Research Paper

Construction and evaluation of a nomogram for predicting survival in patients with lung cancer

Jin Ouyang[1,2,3], Zhijian Hu[4], Jianlin Tong[4], Yong Yang[3], Juan Wang[3], Xi Chen[3], Ting Luo[1], Shiqun Yu[1], Xin Wang[1], Shaoxin Huang[1,3,5]

[1]Laboratory of Precision Preventive Medicine, Medical School, Jiujiang University, Jiujiang, Jiangxi 332000, PR China
[2]Jiangxi Provincial Key Laboratory of Preventive Medicine, Nanchang University, Nanchang 330006, PR China
[3]SpecAlly Life Technology Co. Ltd., Wuhan, Hubei 430075, PR China
[4]Laboratory Department, Jiujiang University Clinical Medical College, Jiujiang University Hospital, Jiujiang, Jiangxi 332000, PR China
[5]School of Public Health, Qingdao University, Qingdao 266100, PR China

Correspondence to: Shaoxin Huang; **email:** huangshaoxinok@163.com, https://orcid.org/0000-0001-9801-4133
Keywords: lung cancer, prognosis, risk score, nomogram, gene expression omnibus
Received: October 11, 2021 **Accepted:** February 28, 2022 **Published:** March 23, 2022

ABSTRACT

Background: Lung cancer is a heterogeneous disease with a severe disease burden. Because the prognosis of patients with lung cancer varies, it is critical to identify effective biomarkers for prognosis prediction.
Methods: A total of 2325 lung cancer patients were integrated into four independent sets (training set, validation set I, II and III) after removing batch effects in our study. We applied the microarray data algorithm to screen the differentially expressed genes in the training set. The most robust markers for prognosis were identified using the LASSO-Cox regression model, which was then used to create a Cox model and nomogram.
Results: Through LASSO and multivariate Cox regression analysis, eight genes were identified as prognosis-associated hub genes, followed by the creation of prognosis-associated risk scores (PRS). The results of the Kaplan-Meier analysis in the three validation sets demonstrate the good predictive performance of PRS, with hazard ratios of 2.38 (95% confidence interval (CI), 1.61–3.53) in the validation set I, 1.35 (95% CI, 1.06–1.71) in the validation set II, and 2.71 (95% CI, 1.77–4.18) in the validation set III. Additionally, the PRS demonstrated superior survival prediction in subgroups by age, gender, p-stage, and histologic type ($p < 0.0001$). The complex model integrating PRS and clinical risk factors also have a good predictive performance for 3-year overall survival.
Conclusions: In this study, we developed a PRS signature to help predict the survival of lung cancer. By combining it with clinical risk factors, a nomogram was established to quantify the individual risk assessments.

INTRODUCTION

Lung cancer remains a highly lethal disease, with a 5-year survival rate of only 19% [1, 2]. Despite progress in treatment strategies, due to late diagnosis, the high mortality rate of lung cancer patients did not drop sharply [2]. Therapy of non-small cell lung cancer (NSCLC) patients has evolved over the past few years with the incorporation of targeted therapy and immune therapy. These changes have increased the importance of prognostic and predictive biomarkers [3]. However, various disease outcomes have been identified in patients with similar clinical and pathological features, suggesting that the current clinical prognostic factors

may be insufficient to consistently predict individual clinical outcomes [4].

With the development of high-throughput technology, RNA-sequencing (RNA-seq) has been broadly used to identify more novel biomarkers in lung cancer research [5]. Talip Zengin et al. used the TCGA database to identify 12 risk genetic features to predict prognosis in patients with lung adenocarcinoma (LUAD), with the AUC values of 0.479 at 1 year, 0.571 at 2 years, 0.622 at 5 years, and 0.676 at 10 years [6]. Shicheng Li et al. identified eight candidate genes related to survival in LUAD. Zuo, S et al. identified the six-gene signature with AUC values of more than 0.650 for 1, 2, 3, 4, and 5-year overall survival (OS) in LUAD [5, 6]. However, the suggested signatures lack consistency among studies and provide limited prognostic information, partially due to the limited sample size and technical factors [7, 8]. To date, all studies that have been executed in an attempt to find prognostic biomarkers for clinical use have failed to achieve higher sensitivity and specificity or are not easily to be validated in external cohorts with relatively small numbers [9–12].

In this study, an eight-gene prognostic signature was identified by evaluating the prognostic value of the related genes to formulate a prognosis-related risk score (PRS). Moreover, we incorporated genes signature and clinical parameters to establish a novel promising prognostic nomogram model with more accurate predictive ability than clinical risk factors for lung cancer patients. Our work may provide a reference for clinicians to formulate more rational treatment strategies, analyze the pathways and possible mechanisms that may affect the prognosis-related lung cancer, and evaluate the differentiation, calibration and clinical value of the model.

