

Screening key lncRNAs for human lung adenocarcinoma based on machine learning and weighted gene co-expression network analysis

Yu Wang, Junfeng Fu, Ze Wang, Zhenyang Lv, Zhe Fan and Ting Lei*

Department of Thoracic Surgery, The Second Hospital of Dalian Medical University, Dalian, Liaoning, China

Abstract.

BACKGROUND: Lung adenocarcinoma (LUAD) accounts for a significant proportion of lung cancer and there have been few diagnostic and therapeutic targets for LUAD due to the lack of specific biomarker. The aim of this study was to identify key long non-coding RNAs (lncRNAs) for LUAD.

METHODS: The lncRNA and mRNA expression profiles of a large group of patients with LUAD were obtained from The Cancer Genome Atlas (TCGA). The differentially expressed lncRNAs (DElncRNAs) and mRNAs (DEmRNAs) were identified. The optimal diagnostic lncRNA biomarkers for LUAD were identified by using feature selection procedure and classification model. We established classification models including random forests, decision tree and support vector machine to distinguish LUAD and normal tissues. The lncRNAs-mRNAs co-expression networks and module identification were established by weighted gene co-expression network analysis (WGCNA). Functional annotation of pink and green modules was performed. The expression of selected DElncRNAs were validated by qRT-PCR.

RESULTS: A total of 1364 DEmRNAs (468 down-regulated and 896 up-regulated mRNAs) and 260 DElncRNAs (88 down-regulated and 172 up-regulated lncRNAs) between LUAD and normal tissue were obtained. LNC11-AS1, MIR3945HG, LINC01270, RP5-1061H2.4, BLACAT1, LINC01703, CTD-2227E11.1 and RP1-244F24.1 were identified as optimal diagnostic lncRNA biomarkers for LUAD. The area under curve (AUC) of the random forests model, decision tree model and SVM model were 0.999, 0.937 and 0.999, and the specificity and sensitivity of the three model were 98.3% and 99.8%, 93.2% and 99% and 100% and 98.4%, respectively. Co-expression networks analysis showed that RP11-389C8.2, CTD-2510F5.4 and TMPO-AS1 were co-expressed with 44, 242 and 241 mRNAs, respectively. Cell cycle, DNA replication and p53 signaling pathway were three significantly enriched pathways. The qRT-PCR results were consistent with our integrated analysis, generally. The GSE32863 and GSE104854 validation was consistent with our integrated analysis, generally.

CONCLUSION: Our study identified eight DElncRNAs as potential diagnostic biomarkers of LUAD. Functional annotation of green module provided new evidences for exploring the precise roles of lncRNA in LUAD.

Keywords: Lung adenocarcinoma, machine learning, weighted gene co-expression network analysis, long non-coding RNAs

1. Introduction

Lung cancer is the one of the leading causes of cancer-related mortality worldwide [1]. Among all lung cancers, more than 85% are categorized as non-small cell lung cancer, of which lung adenocarcinoma (LUAD) is one of the most common types of non-small cell lung cancer [2]. Since LUAD is easy to form

*Corresponding author: Ting Lei, Department of Thoracic Surgery, The Second Hospital of Dalian Medical University, No. 467 Zhongshan Road, Shahekou District, Dalian, Liaoning, China. E-mail: leitong680612@163.com.

metastases at an early stage, the prognosis for patients with LUAD is poor, with a particularly low 5-year survival rate for patients suffering from this disease at its advanced stages [3]. In spite of advancements that have been made for the diagnosis and treatment of LUAD, the 5-year overall survival of patients with LUAD remains low. Hence, it is urgently required to identify accurate indicators in the diagnostic and therapeutic targets of LUAD.

Emerging evidence has disclosed that deregulation of lncRNAs has been well recognized in cancer, and has been suggested to modulate cancer development at transcriptional and posttranscriptional levels [4,5]. Some well-characterized lncRNAs have been reported to possess oncogenic or tumor suppressive roles and function as a biomarker for cancer diagnosis or prognosis [6,7]. The emerging roles of lncRNAs in LUAD have also been investigated in previous studies. For instance, Su et al. demonstrated that LINC00472 contributed to the increase in LUAD cell apoptosis and the inhibition of proliferation [8]. Tian et al. showed that FENDRR and LINC00312 had diagnostic value in patients with LUAD [9]. lncRNA IOC100132354 promotes angiogenesis through VEGFA/VEGFR2 signaling pathway according to Wang et al.'s study on LUAD [10]. However, research for lncRNA underlying biomarkers in LUAD is rarely. Thus, identification of LUAD-related lncRNAs, and investigation of their molecular mechanisms are essential for understanding the occurrence and development of LUAD.

In the current study, we obtained the lncRNA and mRNA expression data of a large number of patients with LUAD from The Cancer Genome Atlas (TCGA), and attempted to identify the optimal diagnostic lncRNA biomarkers by using feature selection procedure and classification model. The mRNAs-lncRNAs co-expression networks and module identification were established by WGCNA. The functions of the potential lncRNA diagnostic biomarkers in LUAD were further analyzed by functional annotation of pink and green modules. To our knowledge, this is first time to find key lncRNAs in LUAD by using machine learning and weighted gene co-expression network analysis (WGCNA).

2. Materials and methods

2.1. Integrated profiles in TCGA

The Cancer Genome Atlas (TCGA) (<http://tcga-data.nci.nih.gov/>) is a website for multidimensional

cancer genomics data, which provide unprecedented opportunities to uncover pathogenesis of cancer. TCGA data portal included 582 LUAD patients, including clinical data of 522 patients, mRNA and lncRNA data of 515 patients up to June 4, 2018. The lncRNA and mRNA expression profiles and clinical data of LUAD and normal samples was downloaded from TCGA. Our study enrolled only patients who were histologically diagnosed as LUAD.

2.2. Identification DElncRNAs and DEmRNAs between LUAD and normal tissues

The undetectable lncRNAs and mRNAs (with read count value = 0 in more than 20% LUAD case or in more than 20% normal tissues) were filtered and deleted. We carried out multiple comparisons by using the Benjamini and Hochberg method [11] to acquire the false discovery rate (FDR). DElncRNAs and DEmRNAs were obtained according to $FDR < 0.05$ and $|\text{Log2fold change}| > 2$ using R-bioconductor package DESeq2 (<http://bioconductor.org/packages/DESeq2/>). DElncRNAs and DEmRNAs were defined with the thresholds of $FDR < 0.05$ and $|\text{Log2fold change}| > 2$. Hierarchical clustering analysis of DElncRNAs and DEmRNAs were further produced by using R package v3.3.3 (<https://www.r-project.org/>).

2.3. Identification of optimal diagnostic lncRNA biomarkers for LUAD

To identify optimal diagnostic lncRNA biomarkers for LUAD, we performed feature selection procedures as follows. LASSO algorithm was conducted by the glmnet package to decrease dimensions of the data. We performed single10-fold cross-validation cycles with the coordinate descent algorithm for each fold and found regularization parameters that result in the smallest average mean squared errors across all folds. The optimal DElncRNAs were selected in LUAD and normal tissue.

To further identify the optimal diagnostic lncRNA biomarkers for LUAD, we performed feature selection procedures as follows. (1) The importance value of each lncRNA was ranked according to mean decrease in accuracy from large to small by random forest algorithm. (2) The optimum number of features was found by adding a DElncRNA at a time in the top down forward-wrapper packaging method. (3) By using support vector machine (SVM) at each increment and the

optimal diagnostic lncRNA biomarkers were identified for LUAD.

