

Received: 2018.10.21
Accepted: 2019.01.30
Published: 2019.05.27

Identification of a Combined RNA Prognostic Signature in Adenocarcinoma of the Lung

Authors' Contribution:

Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

C 1 **Si-Yu He***
D 1 **Wen-Jing Xi***
F 1 **Xin Wang**
D 1 **Chao-Han Xu**
F 1 **Liang Cheng**
B 1 **Si-Yao Liu**
C 1 **Qian-Qian Meng**
F 1 **Boyan Li**
E 1 **Yahui Wang**
D 1 **Hong-Bo Shi**
B 2 **Hong-Jiu Wang**
A 1 **Zhen-Zhen Wang**

1 College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang, P.R. China
2 College of Science, Heilongjiang University of Science and Technology, Harbin, Heilongjiang, P.R. China

* Si-Yu He and Wen-Jing Xi contributed equally to this study

Corresponding Authors: Zhen-Zhen Wang, e-mail: wangzz@ems.hrbmu.edu.cn, Hong-Jiu Wang, e-mail: whj1980329@163.com and Hong-Bo Shi, e-mail: shihongbo@ems.hrbmu.edu.cn

Source of support: Grant No.31701159 from the National Science Foundation of China

Background: Adenocarcinoma of the lung is a type of non-small cell lung cancer (NSCLC). Clinical outcome is associated with tumor grade, stage, and subtype. This study aimed to identify RNA expression profiles, including long noncoding RNA (lncRNA), microRNA (miRNA), and mRNA, associated with clinical outcome in adenocarcinoma of the lung using bioinformatics data.




Material/Methods: The miRNA and mRNA expression profiles were downloaded from The Cancer Genome Atlas (TCGA) database, and lncRNA expression profiles were downloaded from The Atlas of Noncoding RNAs in Cancer (TANRIC) database. The independent dataset, the Gene Expression Omnibus (GEO) accession dataset, GSE81089, was used. RNA expression profiles were used to identify comprehensive prognostic RNA signatures based on patient survival time.

Results: From 7,704 lncRNAs, 787 miRNAs, and 28,937 mRNAs of 449 patients, four joint RNA molecular signatures were identified, including RP11-909N17.2, RP11-14N7.2 (lncRNAs), MIR139 (miRNA), KLHDC8B (mRNA). The random forest (RF) classifier was used to test the prediction ability of patient survival risk and showed a good predictive accuracy of 71% and also showed a significant difference in overall survival (log-rank $P=0.0002$; HR, 3.54; 95% CI, 1.74–7.19). The combined RNA signature also showed good performance in the identification of patient survival in the validation and independent datasets.

Conclusions: This study identified four RNA sequences as a prognostic molecular signature in adenocarcinoma of the lung, which may also provide an increased understanding of the molecular mechanisms underlying the pathogenesis of this malignancy.

MeSH Keywords: **Biological Markers • Carcinoma, Non-Small-Cell Lung • MicroRNAs • RNA, Long Noncoding • Survival Analysis**

Full-text PDF: <https://www.medscimonit.com/abstract/index/idArt/913727>

 4560  2  5  48



Background

Worldwide, adenocarcinoma is the most common type of lung cancer and is classified as a types of non-small cell lung cancer (NSCLC). The clinical outcome is associated with tumor grade, stage, and subtype, and metastases may occur before diagnosis leading to reduced patient survival [1]. Therefore, there is a need to identify prognostic biomarkers of adenocarcinoma of the lung to improve treatment planning. In the era of high-throughput genomics, efforts have been made to identify molecular prognostic biomarkers using data on adenocarcinoma of the lung [2–5]. However, there has been some controversy regarding the validity and reproducibility of molecular prognostic biomarkers. Some valid mRNAs and noncoding RNAs have been identified in lung cancer. For example, an eight microRNA (miRNA) signature was shown to be an independent prognostic marker that predicted overall survival (OS), which was based on a study of miRNA expression in lung cancer samples from 373 lung cancer patients and clinical data from The Cancer Genome Atlas (TCGA) [6].

Dysregulation of long noncoding RNA (lncRNA) is associated with the occurrence of adenocarcinoma of the lung, and some lncRNAs have been identified as prognostic molecular biomarkers. A 64 lncRNA molecular prognostic signature was identified that could distinguish between normal lung tissue and adenocarcinoma of the lung using the Affymetrix Human Genome U133 Plus 2.0 microarray [7]. An eight lncRNA molecular prognostic signature and a nine lncRNA molecular relapse-associated signature were identified in adenocarcinoma of the lung using re-annotated Affymetrix array probe sets to the human genome [8,9]. Until recently, most of the mRNAs, miRNAs, and lncRNAs have been identified by single types of data profiles [10–13], there have been few studies that have integrated multiple RNA expression profiles to identify RNA molecular signatures, which still need to be explored further [14,15].

In the present study, the method of combined RNA expression was used to identify prognostic biomarkers in adenocarcinoma of the lung to develop a prognostic model for patient survival. The basis for the identification of combined molecular prognostic biomarkers is based on the finding that if a gene can act as an independent biomarker of prognosis, a set of genes might represent a combined or more representative prognostic effect. Genes expressed in adenocarcinoma of the lung can be individually selected on the basis of fold-change, log-rank test, and patient spectral similarity methods to obtain candidate genes. Univariate and multivariate Cox regression analysis can then be used to identify the combined gene signatures associated with the development of adenocarcinoma of the lung and to identify the gene biomarkers were found. The random forest classification method tests the effectiveness of the classification in terms of patient prognosis. In the present

study, one validation dataset and one independent dataset, the Gene Expression Omnibus (GEO) accession dataset, GSE81089, was used [16]. Combined biomarkers can be used to identify patients with good and poor prognosis, based on clinical factors, including the tumor grade and stage.

Therefore, this study aimed to identify RNA expression profiles, including lncRNA, miRNA, and mRNA, to develop a combined prognostic molecular signature in adenocarcinoma of the lung.

Material and Methods

Patients cohorts with adenocarcinoma of the lung

Data from patients with adenocarcinoma of the lung, including the microRNA (miRNA), and mRNA expression profiles were downloaded from The Cancer Genome Atlas (TCGA) database [17]. Long noncoding RNA (lncRNA) expression profiles were downloaded from The Atlas of Noncoding RNAs in Cancer (TANRIC) database [18]. The data of mRNAs, miRNAs, and lncRNAs with expression values <1 in two-thirds of the sample were excluded from the profile. Finally, 7,704 lncRNAs, 787 miRNAs, 28,937 mRNAs of 449 patients were analyzed.

Identification of the candidate high-risk genes

There was 70% of the patient expression data selected as the training dataset and the test dataset for cross-validation, and the remainder were used as the validation dataset. Genes were identified as candidate high-risk genes that showed significantly different expression levels among the patients with adenocarcinoma of the lung, which affected patient prognosis, and reduced overall patient survival. The patients were divided into two groups, high-risk and low-risk, according to the median candidate gene expression. For each gene, the patients were divided into two groups according to the median expression level in the training dataset and the test dataset, and differential gene expression between these two patient groups was determined according to its fold-change value (fold-change >2, or fold-change <0.5). For each differentially expressed gene, the gene was evaluated for an associated significant survival difference between the two groups by using the log-rank test ($p < 0.05$), to identify the genes for further analysis. Patients were expected to have high heterogeneity for all genes, and the candidate high-risk genes were expected to show significantly different expression levels and associated patient survival times. Therefore, any two patient groups with high gene expression levels were compared. If there was at least one overlap in the number of patients that was >40% of the total number of the patients, then the gene was identified as a candidate high-risk gene.

Construction of the prognostic prediction model based on the combined RNA signature

Cox regression analysis was used to assess the effect of expression of the candidate high-risk genes on overall survival. Univariate Cox regression analysis was used to test the association between mRNA, miRNA, lncRNA expression levels and overall patient survival in the training dataset. Genes with expression levels that were significantly correlated with the overall survival of the patients ($p < 0.05$) were selected. To assess the relative contribution of predictive genes for survival prediction, they were examined by a multivariate Cox regression analysis with overall survival as the independent variable and the significant genes were subsequently analyzed. Based on the expression of these genes, a combined signature was identified to build a prognostic model of risk by the linear combination of the expression levels of predictive genes with the multivariate Cox regression coefficient as the weight. This prognostic model calculated an expression-based risk score for each patient and was used to classify patients into a high-risk group and a low-risk group using the median risk score, and used receiver operating characteristic (ROC) area under the curve (AUC) to evaluate the performance with 10-fold cross-validation using the R package of pROC [19].