MATERIALS AND METHODS

Dataset preparation and samples information collection

As shown in Figure 1, four sets of subjects were enrolled for preliminary and further verification of screened prognostic biomarkers. In this study, a total of 2325 lung cancer patients who had clinical and

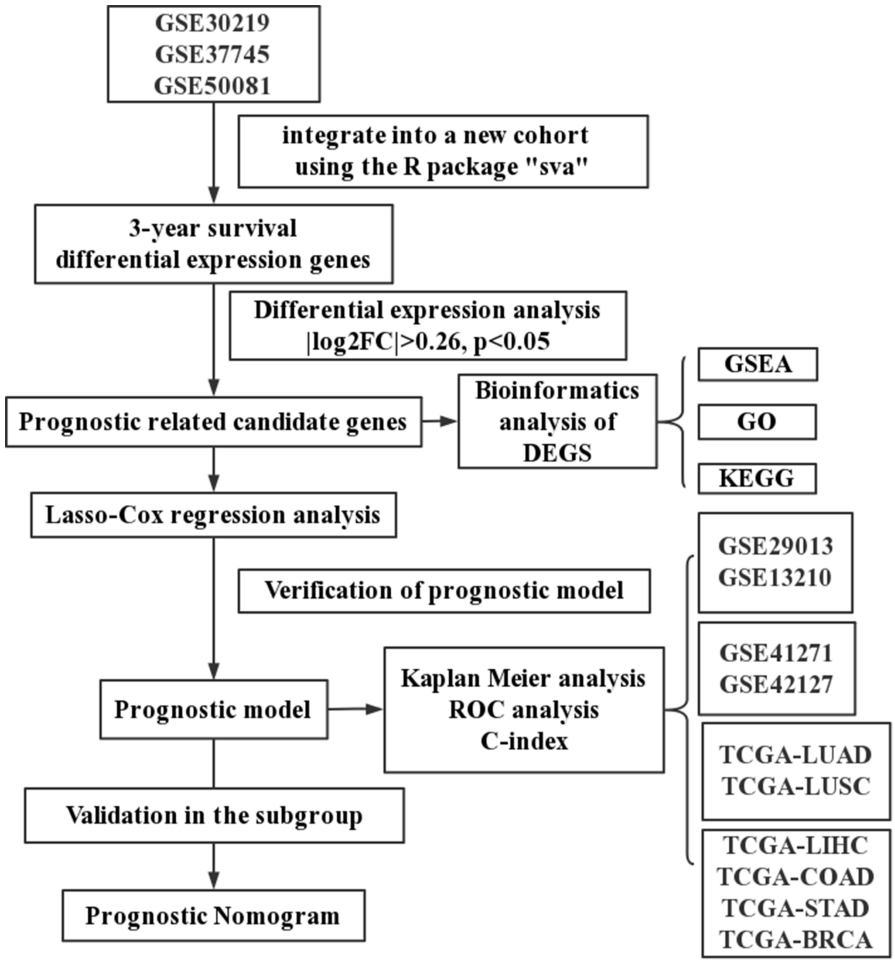


Figure 1. A schematic flowchart for analyzing prognosis-related risk score in lung cancer.

follow-up annotations were included in a training set and three validation sets. Of these, 651 patients in the training set came from GSE30219, GSE37745 and GSE50081 (Affymetrix HG-U133 Plus 2.0 Array). These microarray datasets were downloaded from the gene expression omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) and normalized using a robust multichip average (RMA) algorithm. After batch effects were removed using the combined association test (COMBAT) empirical Bayes method in the surrogate variable analysis (SVA) package, these datasets containing 651 qualified lung cancer patients that were further integrated into a new cohort as the training set. Moreover, similar datasets were processed on the same platform using the identical normalization method and log2 transformation. Validation set I contained a total of 259 lung cancer patients from GSE29013 and GSE31210, and validation set II included 441 lung cancer patients from GSE41271 and GSE42127. The fragments per kilobase per million (FPKM)-normalized RNA-seq data of 494 LUAD and 480 lung squamous cell carcinoma (LUSC) patients retrieved from The Cancer Genome Atlas (TCGA) were integrated as validation set III. We excluded patients with an overall survival (OS) of less than 30 days or with a vague or absent vital status.

In addition, we also established the specificity validation sets of four other cancers, including liver hepatocellular carcinoma (TCGA-LIHC), colon adenocarcinoma (TCGA-COAD), stomach adenocarcinoma (TCGA-STAD), and breast invasive carcinoma (TCGA-BRCA). Supplementary Table 1 shows the dataset information within each cohort.

