage I LUAD. Liu *et al*.[8] proved that increased mRNA expression of TTK and NEK2 improved the risk of smoking-related LUAD death. Moreover, Xie *et al*.[9] identified prognostic signatures containing 6 genes significantly correlated with the overall survival (OS) of NSCLC patients, providing support for the construction of treatment regimens for patients. However, most of these studies considered genes as individual bioinformatics analysis factors and finally combined the screened genes to form a predictive model, which did not make full use of the relationship between genes.

Weighted gene co-expression network analysis (WGCNA) is a systematic biological method for selecting co-expression modules of related genes and the critical module associated with clinical traits[10], providing a new direction to predict the gene signatures. Furthermore, the support vector machine recursive feature elimination feature (SVM-RFE) is a powerful algorithm to establish a gene model based on a large number of sample data. It improves the accuracy of the classifier model and has a wide range of applications in the field of bioinformatics[11-12]. Therefore, an idea of combining WGCNA and SVM-RFE to improve the recognition ability of highly related genes turned out, which was used to establish a cancer diagnosis model and screen candidate gene signatures.

In our study, in order to screen gene signatures related to the diagnosis and prognosis of LUAD, we used WGCNA, SVM-RFE, survival analysis, and other bioinformatics analysis methods to identify and verify the gene signatures highly related to LUAD based on multiple bioinformatics databases. Finally, our study obtained useful models with 5 gene signatures that applied to different sets for the diagnosis and prognosis and provided some valuable insights for the development of novel targets involved in the precision therapy of LUAD.

1 Materials and methods

1.1 Data download and processing

The transcriptome profiling dataset of LUAD and corresponding clinical dataset were obtained from TCGA database, including 344 tumor samples and 38 normal samples, and RNA-seq count data had about 19 430 genes. The mRNA expression of each sample was merged into a matrix with a merge script in the Perl language. Then the matrix of mRNA expression was annotated with the Ensembl database. With the help of the “edgeR” and “DESeq2” package, genes with low read counts were usually not of interest for further differential gene screening. According to the standard that the average expression of a gene in each sample should be ⩾ 1, the mRNA expression data were filtered, a total of 18 127 gene expression data continued to be analyzed. This study conformed to the publication guidelines of TCGA database.

Meanwhile, to improve the accuracy of the diagnostic model constructed by the SVM-RFE algorithm, data of 288 normal LUAD samples were downloaded from the GTEx database and gene count data on 56 754 genes. As the data used herein were freely sourced, approval from the Ethics Committee was not required. However, since only the data from normal samples were collected by GTEx database, it is usually used for bioinformatics analysis combined with TCGA database. Because the types of TCGA data and GTEx data were gene counts and reads per kilobase per million mapped reads (RPKM), respectively. According to the calculation Equation (1), the gene counts of TCGA are converted into RPKM, and then the data sets of TCGA and GTEx are standardized and merged by the Z-socre method to facilitate the further construction of the model. The combination of TCGA and GTEx dataset were divided into a training set (60%) and an interval validation set (40%), including 325 normal samples and 307 tumor samples.

$$FPKM = \frac{total\ exon\ reads}{mapped\ reads \times exon\ length} \quad (1)$$

FPKM standardized total exon reads based on two aspects: mapped reads and exon length.

For further verification, we downloaded LUAD gene expression data from two access sets, GSE32863 and GSE31210, from GEO database. The former one

provided 58 cancer samples and 58 normal samples for verifying the accuracy of the diagnostic model and the latter one provided 226 cancer samples with clinical information which helped to explore the relationship between the gene signatures and clinical prognosis.

1.2 Identification of key gene co-expression module

In order to identify gene modules highly related to tumor patients, a weighted gene co-expression network was constructed by using the "WGCNA" package[13] in R software. WGCNA generated co-expression modules of highly correlated genes based on the interaction between genes and screened the critical module that was highly related to the clinical feature, providing a new insight for gene target prediction in precision medicine[14-16]. According to the variance of gene expression, the first 8 000 genes were selected to construct a co-expression network. After constructing a sample tree based on gene expression, the goodSamplesGenes function was used to delete samples with large outliers. According to the co-expression relationship of genes, Pearson correlation coefficients were calculated between each gene, and their absolute values were used to establish the gene adjacency matrix, the formula is Equation (2). In order to make the distribution of genes conform to the scale-free network base on gene connectivity, the best soft-threshold power value β was chosen to construct the proximity matrix and transformed into a topological overlap matrix (TOM) in Equation (3). According to the TOM of genes, Equation (4) was used to calculate the distance between genes for hierarchical clustering. The network modules were generated using the dynamic shear method and distinguished by colors, with the best soft-threshold power value and a module size cut-off criterion ⩾ 10. Then, the obtained modules were used to screen the key module most relevant to the clinical traits by Pearson correlation test.

$$a_{ij} = |cor(x_i, x_j)^{\beta}| \quad (2)$$

$$TOM_{ij} = \frac{l_{ij} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}} \quad (3)$$

$$D_{ij} = 1 - TOM_{ij} \quad (4)$$

x_i and x_j are the nodes in the scale-free network; β: soft threshold; a_{ij}: gene adjacency matrix; k: the sum of the adjacency coefficient of all nodes connected individually between genes; l_{ij}: the sum of the product of the adjacency coefficients between gene i and j; D_{ij}: the hierarchical clustering distance between gene i and j.

1.3 Analysis of differential expressed genes（DEGs） and the intersection with the key co-expression module

Due to differential analysis of the macro-genome, adj.P was selected to reduce the error rate of differential gene screening. Two DEGs sets were generated by "edgeR" and "DESeq2" packages after normalization and data filter with thresholds$|\log_2$ fold-change$| \geqslant 2$ and adj. $P<0.01$ by comparing LUAD group with normal group, respectively. The two DEGs sets of LUAD group were visualized as volcano plots by using "gplots" package. Then, the overlapped genes obtained by crossing the two DEGs sets and the gene modules with high clinical relevance screened by WGCNA were used to identify potential gene signatures, which were shown as the Venn diagram by "VennDiagram" package.

1.4 Univariate cox regression analysis

The purpose was to independently assess the impact of overlapping genes on the survival time of LUAD patients. The "survival" package was used to analyze the prognostic value of each gene through univariate cox regression[17]. According to the cutoff criterion of $P<0.05$ in TCGA set, genes that were significantly related to the survival time of the patients were considered to have prognostic value. Then, the independent prognostic genes were selected to further analysis.

1.5 Construction of protein-protein interaction（PPI） network

According to the results of the univariate cox regression analysis, we used the Search tool for the Retrieval of Interacting Genes (STRING) database to build PPI networks and then selected the genes that played key roles in the gene network for the further network model analysis. The threshold for minimum interaction score of these genes was 0.7 and built a PPI network model visualized by Cytoscape (version 3.6.1; https://www. cytoscape. org/). According to reports, Maximal Clique Centrality (MCC) algorithm was the most effective way to find hub genes in PPI networks[18]. The MCC score of each node in the PPI network was calculated by using CytoHubba plugin. In our study, according to the results of MCC algorithm analysis, the top 15 genes were considered

as the hub genes that played important roles in the process of tumor formation.