The ‘random Forests’ packet (<https://cran.r-project.org/web/packages/randomForest/>) was used to establish the random forest model. The ‘rpart’ packet (<https://cran.r-project.org/web/packages/rpart/>) was used to build the decision tree mode. The e1071 package (<https://cran.r-project.org/web/packages/e1071/index.html>) in R was used to establish the SVM model. Diagnostic ability of these three models, and each lncRNA biomarker was evaluated by acquiring the area under a receiver operating characteristic (ROC) curve (AUC), sensitivity and specificity. By using pROC package in R language we performed the receiver operating characteristic (ROC) analyses to assess the diagnostic value of lncRNA biomarker. The area under the curve (AUC) under binomial exact confidence interval was calculated and ROC curve was produced.

2.4. Construction of the WGCNA coexpression network

WGCNA is an algorithm used to build co-expression modules and network. Briefly, a co-expression similarity matrix was calculated by the Pearson’s correlation coefficient between mRNAs and lncRNAs pairs. According to the scale-free topology criterion described in the WGCNA package documents, the similarity matrix was raised to a soft thresholding power of power $\beta = 20$ to yield the adjacency matrix. Then, the adjacency matrix was transformed into topological overlap matrix, and was in turn converted into a dissimilarity topological overlap matrix, from which a dendrogram was mapped via hierarchical clustering. The clusters were obtained from the dendrogram by using the dynamic tree cutting technique. The resulting clusters are co-expression modules containing mRNAs and lncRNAs that are considerably interconnected. The modules with absolute value of $r > 0.5$ and $p < 0.001$ were identified to be LUAD associated modules. The Cytoscape software [12] was used to build mRNAs and lncRNAs co-expression network.

2.5. Functional annotation of green and pink module

To display the biological functions and the potential pathways of green and pink module for LUAD, Gene Ontology (GO) classification and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were performed by using g: Profiler (<https://biit.cs.ut.ee/gprofiler/>). $P < 0.05$ was defined as the criterion of statistical significance.

Table 1
Primer sequences used for qRT-PCR

Name	Sequence (5' to 3')
GAPDH-F	GGAGCGAGATCCCTCCAAAAT
GAPDH-R	GGCTGTTGTCATACTTCTCATGG
LANCL1-AS1-F	TGGATCTTCAGCAGCCACCT
LANCL1-AS1-R	TCCAGGCCTGATAGTGTAGCT
MIR3945HG-F	TCAGCCTCAGTGACAGAGCA
MIR3945HG-R	CCCAGGCCGATCTTGGAATG
LINC01270-F	TGACAGGAAGCAGCACAGAG
LINC01270-R	ATGCAAGGGACAGAAAGCCA
RP5-1061H20.4-F	GGGACTGCTGATACTGAGGCT
RP5-1061H20.4-R	CCAAAGAGGCCCAAGTTCC
BLACAT1-F	GGCAAGAGACCCCTGAGAGT
BLACAT1-R	CACGAGTCAGTTCAGCCGAG
LINC01703-F	TACAGGCATGAACCACGGTG
LINC01703-R	GCAGCAATCTCTAATCACGCG
CTD-2227E11.1-F	GAAATGGTTCCTGTGCCCGT
CTD-2227E11.1-R	TCTCCCTCATTGAGCCTTCCT
RP1-244F24.1-F	CAGACTAGGGGCAATCTCGC
RP1-244F24.1-R	GGGAGGGTGAAGAGTGCTGA

2.6. Confirmation by qRT-PCR

According to the results of TCGA integration analysis, eight optimal diagnostic lncRNAs biomarkers (LANCL1-AS1, MIR3945HG, LINC01270, RP5-1061H20.4, BLACAT1, LINC01703, CTD-2227E11.1 and RP1-244F24.1) were screened as candidate lncRNAs. Ten tissues samples of LUAD patients ($n = 5$) and normal adjacent ($n = 5$) were obtained. This study has been approved by the ethics institute of the second hospital of Dalian medical university. The signed informed consents of all the participants were obtained. Total RNA was extracted with a RNA simple total RNA kit (Tiangen, China). Complementary DNAs were generated using the Fast Quant RT Kit (Tiangen, China). Quantitative real-time PCR were conducted using the Super Real PreMix Plus SYBR Green (Tiangen, China) on ABI 7500 real-time PCR system. Relative quantification of mRNA levels was analyzed by using the $2^{-\Delta\Delta Ct}$ method. The PCR primers used are listed in Table 1. The human GAPDH and ACTB were used as endogenous controls for mRNA and lncRNA expression in analysis.

2.7. Validation in the Gene Expression Omnibus (GEO) dataset

GSE32863 dataset was downloaded from the GEO (<https://www.ncbi.nlm.nih.gov/geo/>), which consisted of 60 patients with LUAD and 60 normal controls. GSE104854 dataset was downloaded from the GEO, which consisted of 9 patients with LUAD and 9 normal controls. The dataset GSE32863 and GSE104854

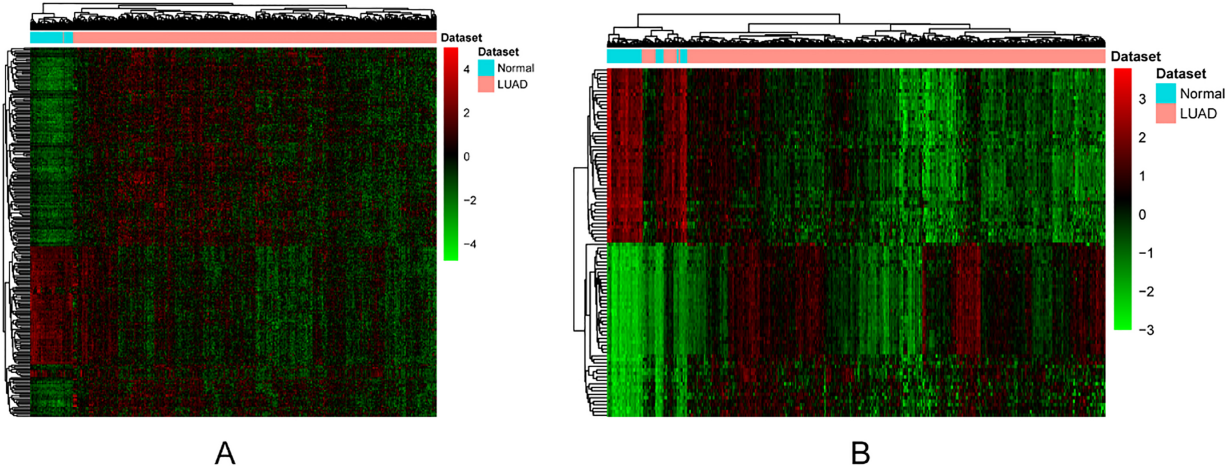


Fig. 1. The heat-maps of DElncRNAs and top 100 DEMRNAs between LUAD and normal tissues. (A) DEMRNAs. (B) DElncRNAs. Row and column represented DElncRNAs/DEMNRNs and tissue samples, respectively. The color scale represented the expression levels.