Statistical analysis

Kaplan-Meier analysis was used to determine the survival time to predict low-risk and high-risk patients, and the two-sided log-rank test was used to assess the differences between the low-risk group and the high-risk group using the R survival package. The hazard ratio (HR) and 95% confidence interval (CI) were determined using the Cox proportional hazard regression model and multivariate analysis was performed to determine whether the prognostic model was independent of other clinical variables, adjusting for risk score, patient age, tumor grade, and tumor stage. The out-of-bag estimate was used as a method of measuring the prediction error of boosted decision trees, random forests, and other models utilizing bootstrap aggregation of subsamples used for training. Statistical analysis was performed using R software (www.r-project.org) and Bioconductor software (www.bioconductor.org).

Implementation of the random forest method and performance evaluation

The R package random forest (RF) method was used to classify the patients according to the identified combined RNA signature. Two parameters required optimization for a given supervised classification problem, the number of possible variables to divide each node of a tree (m_{try}), and the number of decision trees to construct ($trees$). Trees were maintained at a value of 100, as the resulting error, obtained using out-of-bag estimates, was observed to reach a stable minimum.

The performance of the RF classifier was evaluated by several methods, including cross-validation, out-of-bag errors and ROC curves. Cross-validations were used to determine the classifier errors. Out-of-bag errors made use of the unselected samples in each tree in the forest plot to determine the classifier errors and were shown to be accurate empirically. Besides estimating classifier errors, the area under the receiver operating characteristic curve (AUROC) values was calculated to assess the performance of the RF classifiers.

Functional analysis of risk gene lists

Functional analysis was performed using DAVID version 6.8 functional annotation (<https://david.ncifcrf.gov/>) to examine the degree of enrichment of the high-risk gene set in Gene Ontology (GO) terms [20]. GO terms with a q-value false discovery rate-adjusted p-value (q-value) were calculated using the Benjamini-Hochberg procedure. When the q-value is < 0.01 , the GO functional term was considered to represent significant enrichment by using the whole human genome as a reference. The high-risk genes included the mRNA of the combined signature, the target genes of miRNA in the combined signature, and the protein-coding genes that were positively or negatively correlated with lncRNA in the combined signature. The miRNA target genes were obtained from the miRBase database [21]. The protein-coding genes that were positively or negatively correlated with lncRNA were identified by calculating the correlation coefficients.

The construction of the combined signature associated gene network

The target genes of MIR139 were obtained from the StarBase database [22]. The data of the direct and indirect interaction protein were downloaded with the KLHDC8B gene from the STRING biological protein-protein database [23]. The miRNA and its target genes, and protein-protein interaction relationship were used to build an integrated network. The node in the network represented the miRNA and genes, and the edge in the network represented the regulation relationship or protein-protein interaction relationship. The lncRNA target gene data were investigated in the lncRNA2Target version 2.0 database of differentially expressed genes, and there were no lncRNA target genes data for RP11-909N17.2 and RP11-14N7.2 [24].

Results

Identification of integrated RNA sequences as a prognostic signature in adenocarcinoma of the lung

There were 449 patients with adenocarcinoma of the lung who had expression information for combined mRNA,

microRNA (miRNA), long noncoding RNA (lncRNA) simultaneously from The Cancer Genome Atlas (TCGA) database. An integrated gene expression profile was established that included 7,704 lncRNAs, 737 miRNAs and 28,936 mRNAs. Random selection of 70% of the integration profiles of the patients were included in the training and test datasets to identify RNA biomarkers associated with survival in patients with adenocarcinoma of the lung, the other 30% of the integration profile from the patients was selected as a validation dataset, and the patient cohorts from the Gene Expression Omnibus (GEO) accession dataset, GSE81089, was selected as an additional independent dataset [16].

A gene filtration method was used to identify the individual candidate biomarker genes in the integration gene expression profile associated with survival time in patients with adenocarcinoma of the lung to identify the potential prognostic RNA biomarkers. For each gene, patients were classified into two groups according to the median gene expression and this gene was tested if it showed significantly different expression between the two patient groups (fold-change >2 or fold-change <0.5), and differentially expressed mRNAs, lncRNAs, and miRNAs were identified. The log-rank test was performed for each differentially expressed RNA between the two patient groups for their associations with overall survival using $p < 0.01$ as a criterion and identified mRNAs, lncRNAs, miRNAs which were differentially expressed and showed a significant difference between the two patient groups in terms of survival time.

Candidate biomarkers were further identified that were differentially expressed and in which the overlap between the patient groups was >0.3. There were 100 mRNAs, 33 lncRNAs, and 6 miRNAs as the candidate biomarkers that were finally identified. The dot map of those candidate biomarker genes is shown in Figure 1A. The candidate biomarker genes showed differential gene expression levels and survival in the two patient groups and were considered as potential candidate prognostic biomarkers for in patients with adenocarcinoma of the lung.

Identification of the combined prognostic biomarker and the prognostic model

Univariate Cox proportional hazard regression analysis was performed for each candidate prognostic biomarker and identified eight significant RNAs. Multivariate Cox regression analysis for these used overall survival as an independent variable and identified four genes including RP11-909N17.2, RP11-14N7.2 (lncRNAs), MIR139 (miRNA), and KLHDC8B (mRNA), which were significantly associated with overall survival in patients with adenocarcinoma of the lung ($p < 0.05$) (Table 1).

Therefore, a prognostic model was proposed for overall survival with the risk scoring method, which integrated the four

RNAs expression levels and their relative contributions from multivariate analysis. The formula for the prognostic model was defined as follows: Risk score = $(0.041 \times \text{expression value of RP11-909N17.2}) + (0.03 \times \text{expression value of RP11-14N7.2}) + (-0.008 \times \text{expression value of MIR139}) + (-0.053 \times \text{expression value of KLHDC8B})$.

The risk scores were calculated based on the formula of the prognostic model and divided patients into a low-risk group and a high-risk group according to the median risk score. For this prognostic model, a good area under the curve (AUC) value of 0.875 was obtained based on 10-fold cross-validation in the training dataset, as shown in Figure 1B. In the low-risk group ($n=157$), the median survival time was 1.54 years, which was a significantly improved when compared with the high-risk group ($n=158$), and the median survival time was 2.03 years ($p=2e-5$, log-rank test) for all patients in the training dataset (Figure 1C).

The results showed that the four RNA signature had a better prognostic performance for the prediction of outcome in patients with adenocarcinoma of the lung. The patients with high expression levels of RP11-909N17.2 and RP11-14N7.2 and low expression levels of the KLHDC8B gene and the MIR139 gene had high-risk scores, as shown in Figure 1D. The risk score distribution of RNA expression and survival status of 315 patients in the training cohort were ranked according to the four-RNA signature risk score, as shown in Figure 1D.

Validation of the combined RNA prognostic biomarker using Kaplan-Meier survival and building of the classifier by those biomarkers

The prognostic ability of the combined RNA biomarker signature was evaluated in the validation cohort of 134 patients from The Cancer Genome Atlas (TCGA) database. In the validation cohort, 134 patients were divided into a low-risk group ($n=58$) and a high-risk group ($n=76$) by using the risk score model described above and the cut off value derived from the training cohort. The low-risk patient group and the high-risk patient group showed a significant difference in overall survival between two patient groups in the validation cohort (log-rank test, $p=0.031$) by using Kaplan-Meier analysis, and it was consistent with the findings from the training cohort. The median overall survival in the low-risk patient group was significantly greater compared with that in the high-risk group, with a median overall survival of 2.05 years versus 1.49 years, as shown in Figure 2A.