1.6 Functional enrichment analysis

To investigate the biological function of the selected genes, the “clusterProfiler” package in R software[19] was used to analyze the Gene Ontology (GO) annotation in the overlapping genes and hub genes. GO annotation that includes the three aspects: molecular function (MF), cellular component (CC), and biological process (BP), can describe the molecular functions that gene products may perform, the cellular environment, and the biological processes involved[20].

1.7 Gene signatures selection base on SVM−RFE

The “e1071”[21] and “caret” packages[22] in R software were used to select optimized gene signatures based on recursive feature elimination (RFE) algorithm. Furthermore, the expression of genes was considered as the feature, the clinical traits of the sample were considered as the categorical variable, and the support vector machine (SVM) used the linear kernel to predict patients with LUAD to obtain the optimal gene signatures by RFE algorithm. The hub genes that passed the MCC algorithm analysis entered the SVM-RFE algorithm analysis. In the SVM-RFE algorithm, genes were ranked according to the measure of their importance, those genes with lower rankings were removed. And the precision of the gene model was examined by 10-fold cross-validation in the training set. Then, the optimal gene model was examined in the internal validation set and external validation set (GSE32863). Besides, the performance of the classifier in these sets was evaluated by receiver operating characteristic (ROC) curve analysis. The genes selected by the SVM-RFE method were chosen as the gene signatures.

1.8 Constructions and verification of prognostic prediction model

To confirm the prognostic value of gene signatures, the multivariate cox regression model was constructed with the gene signatures as variables based on GSE31210 set. Then, based on the expression of gene signatures and the regression coefficient estimated by the multivariate cox regression model, the risk score (RS) prognostic model was constructed as follows[23]:

$$RS = \sum \beta_{\mathrm{mRNA}} \times Exp_{\mathrm{mRNA}} \quad (5)$$

β_{mRNA} was defined as the independent prognostic coefficient and Exp_{mRNA} represented the expression of corresponding mRNA.

According to the median RS as the cut-off point, all patients in GSE31210 set were distributed to low-risk and high-risk groups. In order to evaluate the survival time difference between low-risk and high-risk groups and verify the prognostic value of RS model, the “survival” (version 3.27) package in R software was used to perform Kaplan-Meier (K-M) survival curve analysis[24].

2 Results

2.1 Weighted gene co−expression network construction

This study was conducted as indicated in Figure 1. The mRNA expression matrix of TCGA-LUAD was obtained (18 127 genes) after data preprocessing. According to the requirements of WGCNA algorithm, we selected the top 8 000 genes in the variance order among genes. Then, in order to ensure the reliability of co-expression network, the outliers of the samples

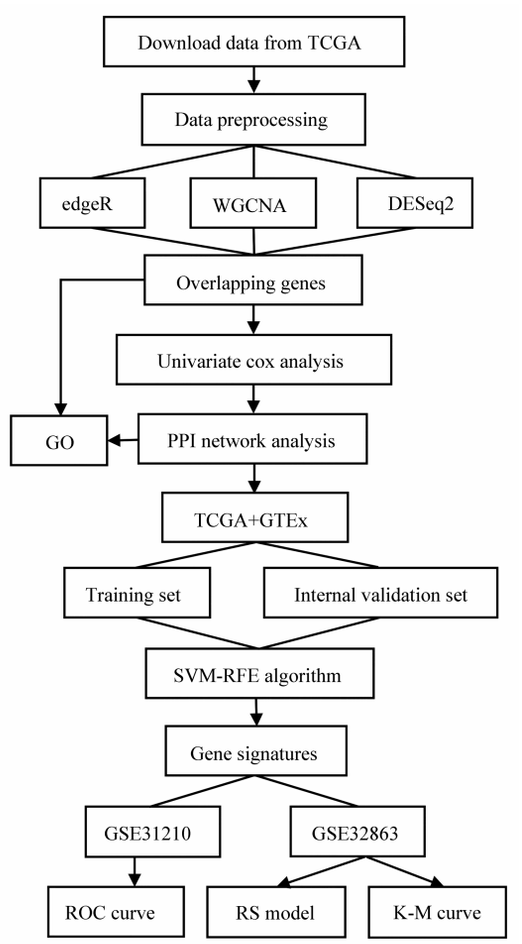

Fig. 1 The workflow of this study

were removed by the method of sample clustering (Figure 2a). The scale-free distribution was shown in Figure 2b, and the appropriate soft-threshold power value β=8 was chosen base on the fitting degree and connectivity of the network. Finally, 11 modules were identified base on dynamic tree clipping and average hierarchical clustering (Figure 2c). In addition, we plotted the heatmap of module-trait relationships to evaluate the association between modules and two clinical traits. The result of the heatmap revealed that the turquoise module (r=0.61, P=2E−35) was found to be the highest associated with tumor tissues (Figure 2d).