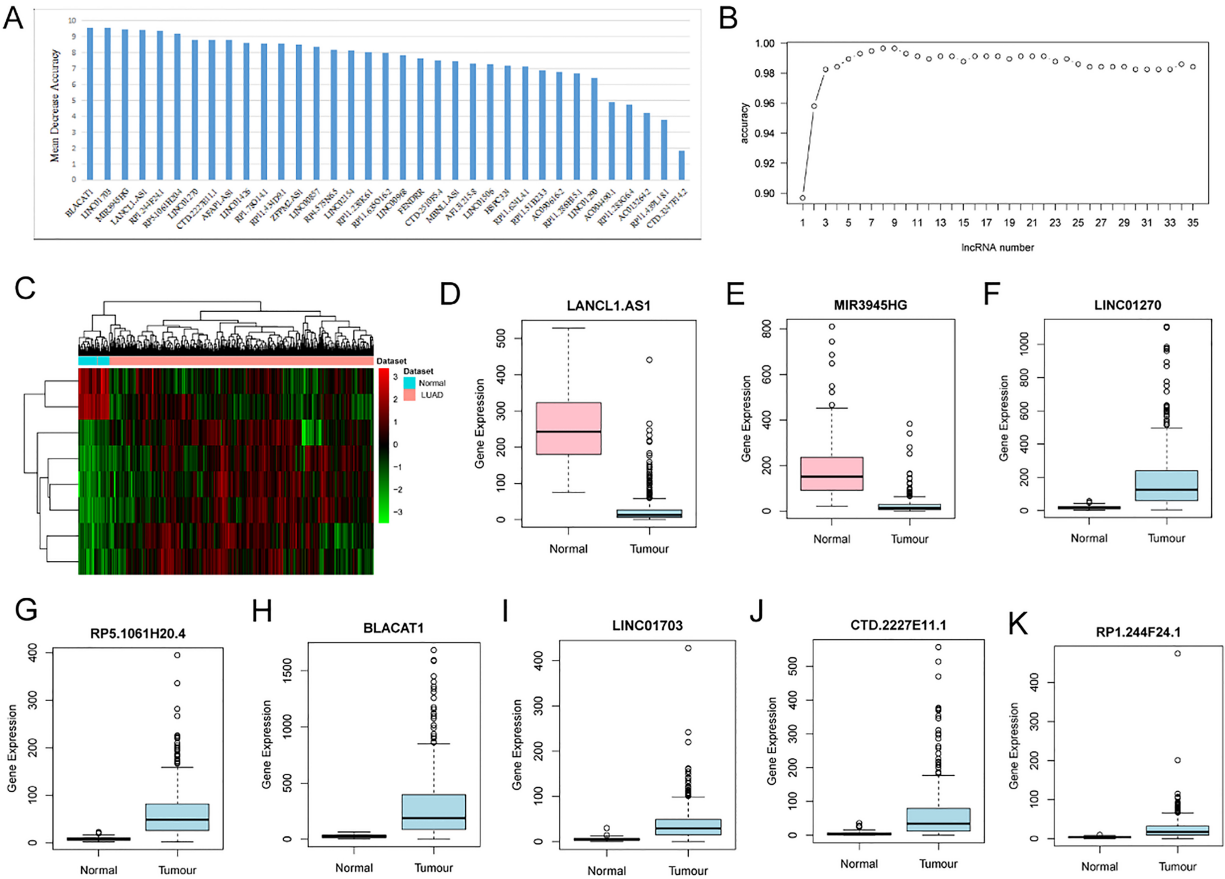


Fig. 2. Identification of optimal lncRNA biomarkers for LUAD. (A) The importance value of each DElncRNA ranked according to the mean decrease in accuracy by using the random forest analysis. (B) The variance rate of classification performance when increasing numbers of the predictive DElncRNAs. (C) Heat-map of LUAD-specific eight lncRNAs biomarkers (LANCL1-AS1, MIR3945HG, LINC01270, RP5-1061H20.4, BLACAT1, LINC01703, CTD-2227E11.1 and RP1-244F24.1). (D-K) Box-plot displayed the expression levels of eight lncRNAs biomarkers between LUAD and normal tissues. The X-axis represented normal and LUAD groups. The Y-axis represented gene expression levels.

examined the tissues. The GEO dataset GSE32863 and GSE104854 to validate the expression of some DEM-RNAs and DElncRNAs, respectively.

3. Results

3.1. DEMRNAs and DElncRNAs in LUAD

The clinical data of 522 LUAD patients were downloaded from TCGA, among which, 513 LUAD patients were enrolled in this study. Males and females accounted for 46.2% and 53.8%, respectively. The medial age of these 513 patients was 65 years old. Whites, Asians and Black or African American accounted for 86.7, 1.6% and 11.7%, respectively. The tumor stages were as follows: Stage I, 54.2%; Stage II, 24%; Stage III, 16.7%; and Stage IV, 5.1%. The mRNA and lncRNA expression profiles of 513 LUAD tumor tissues and 59 normal tissues were obtained. A total of 3177 lncRNAs and 15183 mRNAs were available for analysis after removing those hardly detected lncRNAs and mRNAs.

A total of 1364 DEMRNAs (468 down-regulated and 896 up-regulated mRNAs) and 260 DElncRNAs (88 down-regulated and 172 up-regulated lncRNAs) between LUAD and normal tissue were identified with $FDR < 0.05$ and $|\text{Log2fold change}| > 2$. Heat-maps of DElncRNAs and the top 100 DEMRNAs between LUAD and normal tissue were displayed in Fig. 1A and B, respectively.

3.2. Identification of optimal diagnostic lncRNA biomarkers for LUAD

Based on reduced dimensions of the data, we obtained 35 DElncRNAs between LUAD and adjacent normal tissues by using LASSO algorithm (Table 2). The random forest analysis was used to rank the 35 DElncRNAs according to the mean decrease in accuracy (Fig. 2A). Ten-fold cross-validation result suggested that the average accuracy rate of 8 DElncRNAs (LANCL1-AS1, MIR3945HG, LINC01270, RP5-1061H20.4, BLACAT1, LINC01703, CTD-2227-E11.1 and RP1-244F24.1) reached the higher score for the first time (Fig. 2B). Therefore, these 8 DElncRNAs were determined as the optimal diagnostic lncRNA biomarkers for LUAD which were used to establish the random forests, decision tree and SVM models. The heat-map of these 8 DElncRNAs between LUAD and normal tissue were displayed in

Table 2
DElncRNAs between LUAD and normal tissues after reduced dimensions of data

Symbol	Log2fold change	P value	Up down
AFAP1-AS1	6.147558188	1.26E-121	Up
HSPC324	-3.711651228	3.78E-103	Down
RP4-575N6.5	-3.004823837	1.03E-101	Down
ZFPM2-AS1	5.139179367	2.69E-87	Up
FENDRR	-3.958113208	9.98E-86	Down
LANCL1-AS1	-3.746032429	2.54E-85	Down
LINC00968	-3.58721188	4.86E-79	Down
MIR3945HG	-3.569183879	5.20E-74	Down
LINC01270	2.789264642	6.32E-69	Up
RP11-51B23.3	-2.473610518	2.43E-66	Down
RP5-1061H20.4	2.276375431	5.07E-64	Up
LINC01290	-2.35973866	1.04E-63	Down
BLACAT1	2.865795784	1.20E-56	Up
LINC00857	2.265590784	1.52E-53	Up
MBNL1-AS1	-2.008379964	1.06E-51	Down
LINC01703	2.31228291	1.74E-48	Up
RP11-286H15.1	-3.08225938	6.06E-48	Down
RP1-78O14.1	-3.205548623	3.17E-47	Down
LINC01426	2.396928093	2.39E-46	Up
RP11-283G6.4	-2.189974164	2.27E-44	Down
CTD-2227E11.1	2.750378309	6.94E-43	Up
CTD-2510F5.4	2.387911254	1.64E-39	Up
LINC01506	-2.465002349	2.16E-39	Down
RP11-439L18.1	-2.50043196	1.95E-37	Down
AC004490.1	-2.223065997	8.41E-36	Down
AC090616.2	-2.086527767	2.26E-34	Down
RP1-244F24.1	2.114328604	7.95E-34	Up
RP11-434D9.1	-2.722249168	1.24E-31	Down
LINC02154	-3.173875177	4.04E-29	Down
AF131215.8	-2.319990895	1.74E-28	Down
CTD-3247F14.2	-2.374869893	2.72E-26	Down
RP11-624L4.1	2.300316719	5.98E-24	Up
RP11-635O16.2	-2.476332652	2.15E-16	Down
AC013264.2	-2.033028215	1.28E-14	Down
RP11-238K6.1	-2.391757858	3.94E-14	Down