The robustness of the combined prognostic biomarker for predicting survival in patients with adenocarcinoma of the lung was tested in the independent cohort of 84 patients in the Gene Expression Omnibus (GEO) independent accession dataset,

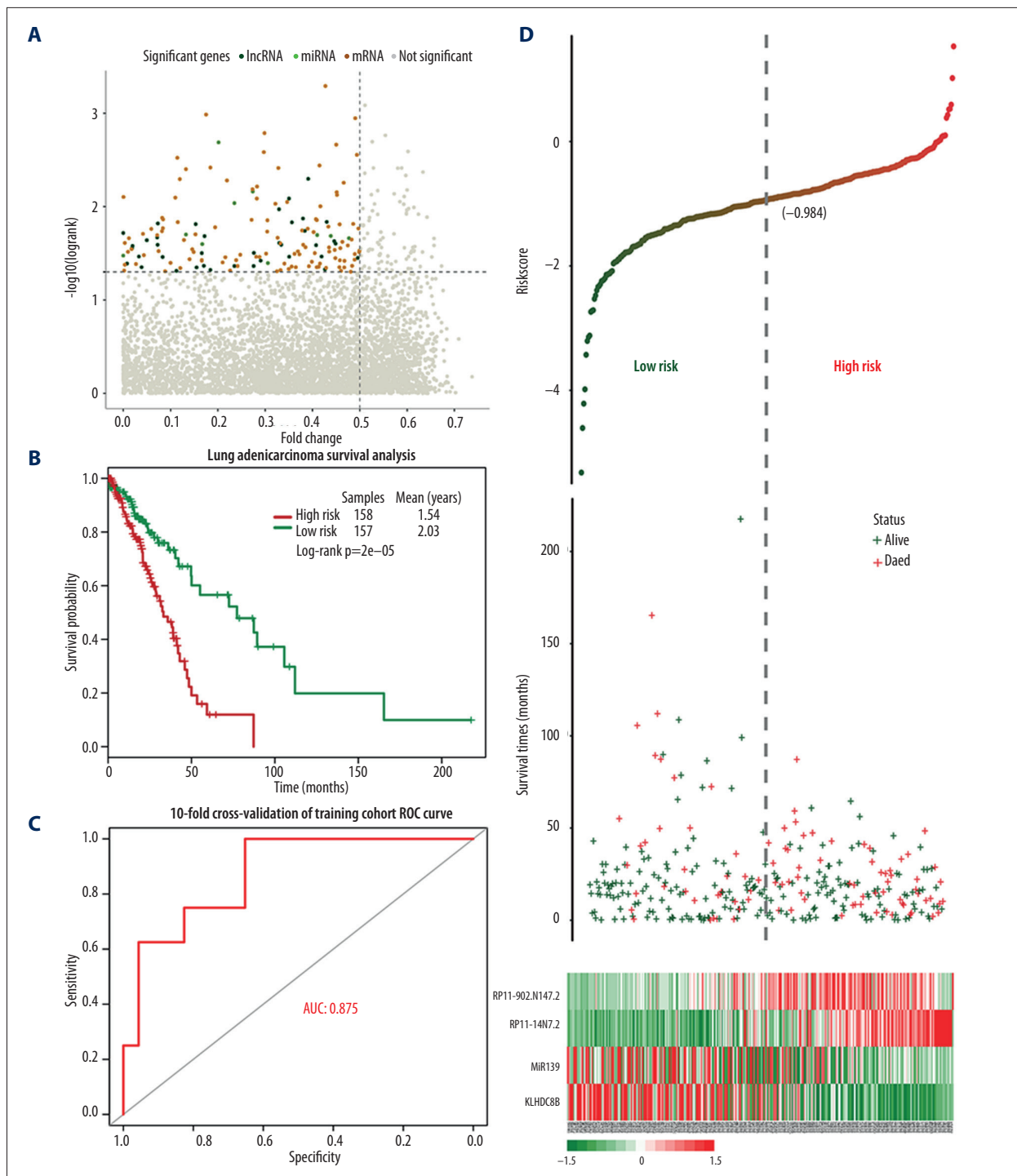


Figure 1. Identification of a combined RNA prognostic signature in adenocarcinoma of the lung. **(A)** The combined signature genes (fold-change <0.05, log-rank $p<0.05$), dark green dots represent the long noncoding RNAs (lncRNAs), Light green dots represent the microRNAs (miRNAs), yellow dots represent mRNA, gray dots represent genes with gene expression levels and survival times with no significant differences between the two patient groups. **(B)** Kaplan-Meier survival plots of patients with low-grade and high-grade adenocarcinoma of the lung in the training dataset. **(C)** Receiver operating characteristic (ROC) curve for the prognostic risk model. The area under the curve (AUC) value of 0.875 showed the good performance of the patient risk prediction. **(D)** The ranking of patient risk scores and survival time, and the expression pattern of prognostic RNAs.

Table 1. General information about the combined molecular prognostic biomarkers.

Gene name	Ensemble ID	Gene type	Chromosome (GRCh38)	Hazard ratio (HR)	Coefficient	P-value
RP11-909N17.2	ENSG00000253931	Antisense RNA	Chromosome 8: 143, 412, 749-143, 417, 054	1.04	4.14E-02	1.00E-02
RP11-14N7.2	ENSG00000232527	lncRNA	Chromosome 1: 144, 227, 030-144, 250, 288	1.03	3.01E-02	7.09E-04
MIR139	ENSG00000272036	miRNA	Chromosome 11: 72, 615, 063-72, 615, 130	0.99	-8.04E-03	4.15E-02
KLHDC8B	ENSG00000185909	Protein coding	Chromosome 3: 49, 171, 611-49, 176, 486	0.95	-5.25E-02	6.35E-04

GSE81089, and obtained similar risk stratification results. As with the training and validation cohorts, the combined prognostic biomarkers were able to classify 84 independent cohorts into low-risk patients (n=26) and high-risk patients (n=58) with significantly different survival times (log-rank test, $p=0.027$) (Figure 2B). The median overall survival in the low-risk patient group compared with the high-risk patient group was significantly greater (median 3.86 years versus 2.98 years). The risk scores for the combined prognostic biomarkers in the validation and independent cohorts and the distribution of survival time and expression levels are shown in Figure 2C and 2D, which shows the similar results observed in the training cohort. The AUC of the prognostic model in the validation and independent dataset were 0.764 and 0.724, as shown in Figure 2E and 2F.

The ability of a single signature to identify patient risk was determined. The median value of each signature gene was used to divide the patients into two groups. Kaplan-Meier survival curve analysis and the log-rank test were used to identify significant survival differences between patient groups. For the training dataset, there was significantly different survival between the patient groups divided by each signature: RP11-909N17.2, $p=0.02791$; RP11-14B7.2, $p=0.04341$; MIR139, $p=0.04023$; KLHDC8B, $p=0.0444$. However, the p-values were less than the p-value of the patient group when divided by the combined signature ($p=2e-05$) as shown in Figure 3A. For the validation dataset, differences in overall survival were not significant between the patient groups when divided by each signature. For independent data, survival differences just reached statistical significance in the patient group divided by KLHDC8B ($p=0.035$), as shown in Figure 3B.