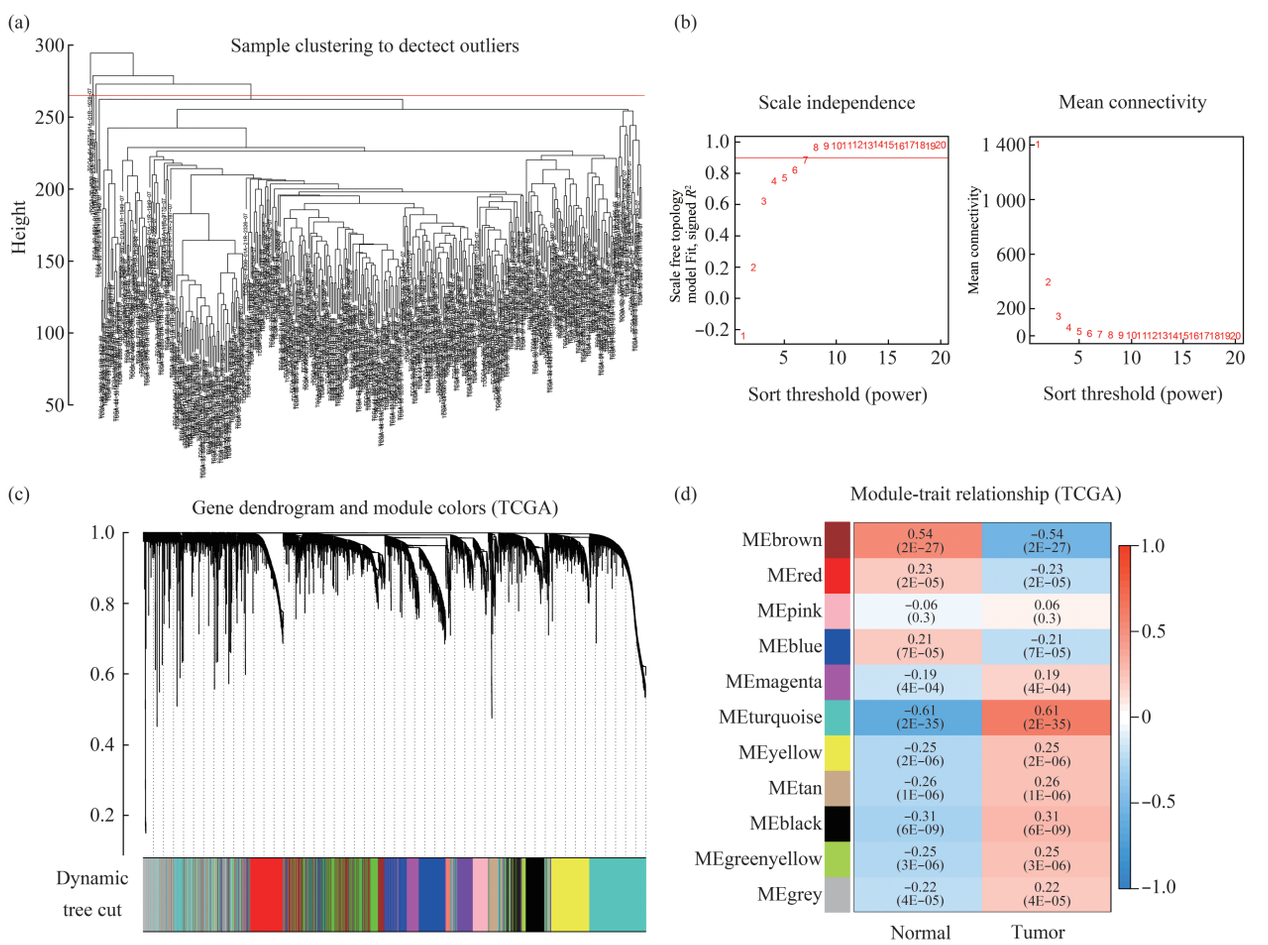

Fig. 2　Construction of weighted gene co−expression network for LUAD

(a) Samples clustering and removal of outliers. (b) Network topology analysis of various soft-threshold power. The left figure shows the scale-free fit index for various soft-threshold powers, signed R^2 (y axis), and the soft threshold power (x axis). Soft-threshold power value β=8 was chosen. The right figure shows that the mean connectivity for various soft-threshold powers, y axis is a decreasing function of the soft-threshold power β (x axis). (c) The cluster dendrogram of the 8 000 genes was ordered by a hierarchical clustering of genes based on the value of dissimilarity (1−*TOM*). Each branch in the figure represents one gene, and each module was assigned different colors. (d) Identification of modules associated with the clinical traits of LUAD. Each module contains the corresponding correlation and *P*. The correlation coefficient represents the correlation between the gene module and clinical characteristics.

2.2　Identification of genes in the DEGs and co−expression modules

Based on the cut-off criteria of $|\log_2$ fold-change$| \geqslant 2$ and adj. P<0.01, a total of 1 137 DEGs (Figure 3a) and 1 011 DEGs (Figure 3b) were found to be dysregulated in tumor samples by the “edgeR” and “DESeq2” packages, respectively. As shown in Figure 2c, 1 799 co-expression genes were found in the turquoise module of TCGA data set. According to the intersection of these three sets, a total of 295 overlapping genes were extracted for gene signature screening (Figure 3c).

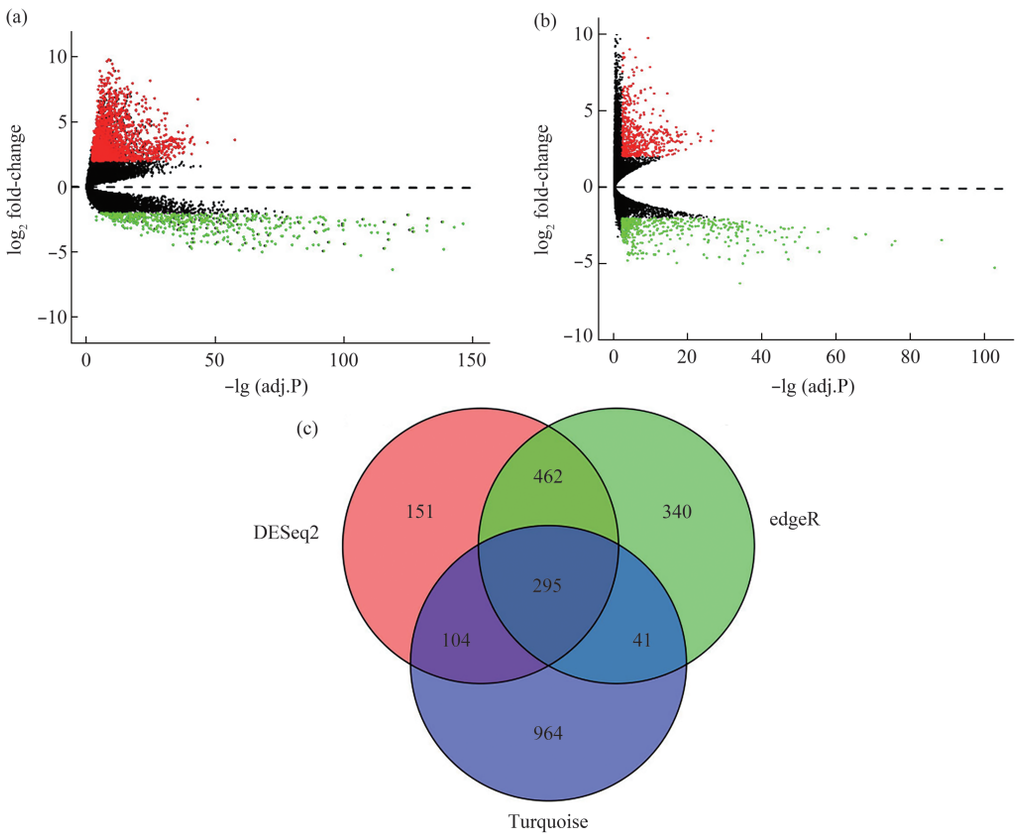

Fig. 3　Screening of overlapping genes

(a, b) Identification of DEGs using "edgeR" and "DESeq2" packages respectively, with the cut-off criteria of $|\log_2 \text{fold-change}| \geqslant 2.0$ and padj (adj. P) < 0.01. (a) Volcano map of DEGs using "edgeR" package. (b) Volcano map of DEGs using "DESeq2" package. Red represents up-regulated genes, and green represents down-regulated genes in the two volcano maps. (c) The Venn diagram of genes among two DEGs sets and co-expression module gene set. Totally, 295 overlapping genes were selected in the intersection of the two DEGs sets and the turquoise module genes set.

2.3　Univariate cox regression and PPI network analysis identifying gene signatures

Univariate cox regression analyses were performed on 306 LUAD patients to evaluate the prognostic relation between the selected 295 genes and OS. A total of 39 genes were obtained with a cut-off criterion of $P<0.05$, which was considered to be significantly associated with OS in LUAD patients. Then, PPI network among the selected genes was constructed by using STRING database (Figure 4a). The cut-off value of the interaction score was 0.7. To further analyze the relationship of the 39 genes, we imported PPI network into Cytoscape. HR (hazard ratio), the co-expression relationship among the selected genes was calculated using the gene expression levels by MCC algorithms in the Cytoscape (Figure 4b) and the result was shown in Table 1. As shown in Figure 3b, PPI network of the top 15 highest-scored genes have 19 nodes and 102 edges, including their expanded sub-network. According to the MCC scores, the top 15 highest-scored genes were considered as the hub genes for further bioinformatics analysis.