Fig. 2C. Box-plot uncovered the expression levels of these 8 DElncRNAs between LUAD and normal tissues (Fig. 2D–K). The AUC of the random forests model was 0.999 and the specificity and sensitivity of this model were 98.3% and 99.8%, respectively (Fig. 3A). The AUC of the decision tree model was 0.937 and the specificity and sensitivity of this model were 93.2.0% and 99%, respectively (Fig. 3B). The AUC of the SVM model was 0.999 and the specificity and sensitivity of this model were 100% and 98.4% (Fig. 3C). The AUC of all these 8 lncRNAs LANCL1-AS1, MIR3945HG, LINC01270, RP5-1061H20.4, BLACAT1, LINC01703, CTD-2227E11.1 and RP1-244F24.1 were also above 0.89 (Fig. 3D–K). Taken together, the AUC of all these eight lncRNAs and their combination were all greater than 0.89 which indicated the LANCL1-AS1, MIR3945HG, LINC01270, RP5-1061H20.4, BLACAT1, LINC01703, CTD-

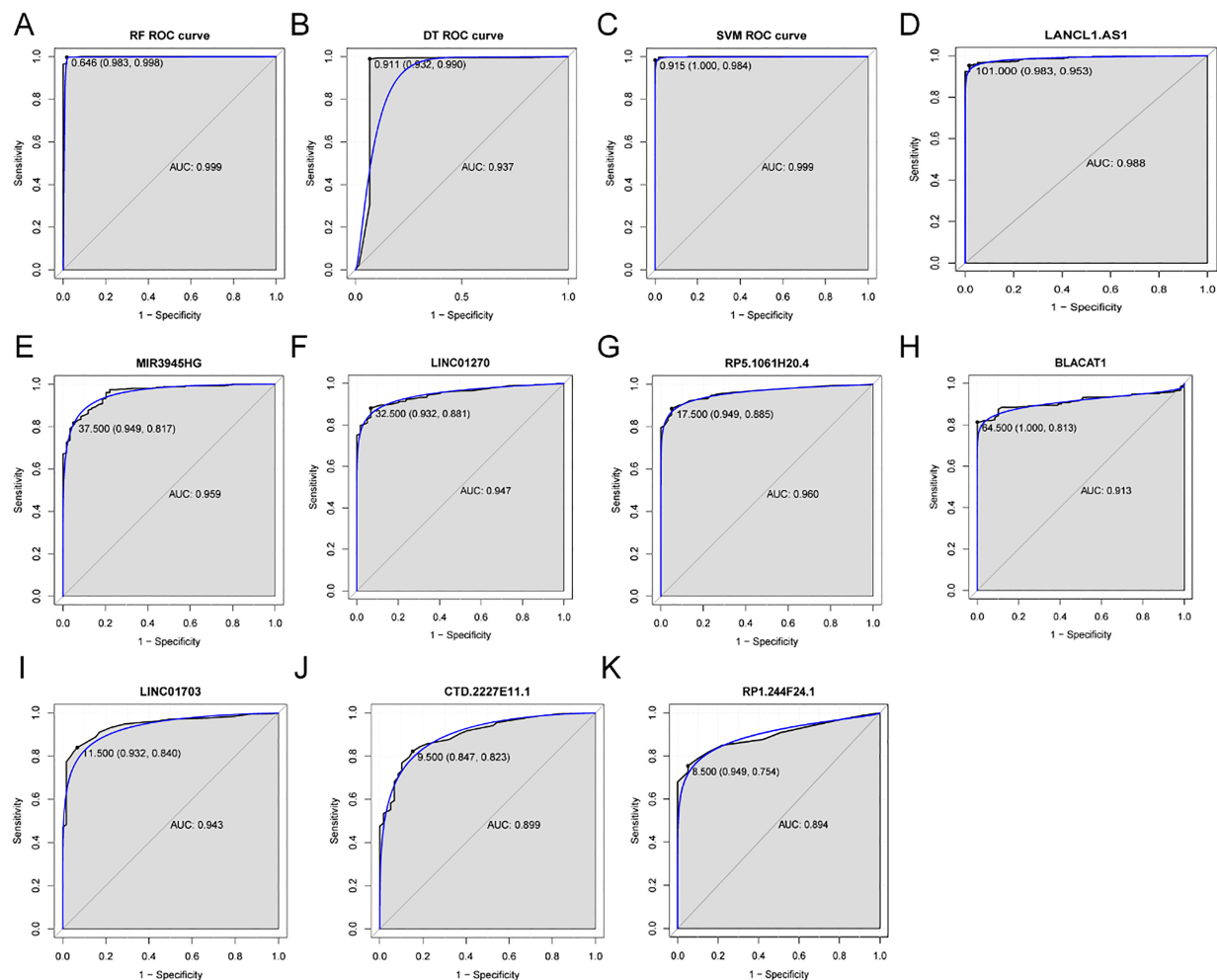


Fig. 3. ROC analysis of eight LUAD-specific lncRNAs biomarkers. The ROC results of these eight diagnostic lncRNAs biomarkers (LANCL1-AS1, MIR3945HG, LINC01270, RP5-1061H20.4, BLACAT1, LINC01703, CTD-222E11.1 and RP1-244F24.1) their combination based on support random forest (Fig. 3A), decision tree model (Fig. 3B) and SVM model model (Fig. 3C) and individual LANCL1-AS1 (Fig. 3D), MIR3945HG (Fig. 3E), LINC01270 (Fig. 3F), RP5-1061H20.4 (Fig. 3G), BLACAT1 (Fig. 3H), LINC01703 (Fig. 3I), CTD-222E11.1 (Fig. 3J) and RP1-244F24.1 (Fig. 3K).

222E11.1 and RP1-244F24.1 and their combination were related to LUAD and could predict the occurrence of LUAD.

3.3. lncRNAs-mRNAs co-expression networks and module identification

WGCNA was performed to identify lncRNAs-mRNAs co-expression networks and module identification. We identified nine module eigengenes (ME), including MEmagenta, MEblue, MEyellow, MEturquoise, MEpink, MERed, MEgreen, MEBrown and MEblack (Fig. 4A). In order to further explore the correlation between modules and tumor-adjacent can-

cer, the correlation analysis was performed, and the modules with the correlation coefficient > 0.5 and $p < 0.01$ were considered to be related. Interestingly, we found that two module (MEpink and MEgreen) were associated with tumor-adjacent cancer (Fig. 4B). Among which, MEpink was negatively with associated with tumor-adjacent cancer ($p = 1.50E-44$ and $|r| = 0.55$) and MEgreen was positively with associated with tumor-adjacent cancer ($p = 1.10E-47$ and $|r| = 0.56$). The pink module contained a total of 44 down-regulated mRNAs and 1 down-regulated lncRNA (RP11-389C8.2) and the green module contained a total of 241 up-regulated mRNAs and 2 up-regulated lncRNAs (CTD-2510F5.4 and TMPO-AS1),