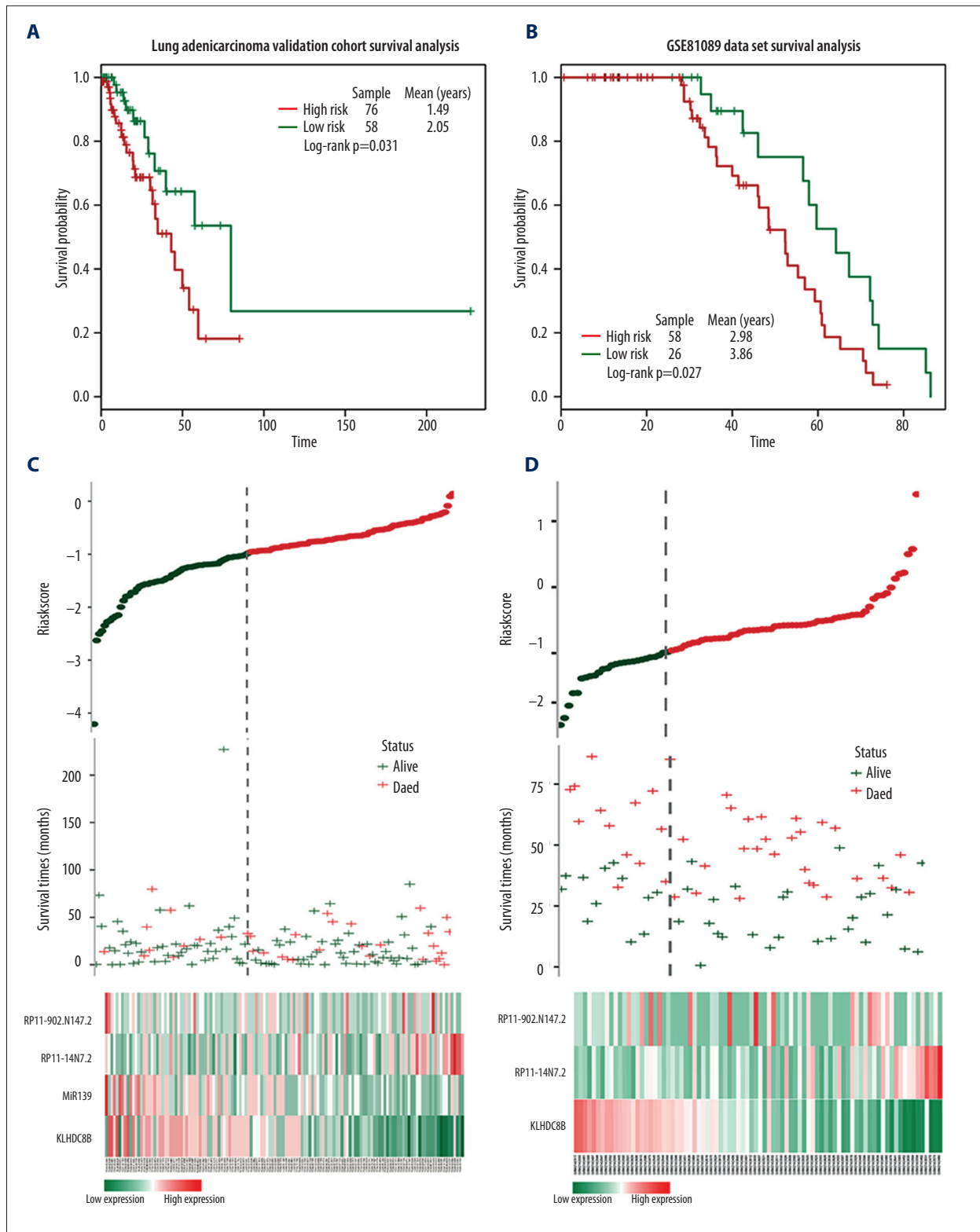
There were significant survival differences between the patient groups divided by age, tumor stage and residual tumor in the training dataset, but the p-values of age ($p=0.043558$) and residual tumor ($p=0.00955$) were less than the p-value of the patient group divided by the combined signature. The p-value of the tumor stage was similar to the p-value of the patient group divided by the combined signature, as shown

in Figure 3C, 3D. Comparison of the results for differences in overall patient survival between the patient groups showed an improved prediction of patient prognosis using the combined RNA expression profile signature when compared with the single signature and other clinical prognostic factors.

In this study, the degree of overlap was evaluated between the patient group with different stage and risk and the detail of the overlap is shown as a Venn diagram in Figure 3E. The overlap between patients with different stages of adenocarcinoma of the lung and patients with different risk was evaluated using a hypergeometric test in the training dataset, the validation dataset, and the independent dataset. The results showed that the patients with stage I and II adenocarcinoma of the lung enriched the low-risk patient group, including the training dataset ($p=0.0015$), the validation dataset ($p=0.038$), and the independent dataset ($p=0.039$). Patients with stage III and IV adenocarcinoma of the lung enriched the high-risk patient group, including the training dataset ($p=0.0007$), the validation dataset ($p=0.038$) and the independent dataset ($p=0.039$), as shown in Figure 3E.

Comparison of the clinical and pathological characteristics of patients in the high-risk and low-risk patient group was divided by the combined signature in the training and validation dataset. The clinical and pathological characteristics of male gender, tumor recurrence, anatomic neoplasm subdivision of left and right upper and middle lobes of the lung showed consistent results for the degree of patient risk. Patients with the characteristics of male gender, tumor recurrence, tumor location in the left upper lobe and right middle lobe, and exon 19 deletion of the EGFR gene, were significantly associated with high risk. Patients with centrally located lung tumors were more likely to have low risk, and patients with increased tumor stage (N0 to N2, T1 to T2, M0 to M1) were more likely to have a higher risk.

The random forest (RF) algorithm was used to test the ability of the combined prognostic biomarker to classify patients into



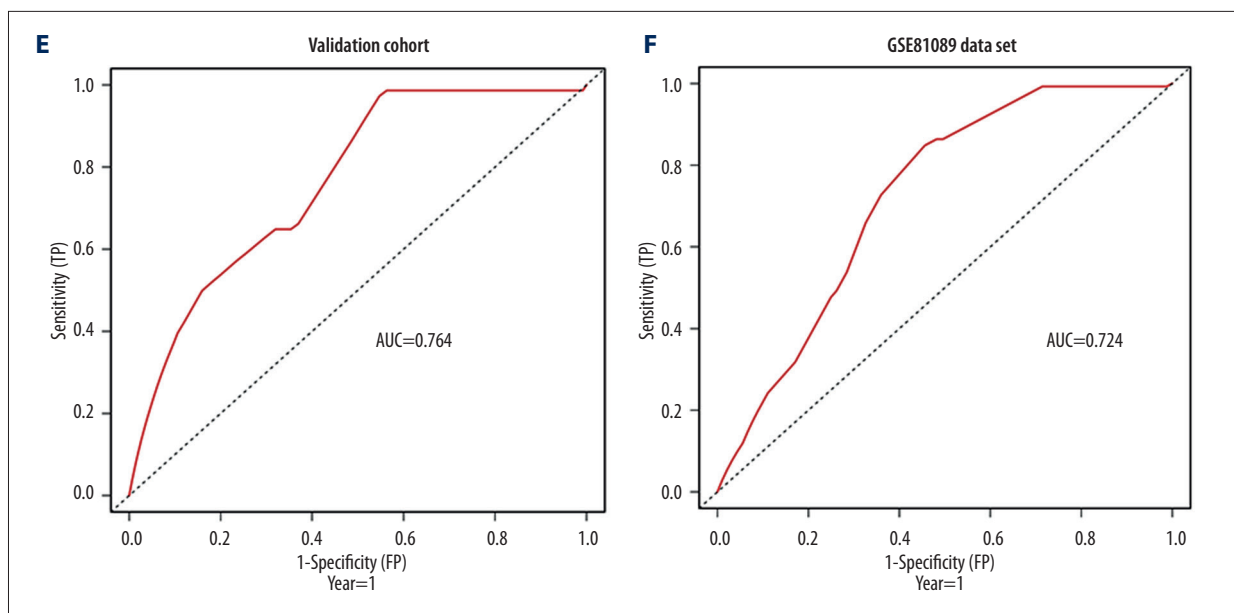
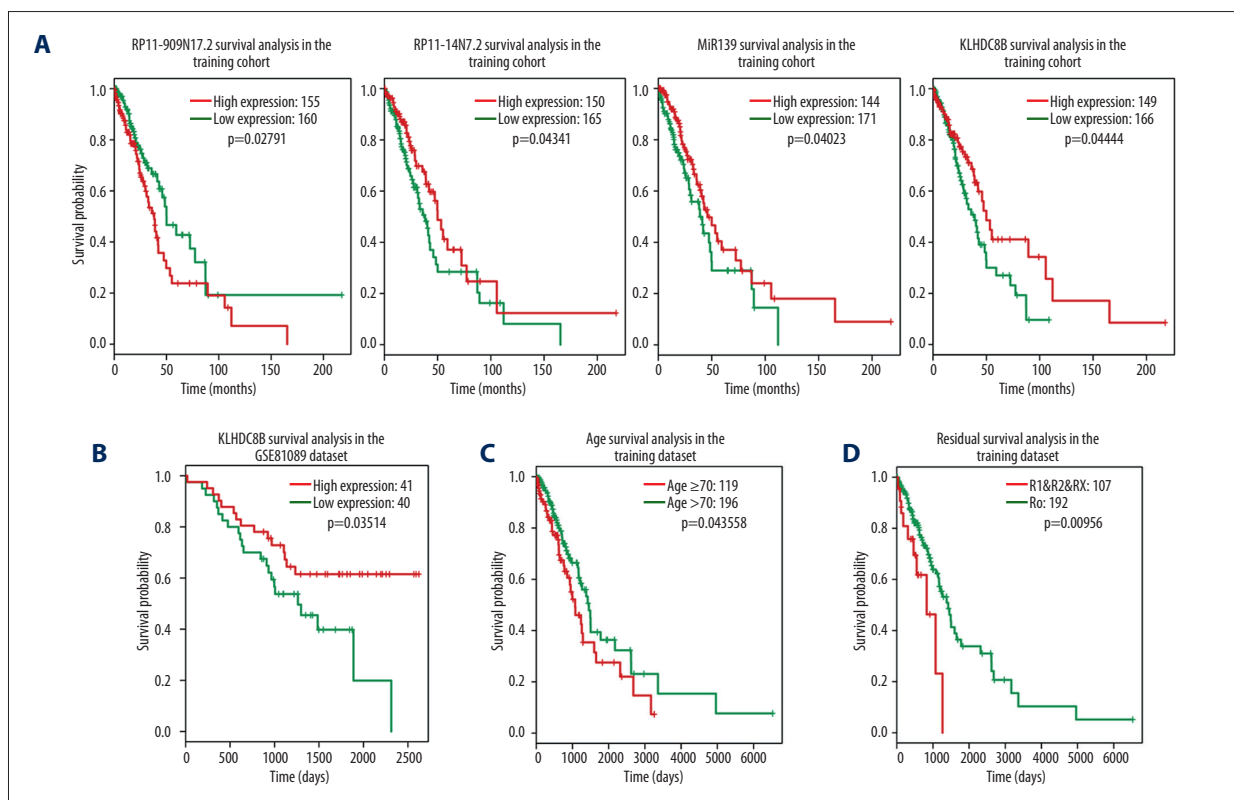