2.4　Functional enrichment analysis

To further analyze the potential biological function of the 295 overlapping genes and hub genes, the functional enrichment analysis was performed by the "clusterProfiler" and "enrichplot" packages in R software. As shown in Figure 5a, several enriched gene sets were obtained by screening from GO enrichment analysis. Biological processes (BP) of the 295 genes were mainly involved in nuclear division and organelle fission. Due to the result of cellular component (CC), these genes were mainly enriched in

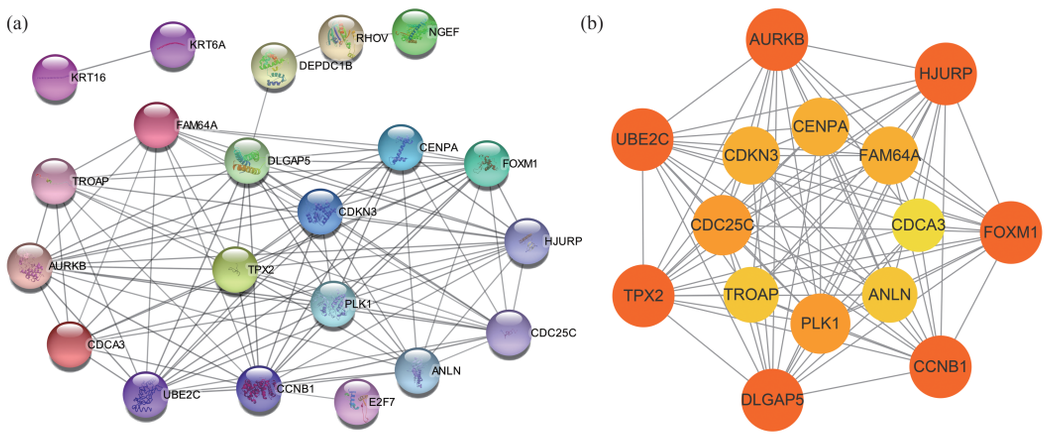

Fig. 4　Screening of the hub genes

(a) PPI network between the selected genes. Edges represent the protein-protein associations. The different colors of the nodes represent different clusters (n≥3). To facilitate observation, different gene clusters connect with different types of lines. (b) Identification of the hub genes from PPI network using MCC algorithm. The nodes from red to green represent the genes with high to low MCC score. Edges represent the protein-protein associations.

Table 1　The top 15 highest-scored genes in PPI network

Gene ID	MCC score	HR	*P* value
ccnb1	4.50E+07	1.17	4.77E-02
dlgap5	4.50E+07	1.16	2.61E-02
tpx2	4.50E+07	1.13	4.53E-02
ube2c	4.50E+07	1.17	1.03E-02
aurkb	4.50E+07	1.15	4.34E-02
hjurp	4.50E+07	1.18	1.70E-02
foxm1	4.50E+07	1.15	2.49E-02
cdc25c	4.43E+07	1.17	3.63E-02
cenpa	4.35E+07	1.13	4.82E-02
cdkn3	4.35E+07	1.18	1.59E-02
plk1	4.06E+07	1.17	2.28E-02
Anln	3.99E+07	1.15	3.44E-02
fam64a	4.35E+06	1.15	3.89E-02
troap	1.45E+06	1.14	4.36E-02
cdca3	7.26E+05	1.20	1.61E-02

the chromosomal region and spindle. Moreover, through the molecular function (MF) analysis, ATPase activity, tubulin binding, and microtubule binding were suggested to be related to these 295 genes. In addition, the result of gene enrichment analysis of the hub genes is shown in Figure 5b. These hub genes were mainly enriched in regulation of mitotic cell cycle phase transition, regulation of cell cycle phase transition, regulation of mitotic nuclear division, mitotic nuclear division, regulation of nuclear division, etc.

2.5　Gene signature selection

The recursive feature elimination (RFE) algorithm was used to identify the most significant gene signatures. In order to solve the class skew caused by the imbalance between normal and tumor samples in TCGA data set, data of 288 normal lung tissue samples were downloaded from GTEx database for SVM-RFE algorithm analysis. Then the RFE algorithm was applied to filter the 15 hub genes in order to identify the optimal combination of gene signatures in the training set. Finally, 5 genes (*anln*, *cdca3*, *cenpa*, *plk1*, *tpx2*) were selected as the gene signatures. Figure 6a shows the optimization process of RFE algorithm. When the number of gene signatures is 5, the accuracy of SVM-RFE model is the highest. Consequently, we constructed a diagnostic gene model based on these 5 genes.

The constructed SVM classifier with 5 gene signatures was applied to the training set (n=391), internal validation set (n=241), and external validation set GSE32863 (n=116). As shown in Figure 6b-d, the classifier could successfully differentiate normal samples and tumor samples in the 3 sets (AUC, area under curve; PPV, positive predictive value; NPV, negative predictive value). The training set generated