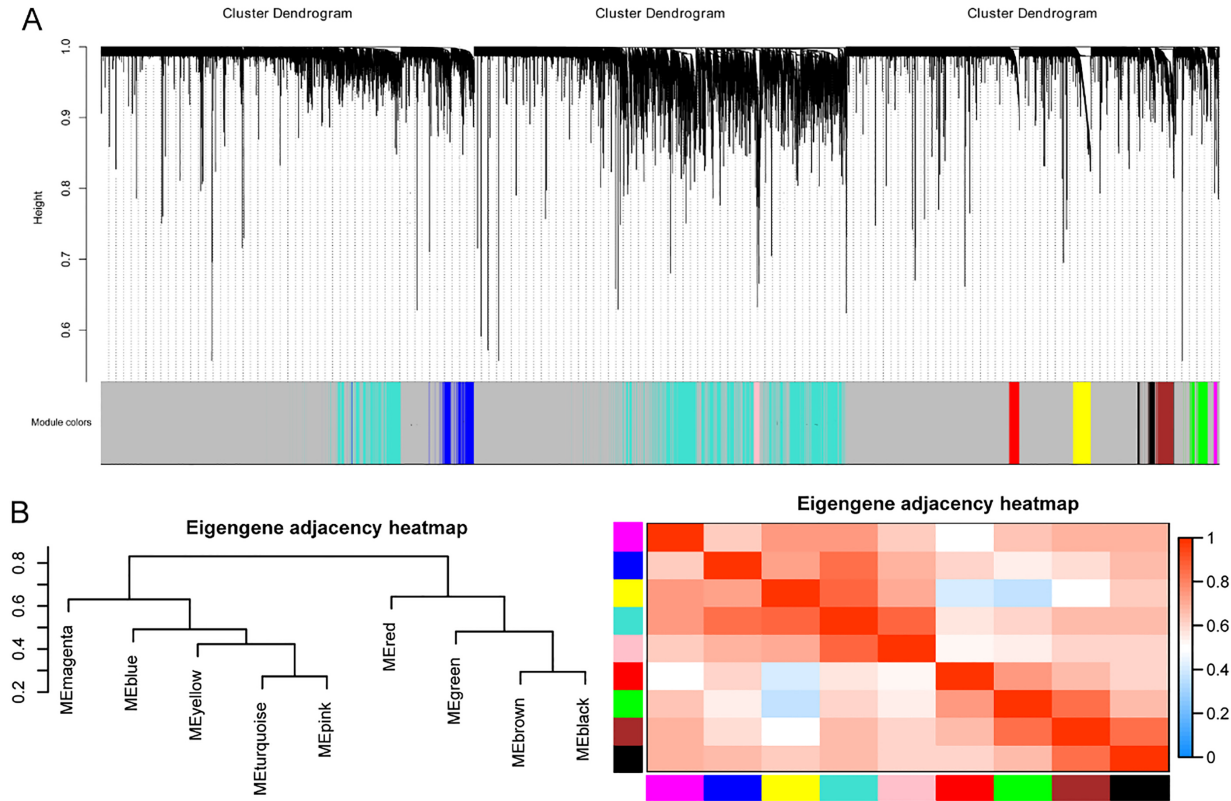


Fig. 4. Network analysis of gene expression in LUAD identifies distinct modules of co-expression data. (A) Cluster dendrogram of the identified lncRNA-mRNA co-expression modules. The lower panel shows colors designated for each module. Note that the leaves presented in the color gray indicates unassigned lncRNAs and mRNAs. (B) The cluster dendrogram of module eigengenes.

which were selected as LUAD associated mRNAs and lncRNAs for the further analysis. Co-expression networks analysis showed that RP11-389C8.2, CTD-2510F5.4 and TMPO-AS1 were co-expressed with 44, 242 and 241 mRNAs, respectively (Fig. 5).

3.4. Functional annotation of pink and green modules

According to the functional annotation of 44 mRNAs in pink modules (Fig. 6), blood vessel morphogenesis ($p = 2.08E-09$), angiogenesis ($p = 3.59E-09$), cell adhesion ($p = 0.000209$) and cell migration ($p = 0.0162$) were significantly enriched GO terms, and Leukocyte transendothelial migration ($p = 0.0282$) and cell adhesion molecules ($p = 0.00222$) were two significantly enriched pathways. Base on the functional annotation of 241 mRNAs in green modules (Fig. 7A–D), cell cycle ($p = 1.92E-145$), mitotic cell cycle ($p = 7.54E-113$), chromosome ($p = 1.24E-81$) and protein binding ($p = 6.88E-21$) were significantly enriched GO terms, and cell cycle ($p = 1.82E-35$), DNA replication ($p = 1.02E-19$) and p53 signaling pathway ($p = 1.71E-05$) were three significantly enriched pathways.

3.5. Confirmation by qRT-PCR

We performed the confirmation of eight optimal diagnostic lncRNAs biomarkers by qRT-PCR. Based on TCGA, LNCCL1-AS1 and MIR3945HG were down-regulated while the other six lncRNAs (LINC01270, RP5-1061H20.4, BLACAT1, LINC01703, CTD-2227-E11.1 and RP1-244F24.1) were up-regulated in LUAD compared to adjacent tissues. According to the qRT-PCR results, except for RP5-1061H20.4 and LINC-01703, two lncRNAs were down-regulated and four lncRNAs were up-regulated which was consistent with the results of TCGA, generally (Fig. 8).

3.6. Validation in GEO dataset

The expression pattern of selected DEmRNAs (ETV4, EPAS1, TUBB3 and PECAM1) and DElncRNAs (BLACAT1 and TMPO.AS1) was verified using GSE32863 and GSE104854 dataset, respectively. As shown in Fig. 9, TMPO.AS1 was down-regulated, which was inconsistent with our integration results.

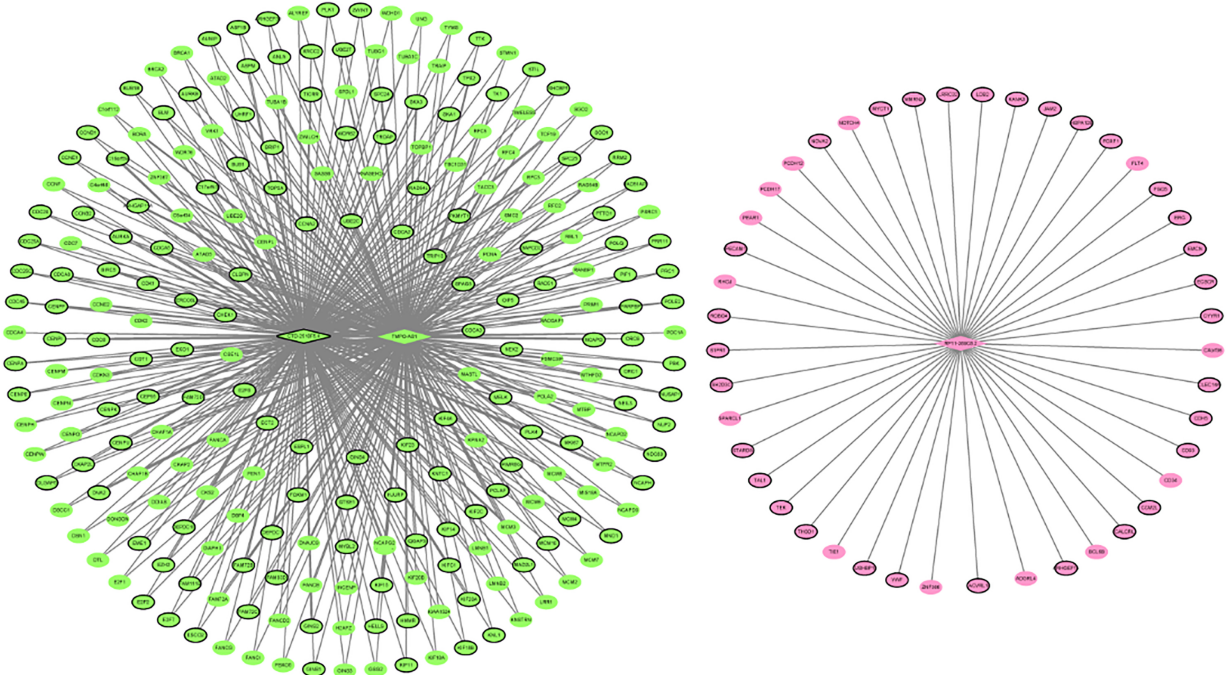


Fig. 5. lncRNAs-mRNAs co-expression network. The ellipses and rhombuses were represented the mRNAs and lncRNAs, respectively. Green and pink color represented green module and pink module, respectively. The black border indicates DEMRNAs and DELncRNAs.