Figure 2. The value of the combined prognostic signature in the validation and additional independent datasets. (A) Kaplan-Meier survival curves of total survival time between high-risk and low-risk patients in the validation dataset of The Cancer Genome Atlas (TCGA). (B) Kaplan-Meier survival curves of total survival time between high-risk and low-risk patients in the independent dataset of GSE81089. (C) The rank of the patient risk score and survival time, and the expression pattern of prognostic RNAs in the validation dataset of TCGA. (D) The rank of the patient risk score and survival time, and the expression pattern of prognostic RNAs in the independent dataset of GSE81089. (E) The receiver operating characteristic (ROC) curve of the prognostic model for lung adenocarcinoma. The area under the curve (AUC) of the prognostic model was 0.764 in the validation dataset. (F) The ROC curve of the prognostic model for adenocarcinoma of the lung. The AUC of the prognostic model was 0.724 in the independent dataset.



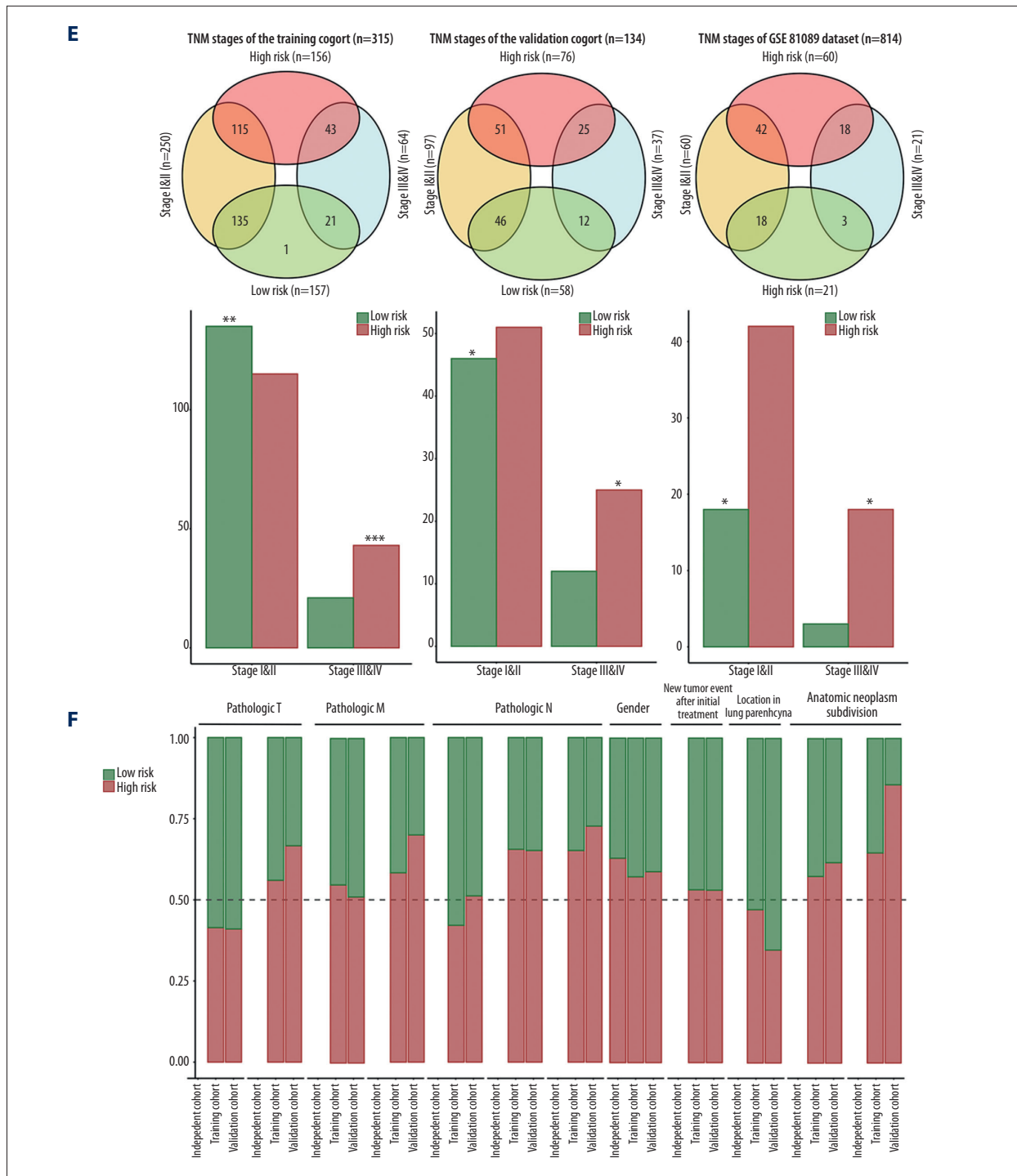


Figure 3. Comparison of the ability of the single, combined signature, and clinical prognostic factors for identifying patient risk. **(A)** Kaplan-Meier survival curve analysis and log-rank test for the single signature in the training dataset. **(B)** Kaplan-Meier survival curve analysis and log-rank test for the KLHC8B gene in the independent dataset. **(C)** Kaplan-Meier survival curve analysis and log-rank test for the age in the training dataset. **(D)** Kaplan-Meier survival curve analysis and log-rank test for residual tumor in the training dataset. **(E)** Venn diagram showed the overlap between the patient group with different stage and risk was shown in all datasets. The significance test of overlap between the patients with different stage and the patients with different risk using a hypergeometric test in the training, validation, and independent datasets. **(F)** Comparison of the clinical and pathological characteristics of patients between the high-risk and low-risk patient groups that were divided by the combined signature in all datasets.