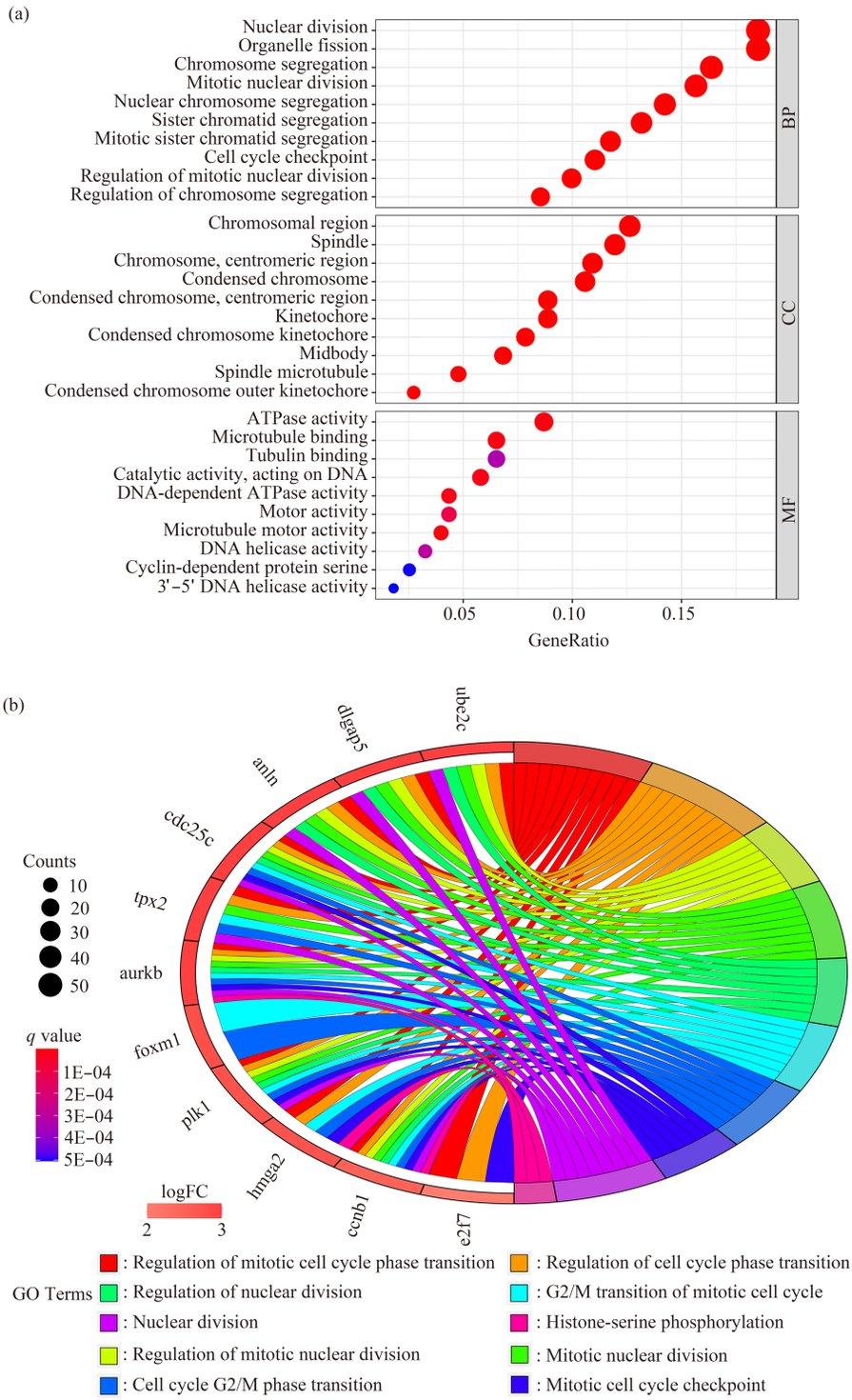

Fig. 5　GO enrichment analysis

(a) GO enrichment analysis of 295 overlapping genes with *q* value (adj.*P*) < 0.01. The color represents *q* value, and the size of the spots represents the gene number. (b) A circular plot of hub genes GO enrichment analysis. The left half of the circle is gene ID, and each line connects to a GO term in the right half of the circle.

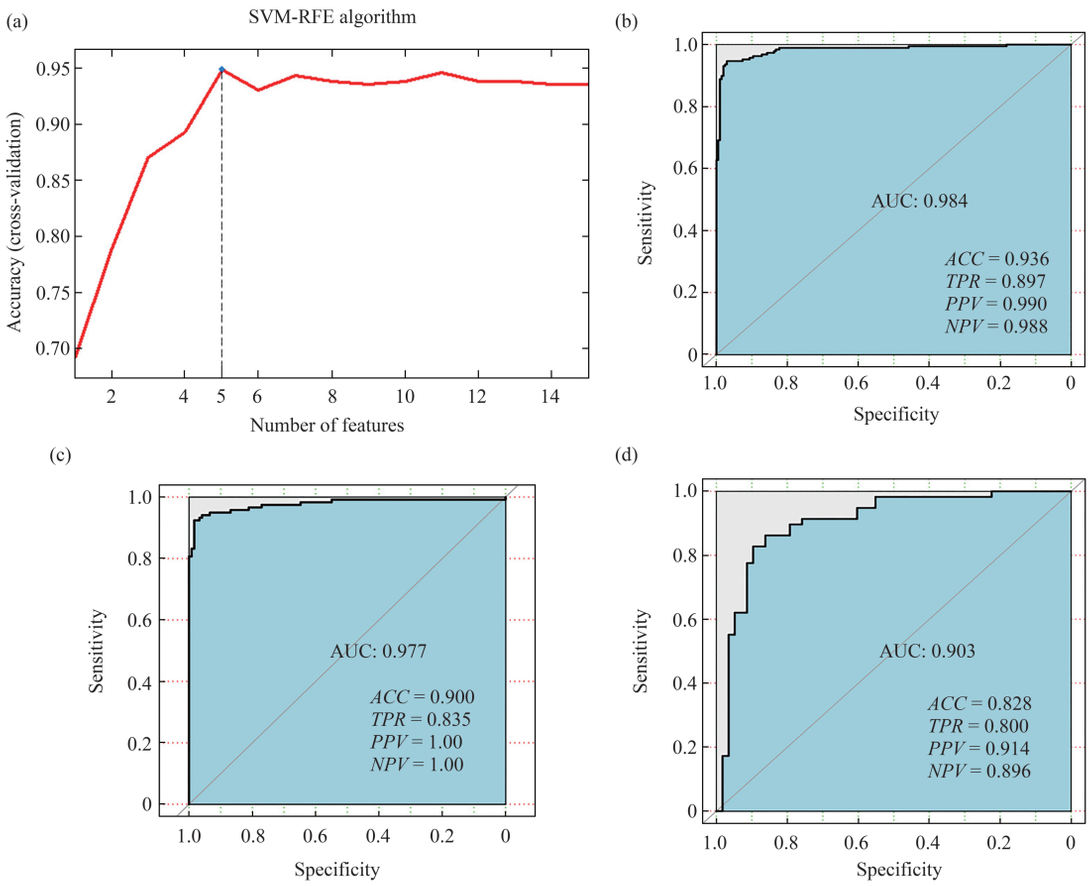

Fig. 6　Construction and validation of the diagnostic model for LUAD

(a) Accuracy curve of the optimized hub genes using RFE algorithm. ROC curves based on the SVM classifier in the training set (b), internal validation set (c), and external validation set GSE32863 (d).

AUC of 0.984, Accuracy of 0.936, and Recall of 0.897, the internal validation set generated AUC of 0.977, Accuracy of 0.900, and Recall of 0.835, and the external validation set GSE32863 generated AUC of 0.903, Accuracy of 0.828, and Recall of 0.800 (Table 2). These results illustrate that the classification model based on the 5 genes could accurately predict the LUAD patients (Figure 6).

Table 2　Effectiveness evaluation of the classifier of 5 gene signatures on the three sets

Set	AUC	Accuracy	Recall	PPV	NPV
Training set	0.984	0.936	0.897	0.990	0.988
Internal validation set	0.977	0.900	0.835	1.000	1.000
GSE32863	0.903	0.828	0.800	0.914	0.896

2.6　Risk score survival model of 5 gene signatures

According to the result of SVM-RFE algorithm analysis, 5 genes (*anln*, *cdca3*, *cenpa*, *plk1*, *tpx2*) were obtained to explore their prognostic value among the patients of the GSE31210 set. The risk score (RS) model was constructed based on the coefficients (Table 3) of the 5 gene signatures and their expression levels in the GSE31210. The RS formula was shown in Equation (6). The concordance index of this model was 0.69, and P=1.717 4E－02 (Figure 7a).