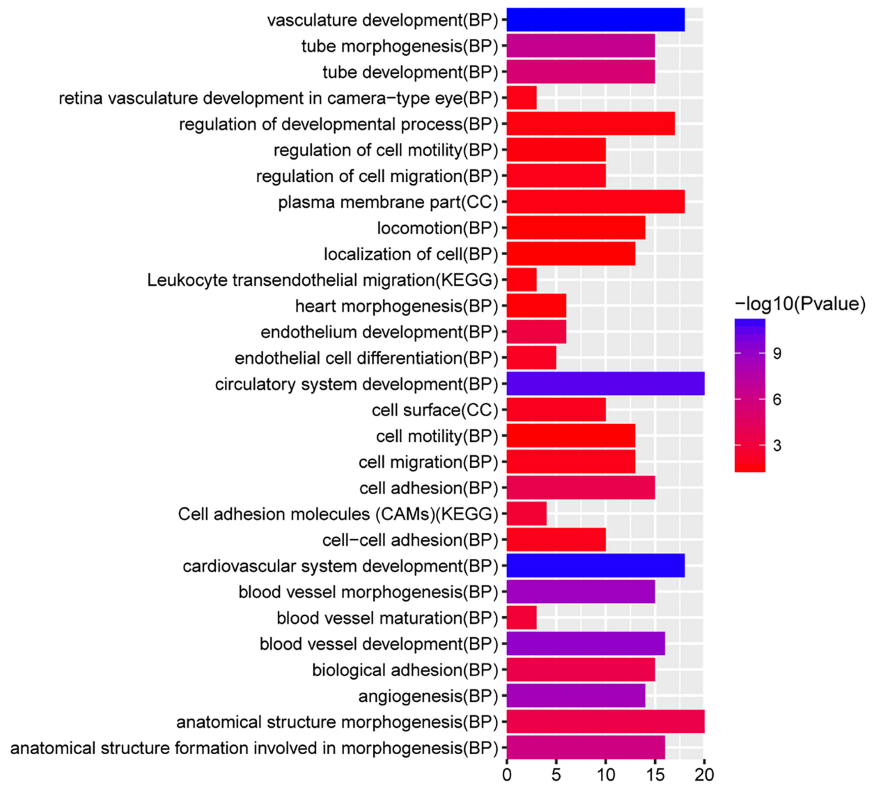


Fig. 6. The function annotation of pink module. The x-axis shows-log FDR and y-axis shows GO terms or KEGG pathways.

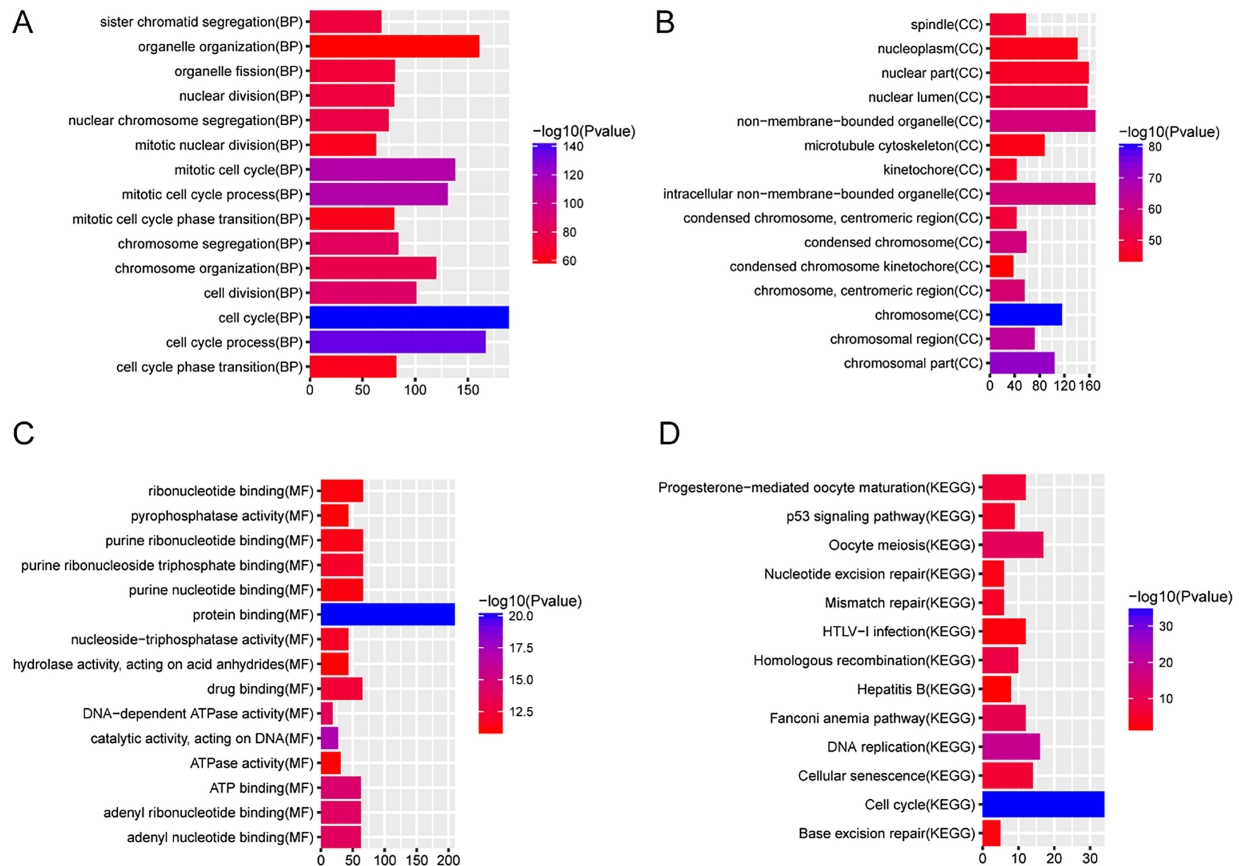


Fig. 7. The function annotation of green module. The x-axis shows -log FDR and y-axis shows GO terms or KEGG pathways. (A) Biological process. (B) Cellular component. (C) Molecular function. (D) KEGG pathways.

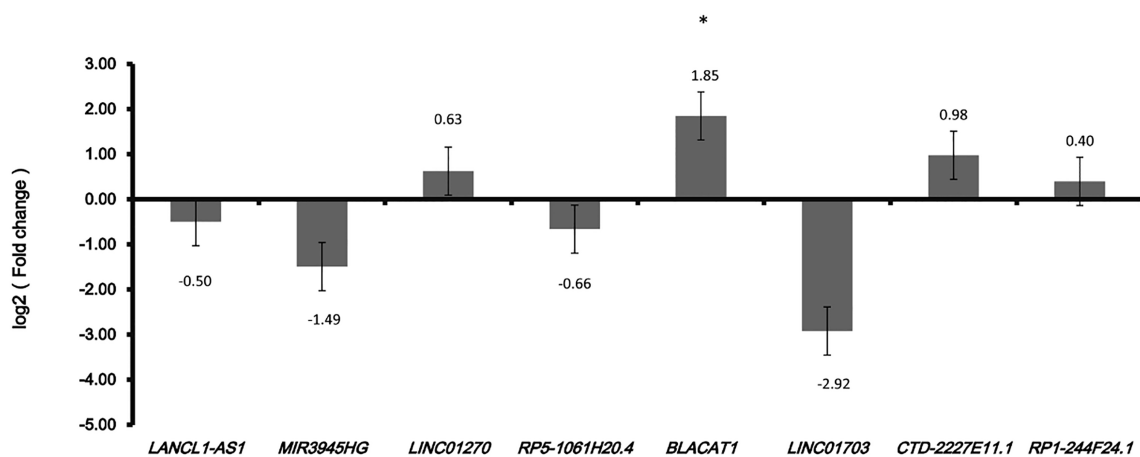


Fig. 8. Validation optimal lncRNA biomarkers in LUAD tissue by qRT-PCR. * $P < 0.05$.

ETV4, TUBB3 and BLACAT1 were up-regulated while EPAS1 and PECAM1 were down-regulated in LUAD, which was consistent with our integration results, suggesting that the results were convincing.