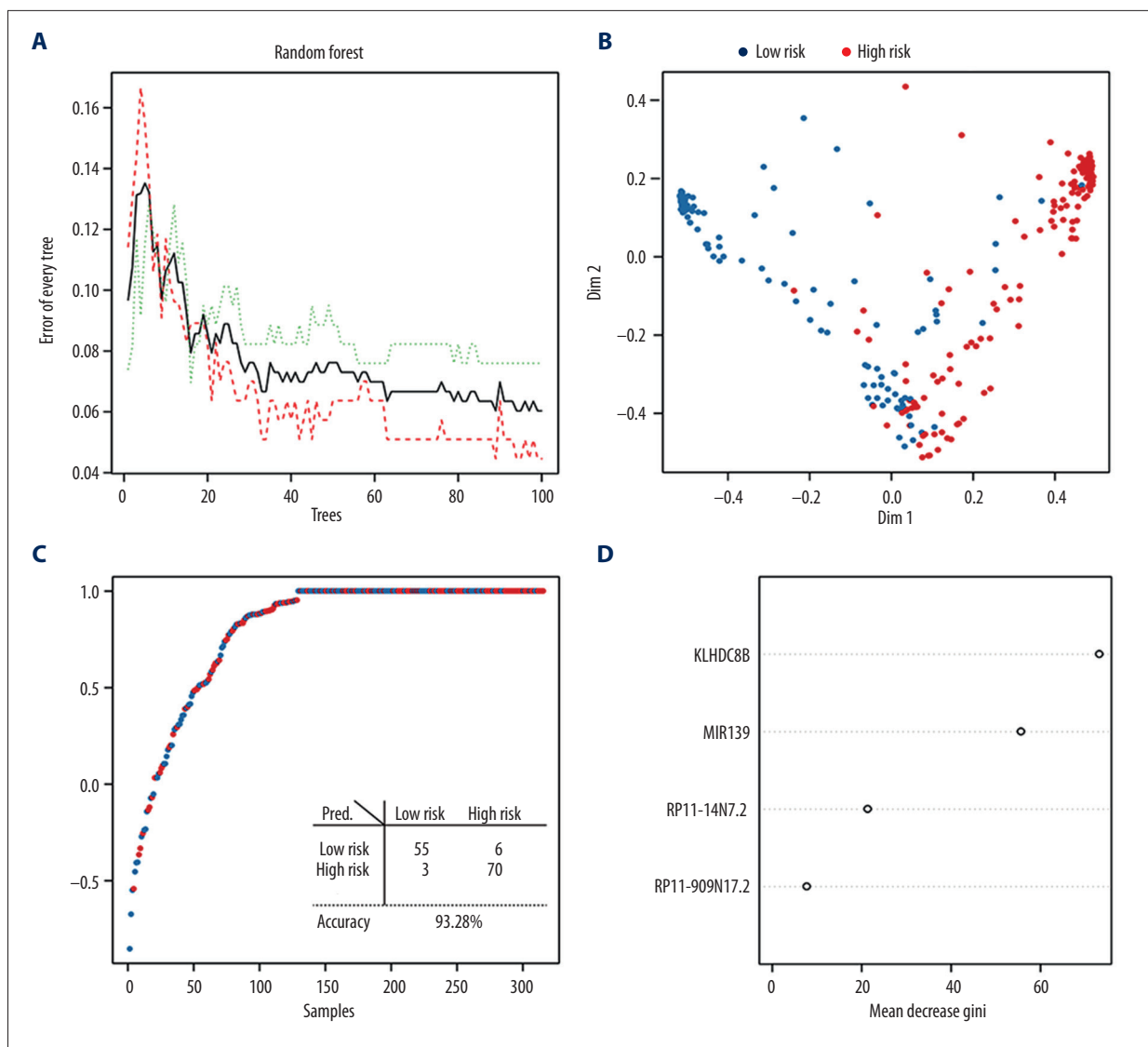


Figure 4. Construction of the classifier using the combined molecular biomarkers with the random forest method. **(A)** Comparison of the training dataset error rates for the random forest method as the number of trees increased. **(B)** The distances between patients in the training dataset in the first two dimensions. Minimum curvilinear embedding from the proximity matrix of the random forest high and low-risk classifier. Each circle represents a patient’s sample with its labeled risk: low risk (blue) or high risk (red). **(C)** Predictive accuracy of the validation dataset. **(D)** Variable importance as given by the mean decrease in the Gini coefficient, which represents the contribution of each variable to the homogeneity of the results of the random forest.

high-risk and low-risk patient groups. The forest plots were created at the value of 100 using out-of-bag samples, and the resulting error was observed to reach a stable minimum, which was equal to 6.03%, with the optimal mtry parameter selected as 4. The values of the Gini index of KLHDC8B, MIR139, RP11-14N7.2, and RP11-909N17.2 were equal to 72.94, 55.42, 21.12, and 7.49, respectively in the training dataset, and the accuracy rate of the RF classifier was 93.28% for validation dataset, as shown in Figure 4. The high values of the Gini index and classification accuracy showed that the combined prognostic

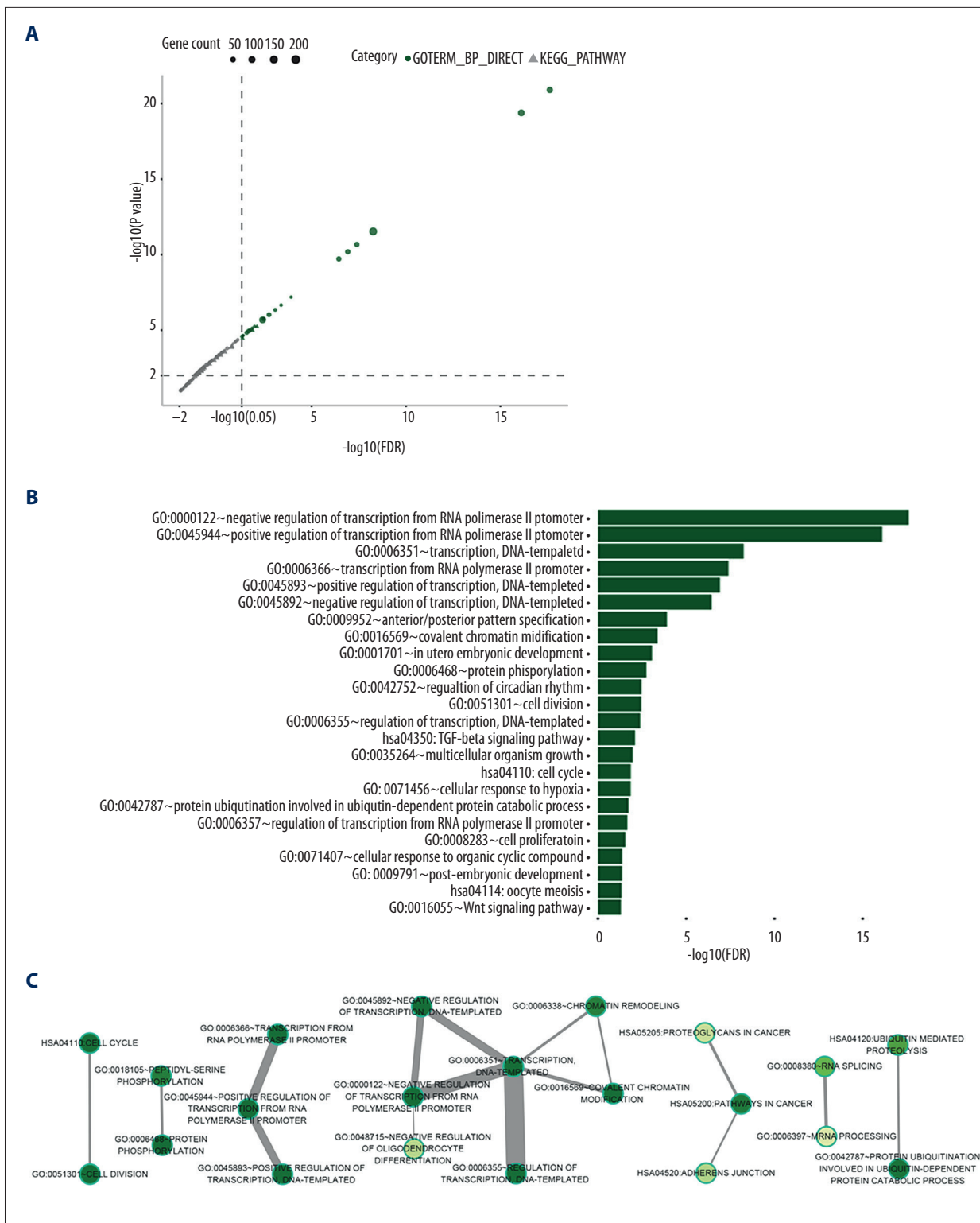
biomarker was able to group the patients into low-risk and high-risk patient groups. These findings supported the potential role of RNA sequences as a prognostic molecular signature in patients with adenocarcinoma of the lung.

Independence of the prognostic role of the combined RNA signature and other clinical variables

The independent prognostic ability of the combined RNA signature in patients with adenocarcinoma of the lung was tested

Table 2. Univariate and multivariate Cox regression analysis of overall survival in all datasets.