Table 3　Cox regression model of the 5 gene signatures in the GSE31210 set

Gene ID	Coefficient	HR	HR.95L	HR.95L	*P* value
cdca3	0.436	1.55	0.808	2.96	0.188
plk1	0.105	1.11	0.635	1.94	0.714
tpx2	0.044	1.05	0.437	2.50	0.921
anln	0.376	1.46	0.824	2.57	0.195
cenpa	−0.331	0.71	0.327	1.56	0.409

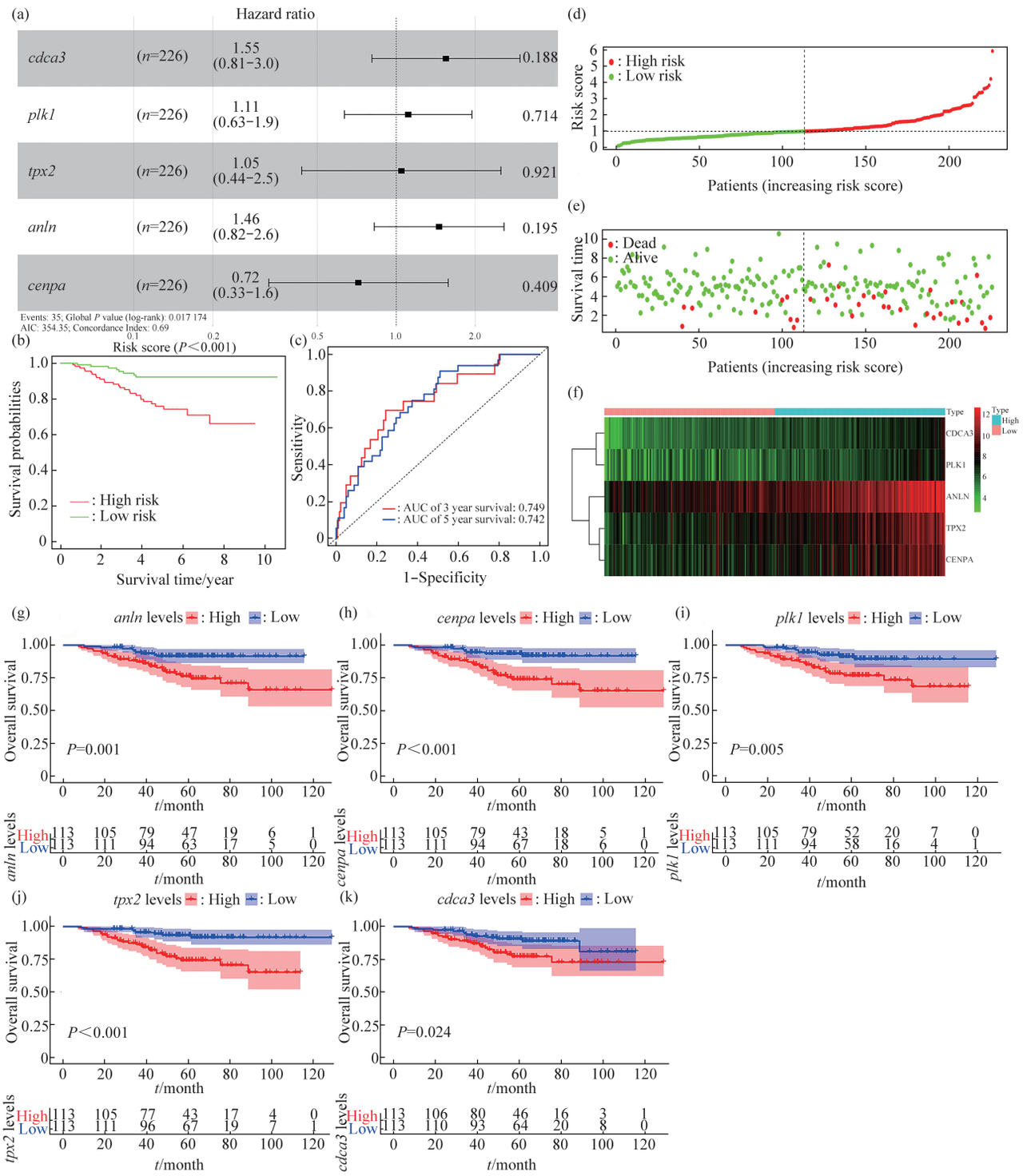

Fig. 7　Prognostic value of the 5 gene signatures in the GSE31210 set

(a) The Hazard ratio, *P*, and risk score distribution of constituents involved in multivariate cox regression and some RS model parameters. (b) Kaplan-Meier curves of the 5 gene signatures for high-risk and low-risk groups. (c) Time-dependent ROC analysis of the 5 gene signatures. (d−f) The risk score distribution of patients, the survival status of patients in the high-risk and low-risk groups, and a heatmap of the 5 genes expression in patients. (g−k) The expression level of the 5 genes was significantly associated with prognosis (P<0.05) in the GSE31210 set. The expression of the 5 genes selected increased, and the overall survival time was significantly reduced.

$$RS = \left(0.436 \times Exp_{cdca3}\right) + \left(0.105 \times Exp_{plk1}\right) + \left(0.044 \times Exp_{tpx2}\right) + (0.376 \times Exp_{anln}) + (-0.331 \times Exp_{cenpa}) \quad (6)$$

The expression of *cdca3* marked as Exp_{cdca3} and it is same to others.

The risk score of each sample was calculated for each LUAD patient, and all 226 patients were divided into two groups containing low-risk and high-risk groups. As shown in Figure 7b, compared with the high-risk group, patients with low-risk scores are demonstrated to have a greater chance of having the same survival time in the GSE31210 set. The AUC was 0.749 and 0.742 for the 3-year and 5-year OS by ROC analysis, respectively (Figure 7c), which indicated the good performance of RS model in survival prediction. Notably, the risk curves, living status of LUAD patients, and the 5 gene expression value associated with the risk score in the GSE31210 set were shown in Figure 7d−f. The mortality of high-risk group was significantly higher than that of low-risk group. Besides, we put the 5 genes into independent survival curve analysis, and the result showed that the expression of *anln*, *cdca3*, *cenpa*, *plk1*, and *tpx2* had a significant impact on the survival time of patients and displayed good prognostic significance (Figure 7g−k).

3 Discussion

NSCLC is the most common type of lung cancer. LUAD is the most common subtype of NSCLC with an extremely high mortality rate[25]. In recent years, the increasing studies of high-tech sequencing which illustrated the importance of gene signatures on determining cancer formation and outcomes provided a novel insight to integrate this bioinformatics into the therapeutic schedule[26-27]. It is necessary for the diagnosis and treatment of LUAD to reveal the molecular mechanism of LUAD and screen out the gene signatures based on the genome information. In this study, 295 overlapping genes were selected by comprehensively differential gene analysis and WGCNA analysis based on TCGA-LUAD database. As shown in functional annotation analysis, these genes were mainly enriched in the nuclear division, organelle fission, and chromosomal region, essential for tumorigenesis. According to the survival status of LUAD patients, 39 prognostic genes were screened by univariate cox regression, and the top 15 hub genes were selected out from the prognostic genes through the MCC algorithm with the help of CytoHubba plugin in Cytoscape. The classification model was constructed based on the 15 hub genes using the SVM-RFE algorithm for choosing the optimal gene signatures in the training set. Due to the results of the model accuracy, a classifier model with 5 gene signatures (*anln*, *cdca3*, *cenpa*, *plk1*, *tpx2*) was obtained and performed well in classifying LUAD samples in the training set, internal validation set, and external validation set GSE32863 through calculating AUC, Accuracy, Recall, PPV and NPV. Moreover, the RS model was constructed to validate the prognostic value of the 5 gene signatures. By applying these signatures to construct the RS model in the GSE31210 set, there were significant differences between the high-risk and low-risk groups, and the high-risk group has a lower survival rate than the low-risk group. Also, we found that these 5 genes had a high prognostic value in LUAD through the survival analysis of each gene signature.