4. Discussion

Lung cancer has the highest morbidity and mortality rate of lethal disease worldwide, and LUAD is one

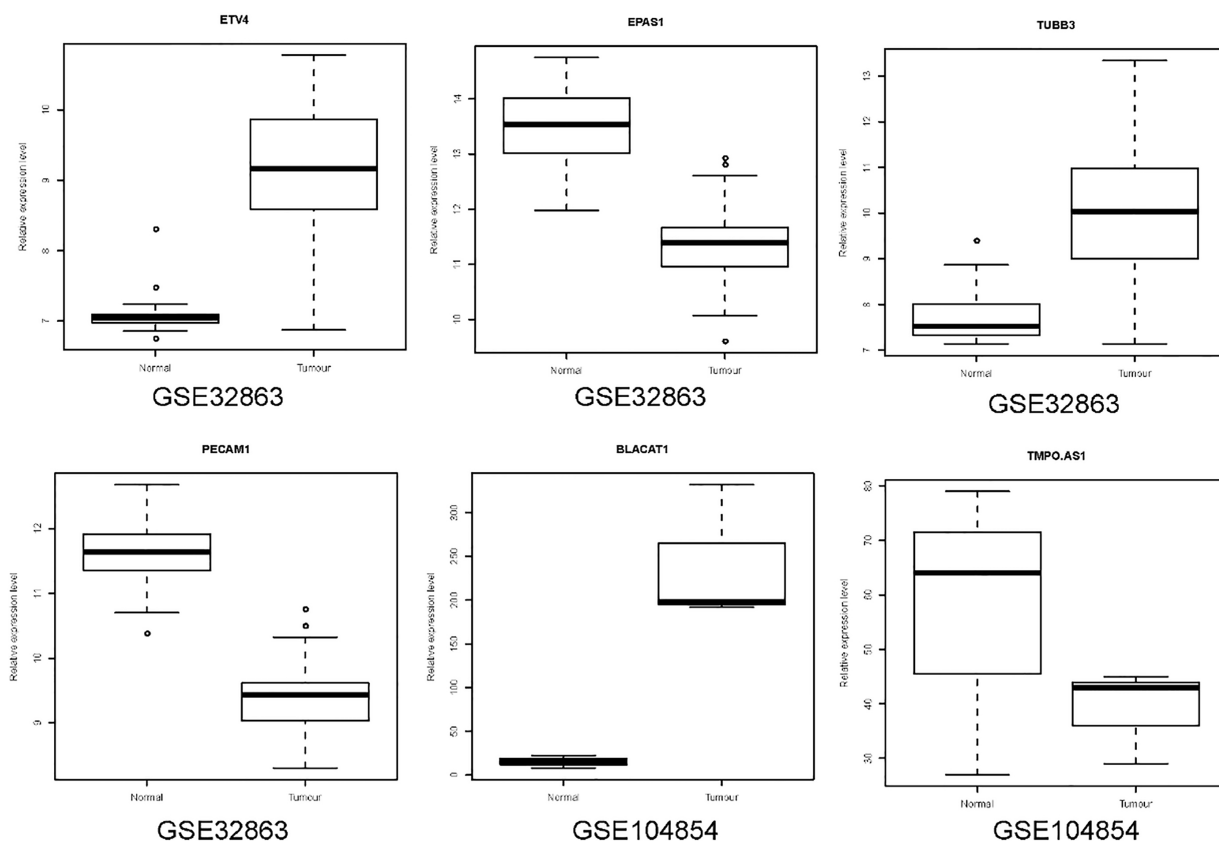


Fig. 9. Validation of selected DE mRNAs and DE lncRNAs in GSE32863 and GSE104854. The x-axis shows healthy normal control and LUAD groups and y-axis shows a log2 transformation to the intensities.

of the most common types of lung cancer [13]. Although the prognosis and diagnosis of LUAD has been achieved through studies, the 5-year overall survival of patients with LUAD remains low, requiring continued research to identify novel biomarkers and decipher the detailed molecular mechanisms underlying of LUAD. lncRNAs have been verified to play a key role in the occurrence and progression of lung cancer [1,8]. Here is a report studying the DE mRNAs and DE lncRNAs in a large group of patients with LUAD from TCGA. A total of 1364 DE mRNAs (468 down-regulated and 896 up-regulated mRNAs) and 260 DE lncRNAs (88 down-regulated and 172 up-regulated lncRNAs) between LUAD and normal tissue were obtained. Additionally, eight optimal diagnostic lncRNAs biomarkers for LUAD were identified by feature selection procedure and classification model assay. Moreover, we used the WGCNA to build lncRNAs-mRNAs co-expression networks and module identification. The functional annotation of pink and green modules were evaluated.

Feature selection methods provided by machine learning approaches are an interesting, flexible and

robust alternative for identifying predictors that contribute to disease occurrence [14]. Compared with similar type of LUAD studies [15–17], we mainly used the machine learning to identify optimal diagnostic lncRNA biomarkers of LUAD. The AUC of all these eight lncRNAs (LANCL1-AS1, MIR3945HG, LINC01270, RP5-1061H20.4, BLACAT1, LINC01703, CTD-2227E11.1 and RP1-244F24.1) and their combination in LUAD were more than 0.89, which indicated eight lncRNAs and their combination could predict the occurrence of LUAD. To our knowledge, except MIR3945HG and BLACAT1, one down-regulated DE lncRNA (LANCL1-AS1) and five up-regulated DE lncRNA (LINC01270, RP5-1061H20.4, LINC01703, CTD-2227E11.1 and RP1-244F24.1) in LUAD was firstly reported and their biological function remain unclear.

It is reported that MIR3945HG was identified as novel candidate diagnostic markers for tuberculosis [18]. Recent research results indicated that MIR3945HG was the highest diagnostic values biomarker in the diagnosis of lung squamous cell car-

cinoma, and was closely related to the survival time of lung squamous cell carcinoma [15]. In the present study, we found that MIR3945HG was down-regulated in both informatics and qRT-PCR validation. Therefore, we hypothesized that MIR3945HG might involve in the development of LUAD.

BLACAT1 (bladder cancer associated transcript 1), also known as linc-UBC1, is a novel identified lncRNA in bladder cancer, and has been tested to function as an oncogenic lncRNA in several types of human cancer [19–21]. The overexpression of BLACAT1 in non-small cell lung cancer tissue and cells, and discovered the oncogenic role of BLACAT1 in non-small cell lung cancer genesis through sponging miR-144, providing a potential biomarker for early detection and prognosis prediction of non-small cell lung cancer [22]. It is reported BLACAT1 was associated with the malignant status and prognosis in patients with small-cell lung cancer, and functions as an oncogenic lncRNA in regulating cell proliferation and motility, suggesting BLACAT1 might act as a potential target for small-cell lung cancer revention and treatment [23]. In this study, we found that BLACAT1 was the top DElncRNA in LUAD, which was consistent with reports of other researchers [24], indicating the TCGA integration analysis data were reliable.

WGCNA results showed that green was positively with associated with tumor-adjacent cancer and contained a total of 241 up-regulated mRNAs and 2 up-regulated lncRNAs (CTD-2510F5.4 and TMPO-AS1). In this study, CTD-2510F5.4 were significantly up-regulated in LUAD. Moreover, CTD-2510F5.4 was also found to be differentially expressed in another study that used RNA-seq data from TCGA and two independent experiments [1,25], which supports the reliability of our results. Li et al. showed that TMPO-AS1 was revealed to be prognostic biomarker for LUAD [17]. Peng et al. reported that TMPO-AS1 could affect the prognosis of LUAD patients through regulating cell cycle and cell adhesion [16]. Herein, we found that TMPO-AS1 was significantly up-regulated in LUAD. Hence, we further confirmed that TMPO-AS1 might be prognostic biomarker for LUAD. According to the functional annotation of 241 mRNAs in green modules, cell cycle, DNA replication and p53 signaling pathway were three significantly enriched pathways. Therefore, we speculated that CTD-2510F5.4 and TMPO-AS1 might involve in the development of LUAD by regulating signaling pathway of cell cycle, DNA replication and p53 signaling pathway.