Variables	Univariate analysis			Multivariate analysis		
	Hazard ratio (HR)	95% CI of HR	P-value	Hazard ratio (HR)	95% CI of HR	P-value
Training cohort (n=315)						
Gene marker risk score						
High		1 (reference)			1 (reference)	
Low	0.39	0.25–0.61	3.89E-05	0.45	0.27–0.76	2.92e-03
Age	1.03	1–0.05	0.037	1.03	1–1.06	5.23e-02
Residual tumor						
R0		1 (reference)			1 (reference)	
R1	2.22	0.797–6.186	0.1272	1.71	0.59–4.93	0.32
R2	27.65	3.440–222.226	0.0018	6.23	0.72–53.93	9.67e-02
RX	2.17	0.859–5.499	0.1012	2.22	0.85–5.8	0.1
Pathologic stage						
I		1 (reference)			1 (reference)	
II	1.59	0.96–2.64	7.01e-02	1.22	0.68–2.18	0.51
III	3.97	2.4–6.57	8.43e-08	3.41	1.84–6.3	9.17e-05
IV	2.32	0.9–5.96	8.02e-02	1.5	0.51–4.42	0.47
Validation cohort (n=134)						
Gene marker risk score						
High		1 (reference)			1 (reference)	
Low	0.46	0.23–0.95	3.5e-02	0.47	0.19–1.18	0.11
Age	0.99	0.96–1.02	0.49	0.96	0.91–1.01	8.16e-02
Residual tumor						
R0		1 (reference)			1 (reference)	
R1	2.57	0.59–11.22	0.21	0.83	0.18–3.97	0.818
R2	0.11	2.38–54.31	2.3e-03	40.99	3.56–471.59	3E-03
RX	4.58e-08	0–Inf	0.998	1e-07	0–Inf	0.998
Pathologic stage						
I		1 (reference)			1 (reference)	
II	6.94	2.47–19.49	2.33e-04	6.54	1.89–22.64	3e-03
III	5.22	2.18–12.5	2.03e-04	6.64	2.09–21.08	1e-03
IV	4.93	1.57–15.52	6.35e-03	0.9	0.14–5.95	0.913
GSE81089 (n=84)						
Gene marker risk score						
High		1 (reference)			1 (reference)	
Low	0.47	0.23–0.93	0.03	0.41	0.18–0.91	0.03
Age	0.99	0.96–1.03	0.82	1	0.97–1.05	0.78
Pathologic stage						
I		1 (reference)			1 (reference)	
II	1.62	0.05–0.56	4e-03	0.13	0.03–0.51	3e-03
III	0.83	0.32–2.27	0.71	1.04	0.38–2.82	0.94
IV	2.3e-07	0–Inf	1	1.9e-07	0–Inf	1



combined signature. In previous studies, target genes for has-miR-139-5P and related genes with lncRNA RP11-909N17.2, RP11-14N7.2 were obtained from the StarBase version 2.0 database, and the LNCipedia database for annotated human lncRNA transcript sequences, respectively [29]. The gene set including targets gene and related genes were used to perform the functional enrichment analysis using the DAVID software and Enrichment Map Cytoscape software. As shown in Figure 5A–5C, the results of functional enrichment analysis showed that the genes associated with the combined signature were involved in negative regulation of transcription from RNA polymerase II promoter, positive regulation of transcription from RNA polymerase II promoter, transcription, and DNA template, which were associated with the occurrence and development of lung cancer.

The relationships between MIR139 and its target genes and between KLHDC8B and its interacting proteins were evaluated with the gene expression information to build an integrated network, as shown in Figure 5D. The network results showed that the expression of the signature genes of MIR139 and KLHDC8B were significantly down-regulated in the high-risk group.

The related gene expression of VAV3, UBE2C, and UBE2S were significantly upregulated in the high-risk group, and the related genes expression of ALK, KLK3 was significantly down-regulated in the high-risk group. Previously published studies have identified the gene for ubiquitin-conjugating enzyme E2C (UBE2C) as a biomarker for survival and its expression level of mRNA and protein has previously been shown to be upregulated in lung cancer and to increase progressively in lung tumors. Patients with stage I adenocarcinoma of the lung with increased expression levels of UBE2C have previously been reported to show significantly reduced overall survival and progression-free survival than patients with lower expression level [30]. Expression levels of UBE2S mRNA and protein levels were have previously been shown to be upregulated and associated with reduced prognosis in patients with lung cancer, and in preclinical studies, inhibition of expression of UBES2 has been shown to reduce cell proliferation [31]. In the present study, as shown in Figure 5D, mRNA expression of UBE2C and UBE2S showed significant upregulation in the high-risk patient group, which is consistent with previously published findings. The genes UBE2C and UBE2S have been reported to be indirectly regulated by MIR139, and so this result suggested that the down-regulation of expression of MIR139 could cause the upregulation of UBE2C and UBE2S. MIR139 as the regulator of UBE2C and UBE2S could be a potential target for adenocarcinoma of the lung, as AKL was previously reported to be a validated therapeutic target of lung cancer that showed significant down-regulation in high-risk patients groups [32].

The findings of this study identified three important molecular biomarker pathways from the node MIR139 to node KLHDC8B in the network as follows: MIR139-VAV3-SHC1-CSF1R-KLHDC8B, MIR139-UBE2C-UBE2D1-KLHK2-KLHDC8B; MIR139-UBE2S-RPS27A-REL-KLHDC8B. Most of the genes in the three pathways have been previously reported as being biomarkers for lung cancer [33]. These findings may have implications not only for the prognosis of adenocarcinoma of the lung but might provide insights into its pathogenesis.

Discussion

There remains a need for improved prognostic biomarkers that can guide treatment options for patients with adenocarcinoma of the lung. Therefore, this study aimed to identify RNA expression profiles, including long noncoding RNA (lncRNA), microRNA (miRNA), and mRNA, associated with clinical outcome in adenocarcinoma of the lung using bioinformatics data. The approaches used in this study are supported by previous studies that have shown that molecular biomarkers for clinical outcome can be used for patients with malignancy [34–39]. Previous studies have identified the cancer-related genes that were verified in the present study as the candidate prognostic genes and integrated them into expression and clinical data [40–45]. Other studies have previously clustered the whole gene expression into two groups to identify the association between abnormal regulation of genes as molecular prognostic biomarkers [46–48]. However, few studies have evaluated the regulation of these genes in terms of patient overall survival or provided an integrated RNA signature or profile as a comprehensive prognostic biomarker in patients with adenocarcinoma of the lung.

The findings of the present study demonstrated the use of a novel method for the identification of four RNA sequences as a prognostic molecular signature in adenocarcinoma of the lung by integrating mRNA, miRNA, and lncRNA of patients from The Cancer Genome Atlas (TCGA) database. Four genes were identified that showed different expression in two groups, a high-risk and a low-risk group, which were significantly different in terms of survival time by using individual gene filtering and Cox regression analysis. The combined signature included two mRNAs, one miRNA, and one lncRNA, and this study showed that one gene was correlated with improved patient survival and two genes were correlated with poor patient survival. A prognostic prediction model was created based on the combined RNA prognostic signature to predict the prognosis of patients and also used one validation dataset and one independent dataset to examine the accuracy of the predictive model. The combined signature was shown to be useful in terms of its predictive role in the validation and independent dataset. Furthermore, each gene in the combined signature

was studied for its ability to predict patient prognosis. The results showed that these genes were independent of the other clinicopathological factors, including patient age, tumor stage, and tumor grade.

Conclusions

This study identified four RNA sequences as a prognostic molecular signature in adenocarcinoma of the lung, which may also provide an increased understanding of the molecular mechanisms underlying the pathogenesis of this

malignancy. The combined RNA prognostic signature was identified by integrating RNA profiles in adenocarcinoma of the lung. Bioinformatics data analysis using this method might lead to the identification of a potential prognostic molecular biomarker based on combined RNA signatures. Further studies using this approach might aid clinical diagnosis, prognosis, and improve the understanding of the pathogenesis of complex diseases such as adenocarcinoma of the lung.

Conflict of interest

None.