The anillin actin binding protein (*anln*) gene is located on chromosome 7p14.2 and encodes a protein composed of 1 124 amino acids that contain four domains, including a RhoA-binding domain, a C-terminal pleckstrin homology domain, and myosin- and actin-binding domain[28]. *Anln* plays an important role in the process of cell cycle in the assembly of actin and myosin contractile rings in separates daughter cells[29]. Moreover, anillin is a substrate for the anaphase-promoting complex/cyclosome (APC/C), a ubiquitin ligase that controls the mitotic progression[30]. In addition, the inheritance and mitotic proliferation of defective genome can cause pathological conditions including a variety of cancers[31]. Aniline, a cell cycle regulator, has been proved to play a key role in tumor invasion[32-33]. In our study, compared with normal samples, the expression of *anln* is up-regulated in tumor tissues, which is significantly correlated with LUAD. Mechanism of *anln* function showed that active cell division in tumor tissue results in higher levels of *anln*, which is consistent with our findings.

Cenpa (centromere protein-A), a centromere-specific 17-ku protein, is a unique histone H3, likes the protein found in the active centromeres, relates to the major epigenetic tension of centromeric identity[34]. *Cenpa* plays an important role in cell cycle

regulation and cell survival[35]. According to previous reports, when *cenpa* and *cenpb* were knocked out at the same time, the inhibitory effect on cell proliferation was more significant[36]. A recent study showed that high expression of *cenpa* was closely associated with LUAD tumorigenesis proved by real-time polymerase chain reaction (RT-PCR) and Western blotting analysis[37]. Therefore, inhibitors targeting *cenpa* may be a promising anticancer strategy.

Plk1 (polo-like kinase 1), a member of the mitotic serine/threonine kinase family, is closely related to spindle formation and chromosome segregation during mitosis[38]. Previous studies demonstrated that *plk1* mRNA expression was elevated in proliferating cells including tumors of different origins and various cancer cell lines[39]. Wang *et al*. [40] revealed that the overexpression of plk1 protein was an independent prognostic biomarker in NSCLC patients. In the present study, we further demonstrate that LUAD patients who express a high level of plk1 protein have low overall survival, and the high expression of plk1 protein is a prognostic factor validated by the RS model and survival analysis.

The targeting protein for Xenopus kinesin-like protein 2 (*tpx2*), which is required for targeting Aurora-A kinase to the spindle apparatus, had been reported as gene signatures for human lung cancer prognosis *in vitro* lung carcinogenesis system[41-42]. Li *et al*. [43] demonstrated that *tpx2* was a potential candidate targeted for amplification and over-expression in NSCLC. In our study, *tpx2* is mainly plays a role in the regulation of mitotic cell cycle phase transition and mitotic nuclear division. The survival time of samples with a high *tpx2* gene expression level is significantly lower than that with a low *tpx2* gene expression level, which is consistent with previous studies.

Cell division cycle-associated protein-3 (*cdca3*), is required for mitosis entry as a part of the SKP1-Cullin RING-F-box (SCF) ubiquitin ligase complex to degrade the endogenous cell cycle inhibitor Wee1[44]. Some studies showed that unregulated *cdca3* was associated with the carcinogenic process and malignant patterns of certain tumors[45-46], but the relationship between the gene and formation of LUAD has not been found. *Cdca3* may be a new potential target for NSCLC to inhibit tumor growth and promote tumor aging, which may play an important role in tumor cell proliferation.

In conclusion, we obtained 5 gene signatures of LUAD by bioinformatics and machine learning analysis methods. The diagnostic and prognostic models constructed by the 5 gene signatures could had an outstanding performance in predicting and prognostic among different sets. PPI network analysis and GO analysis confirmed that these genes were positively related to the tumorigenesis and development of LUAD. The combined application of multiple sets provided more robust supports for our research. Therefore, our study may provide new insight into the diagnosis and treatment of LUAD.

References

[1] Li S, Zhu R, Li D, *et al*. Prognostic factors of oligometastatic non-small cell lung cancer: a meta-analysis. J Thorac Dis, 2018, **10**(6): 3701-3713

[2] Barlesi F, Mazieres J, Merlio J, *et al*. Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT). Lancet, 2016, **387**(10026): 1415-1426

[3] Nanavaty P, Alvarez M S, Alberts W M. Lung cancer screening: advantages, controversies, and applications. Cancer Control, 2014, **21**(1): 9-14

[4] Zhang H, Shen J, Yi L, *et al*. Efficacy and safety of ipilimumab plus chemotherapy for advanced lung cancer: a systematic review and meta-analysis. J Cancer, 2018, **9**(23): 4556-4567

[5] Ayyildiz D, Piazza S. Introduction to bioinformatics. Methods Mol Biol, 2019, **1986**: 1-15

[6] Parikh A R. Lung cancer genomics. Acta Med Acad, 2019, **48**(1): 78-83

[7] Dama E, Melocchi V, Dezi F, *et al*. An aggressive subtype of stage I lung adenocarcinoma with molecular and prognostic characteristics typical of advanced lung cancers. Clin Cancer Res, 2017, **23**(1): 62-72

[8] Liu Y, Ni R, Zhang H, *et al*. Identification of feature genes for smoking-related lung adenocarcinoma based on gene expression profile data. Oncotargets Ther, 2016, **9**: 7397-7407

[9] Xie H, Xie C. A six-gene signature predicts survival of adenocarcinoma type of non-small-cell lung cancer patients: a comprehensive study based on integrated analysis and weighted gene coexpression network. Biomed Res Int, 2019, **2019**(4250613): 1-16

[10] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol, 2005, **4**(1): 1-45

[11] Tang Y, Zhang Y Q, Huang Z. Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Trans Comput Biol Bioinform, 2007, **4**(3): 365-381

[12] Su R, Zhang J, Liu X, *et al*. Identification of expression signatures for non-small-cell lung carcinoma subtype classification. Bioinformatics, 2020, **36**(2): 339-346

[13] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 2008, **9**: 599

[14] Fan Z, Xue W, Li L, *et al*. Identification of an early diagnostic biomarker of lung adenocarcinoma based on co-expression similarity and construction of a diagnostic model. J Transl Med, 2018, **16**(1): 205

[15] Zhang L, Tan J, Han D, *et al*. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov Today, 2017, **22**(11): 1680-1685

[16] Ding T, Zhang R, Zhang H, *et al*. Identification of gene co-expression networks and key genes regulating flavonoid accumulation in apple (Malus × domestica) fruit skin. Plant Sci, 2021, **304**: 110747

[17] Mermut O, Inanc B. Prognostic factors and survival of elder women with breast cancer aged >= 70 years. Turk J Geriatr, 2019, **22**(4): 426-433

[18] Chin C, Chen S, Wu H, *et al*. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol, 2014, **8**(Suppl 4): S11

[19] Yu G, Wang L, Han Y, *et al*. clusterprofiler: an r package for comparing biological themes among gene clusters. OMICS, 2012, **16**(5): 284-287

[20] Harris M A, Clark J I, Ireland A, *et al*. The Gene Ontology (GO) project in 2006. Nucleic Acids Res, 2006, **34**(SI): D322-D326

[21] Becker N, Werft W, Toedt G, *et al*. penalizedSVM: a R-package for feature selection SVM classification. Bioinformatics, 2009, **25**(13): 1711-1712