In summary, we identified 1364 DEmRNAs and 260 DElncRNAs in LUAD compared to normal tis-

suess. The feature selection procedure and classification model assay was to obtain eight optimal diagnostic lncRNAs biomarkers for LUAD. WGCNA was used to identify mRNAs-lncRNAs co-expression networks, and we obtained three lncRNAs (RP11-389C8.2, CTD-2510F5.4 and TMPO-AS1), which were related to LUAD, and built lncRNAs-mRNAs co-expression networks. However, there are limitations to our study. Firstly, the sample size in the confirmation by qRT-PCR was small and large numbers of samples of LUAD are needed for further research. Secondly, these DEmRNAs and DElncRNAs in LUAD were identified and the bio-function was not studied. Therefore, *in vivo* and *in vitro* experiments were necessary to illuminate the biological roles of DEmRNAs and DElncRNAs in LUAD in the future work.

References

- [1] X. Wang, G. Li, Q. Luo, J. Xie and C. Gan, Integrated TCGA analysis implicates lncRNA CTB-193M12.5 as a prognostic factor in lung adenocarcinoma, *Cancer Cell Int* **18** (2018), 27.
- [2] R. Siegel, D. Naishadham and A. Jemal, Cancer statistics, 2012, *CA Cancer J Clin* **62** (2012), 10–29.
- [3] J.J. Lin, S. Cardarella, C.A. Lydon, S.E. Dahlberg, D.M. Jackman, P.A. Janne and B.E. Johnson, Five-year survival in EGFR-mutant metastatic lung adenocarcinoma treated with EGFR-TKIs, *J Thorac Oncol* **11** (2016), 556–565.
- [4] G. Yang, X. Lu and L. Yuan, lncRNA: a link between RNA and cancer, *Biochim Biophys Acta* **1839** (2014), 1097–1109.
- [5] W. Xie, S. Yuan, Z. Sun and Y. Li, Long noncoding and circular RNAs in lung cancer: advances and perspectives, *Epigenomics* **8** (2016), 1275–1287.
- [6] S.W. Cheetham, F. Gruhl, J.S. Mattick and M.E. Dinger, Long noncoding RNAs and the genetics of cancer, *Br J Cancer* **108** (2013), 2419–2425.
- [7] J.R. Prensner and A.M. Chinnaiyan, The emergence of lncRNAs in cancer biology, *Cancer Discov* **1** (2011), 391–407.
- [8] C. Su, K. Shi, X. Cheng, Y. Han, Y. Li, D. Yu and Z. Liu, Long noncoding RNA LINC00472 inhibits proliferation and promotes apoptosis of lung adenocarcinoma cells via regulating miR-24-3p/DEDD, *Technol Cancer Res Treat* **17** (2018), 1533033818790490.
- [9] Z. Tian, S. Wen, Y. Zhang, X. Shi, Y. Zhu, Y. Xu, H. Lv and G. Wang, Identification of dysregulated long non-coding RNAs/microRNAs/mRNAs in TNM I stage lung adenocarcinoma, *Oncotarget* **8** (2017), 51703–51718.
- [10] Y. Wang, F. Zhang, J. Wang, L. Hu, F. Jiang, J. Chen and L. Wang, lncRNA LOC100132354 promotes angiogenesis through VEGFA/VEGFR2 signaling pathway in lung adenocarcinoma, *Cancer Manag Res* **10** (2018), 4257–4266.
- [11] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)* **57** (1995), 289–300.
- [12] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, Cytoscape: a software environment for integrated models of biomolecu-

- lar interaction networks, *Genome Research* **13** (2003), 2498–2504.
- [13] X. Li, B. Li, P. Ran and L. Wang, Identification of ceRNA network based on a RNA-seq shows prognostic lncRNA biomarkers in human lung adenocarcinoma, *Oncol Lett* **16** (2018), 5697–5708.
- [14] G. Machado, M.R. Mendoza and L.G. Corbellini, What variables are important in predicting bovine viral diarrhea virus? A random forest approach, *Vet Res* **46** (2015), 85.
- [15] W.J. Chen, R.X. Tang, R.Q. He, D.Y. Li, L. Liang, J.H. Zeng, X.H. Hu, J. Ma, S.K. Li and G. Chen, Clinical roles of the aberrantly expressed lncRNAs in lung squamous cell carcinoma: a study based on RNA-sequencing and microarray data mining, *Oncotarget* **8** (2017), 61282–61304.
- [16] F. Peng, R. Wang, Y. Zhang, Z. Zhao, W. Zhou, Z. Chang, H. Liang, W. Zhao, L. Qi, Z. Guo and Y. Gu, Differential expression analysis at the individual level reveals a lncRNA prognostic signature for lung adenocarcinoma, *Mol Cancer* **16** (2017), 98.
- [17] D.S. Li, J.L. Ainiwaer, I. Sheyhiding, Z. Zhang and L.W. Zhang, Identification of key long non-coding RNAs as competing endogenous RNAs for miRNA-mRNA in lung adenocarcinoma, *Eur Rev Med Pharmacol Sci* **20** (2016), 2285–2295.
- [18] X. Yang, J. Yang, J. Wang, Q. Wen, H. Wang, J. He, S. Hu, W. He, X. Du, S. Liu and L. Ma, Microarray analysis of long noncoding RNA and mRNA expression profiles in human macrophages infected with *Mycobacterium tuberculosis*, *Sci Rep* **6** (2016), 38963.
- [19] X. Chen, M. Dai, H. Zhu, J. Li, Z. Huang, X. Liu, Y. Huang, J. Chen and S. Dai, Evaluation on the diagnostic and prognostic values of long non-coding RNA BLACAT1 in common types of human cancer, *Mol Cancer* **16** (2017), 160.
- [20] C.H. Wang, Y.H. Li, H.L. Tian, X.X. Bao and Z.M. Wang, Long non-coding RNA BLACAT1 promotes cell proliferation, migration and invasion in cervical cancer through activation of Wnt/beta-catenin signaling pathway, *Eur Rev Med Pharmacol Sci* **22** (2018), 3002–3009.
- [21] D. Liao, G. Lv, T. Wang, J. Min, Y. Wang and S. Liu, Prognostic value of long non-coding RNA BLACAT1 in patients with papillary thyroid carcinoma, *Cancer Cell Int* **18** (2018), 47.
- [22] J.R. Ye, L. Liu and F. Zheng, Long noncoding RNA bladder cancer associated transcript 1 promotes the proliferation, migration, and invasion of nonsmall cell lung cancer through sponging miR-144, *DNA Cell Biol* **36** (2017), 845–852.
- [23] W. Chen, Y. Hang, W. Xu, J. Wu, L. Chen, J. Chen, Y. Mao, J. Song and H. Wang, BLACAT1 predicts poor prognosis and serves as oncogenic lncRNA in small-cell lung cancer, *J Cell Biochem* (2018).
- [24] X. Ding, S. Zhang, X. Li, C. Feng, Q. Huang, S. Wang, W. Xia, F. Yang, R. Yin, L. Xu, M. Qiu, M. Li and J. Wang, Profiling expression of coding genes, long noncoding RNA, and circular RNA in lung adenocarcinoma by ribosomal RNA-depleted RNA sequencing, *FEBS Open Bio* **8** (2018), 544–555.
- [25] A. Ashouri, V.I. Sayin, J. Van den Eynden, S.X. Singh, T. Papiannakopoulos and E. Larsson, Pan-cancer transcriptomic analysis associates long non-coding RNAs with key mutational driver events, *Nat Commun* **7** (2016), 13197.