References:

1. Cancer Genome Atlas Research Network: Author correction: Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 2018; 559: E12
2. Verri C, Borzi C, Holscher T et al: Mutational profile from targeted NGS predicts survival in LDCT screening-detected lung cancers. *J Thoracic Oncol*, 2017; 12: 922–31
3. Wong SQ, Fellowes A, Doig K et al: Assessing the clinical value of targeted massively parallel sequencing in a longitudinal, prospective population-based study of cancer patients. *Br J Cancer*, 2015; 112: 1411–20
4. Chen B, Zhang R, Gan Y et al: Development and clinical application of radiomics in lung cancer. *Radiat Oncol*, 2017; 12: 154
5. Hamblin A, Wordworth S, Fermont JM et al: Clinical applicability and cost of a 46-gene panel for genomic analysis of solid tumours: Retrospective validation and prospective audit in the UK National Health Service. *PLoS Med*, 2017; 14: e1002230
6. Li X, Shi Y, Yin Z et al: An eight-miRNA signature as a potential biomarker for predicting survival in lung adenocarcinoma. *J Transl Med*, 2014; 12: 159
7. Yu H, Xu Q, Liu F et al: Identification and validation of long noncoding RNA biomarkers in human non-small-cell lung carcinomas. *J Thoracic Oncol*, 2015; 10: 645–54
8. Zhou M, Guo M, He D et al: A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J Transl Med*, 2015; 13: 231
9. Zhou M, Xu W, Yue X et al: Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma. *Oncotarget*, 2016; 7: 29720–38
10. Zhao J, Li L, Wang Q et al: CircRNA expression profile in early-stage lung adenocarcinoma patients. *Cell Physiol Biochem*, 2017; 44: 2138–46
11. Zhao K, Li Z, Tian H: Twenty-gene-based prognostic model predicts lung adenocarcinoma survival. *Onco Targets Ther*, 2018; 11: 3415–24
12. Zhao S, Gao X, Zang S et al: MicroRNA-383-5p acts as a prognostic marker and inhibitor of cell proliferation in lung adenocarcinoma by cancerous inhibitor of protein phosphatase 2A. *Oncology Lett*, 2017; 14: 3573–79
13. Zhou W, Chen X, Hu Q et al: Galectin-3 activates TLR4/NF-kappaB signaling to promote lung adenocarcinoma cell proliferation through activating lncRNA-NEAT1 expression. *BMC Cancer*, 2018; 18: 580
14. Chen M, Zhao H, Lind SB, Pettersson U: Data on the expression of cellular lncRNAs in human adenovirus infected cells. *Data Brief*, 2016; 8: 1263–79
15. Dong L, Xia J, Zhang J et al: Long-term progression-free survival in an advanced lung adenocarcinoma patient harboring EZR-ROS1 rearrangement: A case report. *BMC Pulm Med*, 2018; 18: 13
16. Mezheyeuski A, Bergslund CH, Backman M et al: Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify lung cancer patients. *J Pathol*, 2018; 244: 421–31
17. Chandran UR, Medvedeva OP, Barmada MM et al: TCGA expedition: A data acquisition and management system for TCGA data. *PLoS One*, 2016; 11: e0165395
18. Li J, Han L, Roebuck P et al: TANRIC: An interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res*, 2015; 75: 3728–37
19. Simon RM, Subramanian J, Li MC, Menezes S: Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief Bioinform*, 2011; 12: 203–14
20. Huang DW, Sherman BT, Tan Q et al: The DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*, 2007; 8: R183
21. Kozomara A, Griffiths-Jones S: miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, 2014; 42: D68–73
22. Li JH, Liu S, Zhou H et al: StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*, 2014; 42: D92–97
23. Szklarczyk D, Morris J H, Cook H et al: The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*, 2017; 45: D362–68
24. Cheng L, Wang P, Tian R et al: LncRNA2Target v2.0: A comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res*, 2019; 47(D1): D140–44
25. Wang Y, Chen W, Chen J et al: LncRNA expression profiles of EGFR exon 19 deletions in lung adenocarcinoma ascertained by using microarray analysis. *Med Oncol*, 2014; 31: 137
26. Edmonds MD, Eischen CM: Differences in miRNA expression in early stage lung adenocarcinomas that did and did not relapse. *PLoS One*, 2014; 9: e101802
27. Xing L, Todd NW, Yu L et al: Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Mod Pathol*, 2010; 23: 1157–64
28. Zhang H, Sun Z, Yu L, Sun J: MiR-139-5p inhibits proliferation and promotes apoptosis of human airway smooth muscle cells by downregulating the Brg1 gene. *Respir Physiol Neurobiol*, 2017; 246: 9–16
29. Volders PJ, Helsens K, Wang X et al: LNCipedia: A database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res*, 2013; 41: D246–51
30. 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: 702–11
31. Liu Z, Xu L: UBE2S promotes the proliferation and survival of human lung adenocarcinoma cells. *BMB Rep*, 2018; 51: 642–47
32. Shaw AT, Engelman JA: ALK in lung cancer: Past, present, and future. *J Clin Oncol*, 2013; 31: 1105–11
33. Lorenzo-Martin LF, Citterio C, Menacho-Marquez M et al: Vav proteins maintain epithelial traits in breast cancer cells using miR-200c-dependent and independent mechanisms. *Oncogene*, 2019; 38(2): 209–27
34. Xie L, Yao Z, Zhang Y et al: Deep RNA sequencing reveals the dynamic regulation of miRNA, lncRNAs, and mRNAs in osteosarcoma tumorigenesis and pulmonary metastasis. *Cell Death Dis*, 2018; 9: 772
35. Peng R, Liu Y, Cai Z et al: Characterization and analysis of whole transcriptome of giant panda spleens: Implying critical roles of long non-coding RNAs in immunity. *Cell Physiol Biochem*, 2018; 46: 1065–77
36. Lin J, Xia J, Zhang K, Yang Q: Genome-wide profiling of chicken dendritic cell response to infectious bursal disease. *BMC Genomics*, 2016; 17: 878

37. Dai F, Mei L, Meng S et al: The global expression profiling in esophageal squamous cell carcinoma. *Genomics*, 2017; 109: 241–50
38. Ning P, Wu Z, Hu A et al: Integrated genomic analyses of lung squamous cell carcinoma for identification of a possible competitive endogenous RNA network by means of TCGA datasets. *Peer J*, 2018; 6: e4254
39. Sui J, Li YH, Zhang YQ et al: Integrated analysis of long non-coding RNA associated ceRNA network reveals potential lncRNA biomarkers in human lung adenocarcinoma. *Int J Oncol*, 2016; 49: 2023–36
40. Svoboda M, Meshcheryakova A, Heinze G et al: AID/APOBEC-network reconstruction identifies pathways associated with survival in ovarian cancer. *BMC Genomics*, 2016; 17: 643
41. Cockburn JG, Hallett RM, Gillgrass AE et al: The effects of lymph node status on predicting outcome in ER+/HER2– tamoxifen treated breast cancer patients using gene signatures. *BMC Cancer*, 2016; 16: 555
42. Hayashi M, Nomoto S, Hishida M et al: Identification of the collagen type 1 alpha 1 gene (*COL1A1*) as a candidate survival-related factor associated with hepatocellular carcinoma. *BMC Cancer*, 2014; 14: 108
43. Bae T, Rho K, Choi JW et al: Identification of upstream regulators for prognostic expression signature genes in colorectal cancer. *BMC Syst Biol*, 2013; 7: 86
44. O'Leary PC, Penny SA, Dolan RT et al: Systematic antibody generation and validation via tissue microarray technology leading to identification of a novel protein prognostic panel in breast cancer. *BMC Cancer*, 2013; 13: 175
45. Shi H, Zhou Y, Liu H et al: Expression of CIAPIN1 in human colorectal cancer and its correlation with prognosis. *BMC Cancer*, 2010; 10: 477
46. Dong L, Jensen RV, De Rienzo A et al: Differentially expressed alternatively spliced genes in malignant pleural mesothelioma identified using massively parallel transcriptome sequencing. *BMC Med Genet*, 2009; 10: 149
47. Sridhar S, Schembri F, Zeskind J et al: Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. *BMC Genomics*, 2008; 9: 259
48. Wu DQ, Ye J, Ou HY et al: Genomic analysis and temperature-dependent transcriptome profiles of the rhizosphere originating strain *Pseudomonas aeruginosa* M18. *BMC Genomics*, 2011; 12: 438