[22] Kuhn M. Building predictive models in r using the caret package. J Stat Softw, 2008, **28**(5): 1-26

[23] Gong C, Tan W, Chen K, *et al*. Prognostic value of a BCSC-associated microRNA signature in hormone receptor-positive HER2-negative breast cancer. Ebiomedicine, 2016, **11**: 199-209

[24] Zhao Z, He B, Cai Q, *et al*. A model of twenty-three metabolic-related genes predicting overall survival for lung adenocarcinoma. PeerJ, 2020, **8**: e10008

[25] Bray F, Ferlay J, Soerjomataram I, *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, 2018, **68**(6): 394-424

[26] Zhao Q P, Xiong T, Xu X J, *et al*. *De Novo* transcriptome analysis of Oncomelania hupensis after molluscicide treatment by next-generation sequencing: implications for biology and future snail interventions. PLoS One, 2015, **10**(3): e118673

[27] 门婧睿,谭建军,孙洪亮 . 肝癌预后 miRNA 风险评分模型的鉴定和分析 . 生物化学与生物物理进展 . 2020, **47**(4):344-360
Men J R, Tan J J, Sun H L. Prog Biochem Biophys, 2020, **47**(4): 344-360

[28] Strausberg R L, Feingold E A, Grouse L H, *et al*. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci USA, 2002, **99**(26): 16899-16903

[29] Straight A F, Field C M, Mitchison T J. Anillin binds nonmuscle myosin II and regulates the contractile ring. Mol Biol Cell, 2005, **16**(1): 193-201

[30] Monzo P, Gauthier N C, Keslair F, *et al*. Clues to CD2-associated protein involvement in cytokinesis. Mol Biol Cell, 2005, **16**(6): 2891-2902

[31] Brinkley B R. Managing the centrosome numbers game: from chaos to stability in cancer cell division. Trends Cell Biol, 2001, **11**(1): 18-21

[32] Williams G H, Stoeber K. The cell cycle and cancer. J Pathol, 2012, **226**(2): 352-364

[33] DelSal G, Loda M, Pagano M. Cell cycle and cancer: critical events at the G1 restriction point. Crit Rev Oncog, 1996, **7**(1-2): 127-142

[34] Black B E, Jansen L E, Maddox P S, *et al*. Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. Mol Cell, 2007, **25**(2): 309-322

[35] Li Y, Zhu Z, Zhang S, *et al*. ShRNA-targeted centromere protein A inhibits hepatocellular carcinoma growth. PLoS One, 2011, **6**(3): e17794

[36] Cheng Z, Yu C, Cui S, *et al*. circTP63 functions as a ceRNA to promote lung squamous cell carcinoma progression by upregulating FOXM1. Nat Commun, 2019, **10**: 3200

[37] Wu Q, Qian Y M, Zhao X L, *et al*. Expression and prognostic significance of centromere protein A in human lung adenocarcinoma. Lung Cancer, 2012, **77**(2): 407-414

[38] Degenhardt Y, Lampkin T. Targeting Polo-like kinase in cancer therapy. Clin Cancer Res, 2010, **16**(2): 384-389

[39] Zhou Q, Su Y, Bai M. Effect of antisense RNA targeting Polo-like kinase 1 on cell growth in A549 lung cancer cells. J Huazhong Univ Sci Technolog [Med Sci], 2008, **28**(1): 22-26

[40] Wang Z X, Xue D, Liu Z L, *et al*. Overexpression of polo-like kinase 1 and its clinical significance in human non-small cell lung cancer. Int J Biochem Cell Biol, 2012, **44**(1): 200-210

[41] Tonon G, Wong K, Maulik G, *et al*. High-resolution genomic profiles of human lung cancer. Proc Natl Acad Sci USA, 2005, **102**(27): 9625-9630

[42] Kadara H, Lacroix L, Behrens C, *et al*. Identification of gene signatures and molecular markers for human lung cancer prognosis using an *in vitro* lung carcinogenesis system. Cancer Prev Res, 2009, **2**(8): 702-711

[43] Li Y, Tang H, Sun Z, *et al*. Network-based approach identified cell cycle genes as predictor of overall survival in lung adenocarcinoma patients. Lung Cancer, 2013, **80**(1): 91-98

[44] Ayad N G, Rankin S, Murakami M, *et al*. Tome-1, a trigger of mitotic entry, is degraded during G1 *via* the APC. Cell, 2003, **113**(1): 101-113

[45] O'Byrne K, Adams M, Burgess J, *et al*. 24P CDCA3 regulates the cell cycle and modulates cisplatin sensitivity in non-small cell lung cancer. J Thorac Oncol, 2016, **11**(4 Suppl): S65

[46] Uchida F, Uzawa K, Kasamatsu A, *et al*. Overexpression of cell cycle regulator CDCA3 promotes oral cancer progression by enhancing cell proliferation with prevention of G1 phase arrest. BMC Cancer, 2012, **12**: 321

基于WGCNA和SVM-RFE算法挖掘肺腺癌诊断和预后基因标志物*

王　美　王可心　谭建军**　王京京

（北京工业大学环境与生命学部生物医学工程系，智能化生理测量与临床转化北京市国际科研合作基地，北京 100124）

摘要　目的　肺癌是世界上最常见的癌症之一，在众多肺癌患者中，肺腺癌（lung adenocarcinoma，LUAD）的死亡率最高。基因表达谱的变化与肿瘤的发生和发展过程有关，通过识别与LUAD患者相关的诊断和预后基因标志物，可以为肺腺癌的预防和治疗提供理论依据。**方法**　本研究以肿瘤基因组图谱（The Cancer Gene Atlas，TCGA）数据库为基础，采用加权基因共表达网络分析（weighted gene co-expression network analysis，WGCNA）、差异基因分析、cox回归分析、蛋白质互作网络（protein-protein interaction，PPI）分析等方法筛选与LUAD形成过程高度相关的hub基因。将TCGA和基因型组织表达（GTEx genotype tissue expression，GTEx）数据库中的RNA数据合并划分为训练集和内部验证集，利用基于支持向量机的递归特征消除算法（support vector machine recursive feature elimination feature，SVM-RFE）构建诊断模型并进行验证。GSE32863和GSE31210数据集分别用于验证诊断模型的准确性和基因标志物的预后价值。**结果**　SVM-RFE算法得到的5个基因标志物（*anln*、*cenpa*、*plk1*、*tpx2*、*cdca3*）模型在LUAD患者分类中具有显著的诊断能力。功能富集分析表明，这5个基因与肿瘤发生发展的生物学过程密切相关。此外，这5个基因高表达的LUAD患者的预后表现不良，死亡率显著高于低表达的患者。**结论**　我们的研究为LUAD的诊断和预后提供了具有5个基因特征的模型，这对于开发用于精确治疗的新靶点具有重要意义。

关键词　肺腺癌，基因标志物，加权基因共表达网络分析，递归特征消除算法

中图分类号　R735.7，R318.04　　**DOI**：10.16476/j.pibb.2021.0010

* 北京市自然科学基金（2202002）和国家自然科学基金（21173014）资助项目。

** 通讯联系人。

Tel：010-67392001，E-mail：tanjianjun@bjut.edu.cn

收稿日期：2021-01-13，接受日期：2021-05-14