Identification of survival-related gene and functional enrichment analysis

The average OS period of the patients in the training set was approximately 3 years, which is a critical point in time. DEGs were identified between alive and deceased subjects within 3 years using the linear models for microarray (LIMMA) package, with difference multiples >1.2 and p-value < 0.05 and were selected for further analysis. Furthermore, we performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis using the R package clusterpro filer at the level of p-value < 0.05 and false discovery rate <0.05. Additionally, the gene set enrichment analysis (GSEA) algorithm in R package gene set variation analysis (GSVA) was used to evaluate the biomarker performance in the training sets retrieved from the Molecular Signature Database (MSigDB) [13].

Candidate selection and signature establishment

The least absolute shrinkage selection operator (LASSO) algorithm was used to identify the 306 DEG candidate genes with the best survival prediction features in the training sets. Subsequently, we performed multivariate Cox regression analysis based on the results of LASSO analysis. Cox proportional risk regression models were used to assess the importance of each candidate for OS. PRS were calculated as follows: $PRS = exp_{gene1} \times \beta_{gene1} + exp_{gene2} \times \beta_{gene2} + \cdots + exp_{geneN} \times \beta_{geneN}$ by weighting normalized gene expression values according to their Cox coefficients.

Study subjects in each dataset were divided into high- and low-risk groups according to the cut-off points of median risk scores. Kaplan-Meier (K-M) survival curves and time-dependent receiver operating characteristic (survival-ROC) were conducted to evaluate the prognostic value of the risk score model. The higher the calculated C-index, the more precise the prediction.

Validation of the prognostic signature

In the training set, a stratification analysis was performed to determine whether the prognostic signature could accurately predict patient survival in different clinical factor subgroups. The model's performance was further evaluated using the independent validation set. In addition, the specificity of the model was tested in four other vital cancers. Gene expression data from different sets were adjusted individually by subtracting the median expression value after log2 transformation.

The PRS was combined with clinically informative variables to create a multivariate cox regression model (complex model) and a nomogram to visualize the predicted outcome for each patient. Additionally, the Hosmer-Lemeshow test was used to validate calibration curves that were established to improve the accuracy of nomogram prediction [14]. To accomplish this, we calculated the total score derived from the established nomogram for each patient in the validation set and generated a calibration curve with Cox regression [15].

Statistical analysis

IBM SPSS Statistics 26 (IBM Corp., Armonk, NY, USA), GraphPad Prism 7.0 (GraphPad Software Inc, San Diego, CA, USA), the EmpowerStats software (http://www.empowerstats.com, X&Y solutions, Inc. Boston MA, USA) and R software (version 4.1.0,

http://www.r-project.org) were used to analyze data and plot graphs. LASSO logistic regression analysis was conducted using the glmnet package in R. Nomogram plots were established by the root mean squares (RMS) package. The pROC and survival-ROC packages were applied to analyze ROC and time-dependent ROC (tROC) curves. Independent sample *t*-tests or Mann-Whitney *U*-tests were used to compare continuous variables, and chi-square tests were used to compare categorical variables. Statistical significance was defined as a p-value < 0.05.

RESULTS

Identification of the DEGs and the hub markers associated with prognosis in the training set

A total of 651 lung cancer samples with the OS time over 30 days were analyzed. The LIMMA software was used to identify the DEGs between samples from alive or deceased patients with 3-year survival in lung cancer. Out of 14,052 genes, 306 met the threshold set (adjusted p-value < 0.05, |log2FC| >0.26), with 146 DEGs being down-expression and 160 being up-expression in the deceased group (Figure 2A, 2B).

To determine the most robust prognostic indicators, LASSO-Cox regression models were used. To overcome over-fitting, tenfold cross-validation was used on the 306 DEGs associated with 3-year survival. The more robust prognostic candidates were investigated using the LASSO regression method with an optimal value of 0.1618 (Figure 2C). The results showed that all 26 prognosis-related candidates had non-zero LASSO coefficients (Figure 2D). Subsequently, multiple stepwise Cox regression was used to determine the impact of the candidate genes, and eight hub markers were chosen to construct the risk model in lung cancer patients (Table 1). The expression profiles of these eight genes showed that elevated expression of secreted phosphoprotein 1 (SPP1), sodium-dependent phosphate transporter 1 (SLC20A1), and centromere protein H (CENPH) in lung cancer samples were risk factors for prognosis. In contrast, high-expression of MAGE family member E1 (MAGEE1), chromosome 16 open reading frame 54 (C16ORF54), potassium voltage-gated channel subfamily S member 3 (KCNS3), tripartite motif containing 68 (TRIM68) and cytochrome B5 domain containing 2 (CYB5D2) were protective factors for prognosis (Figure 2E). In addition, the expression of the eight genes was significantly different between

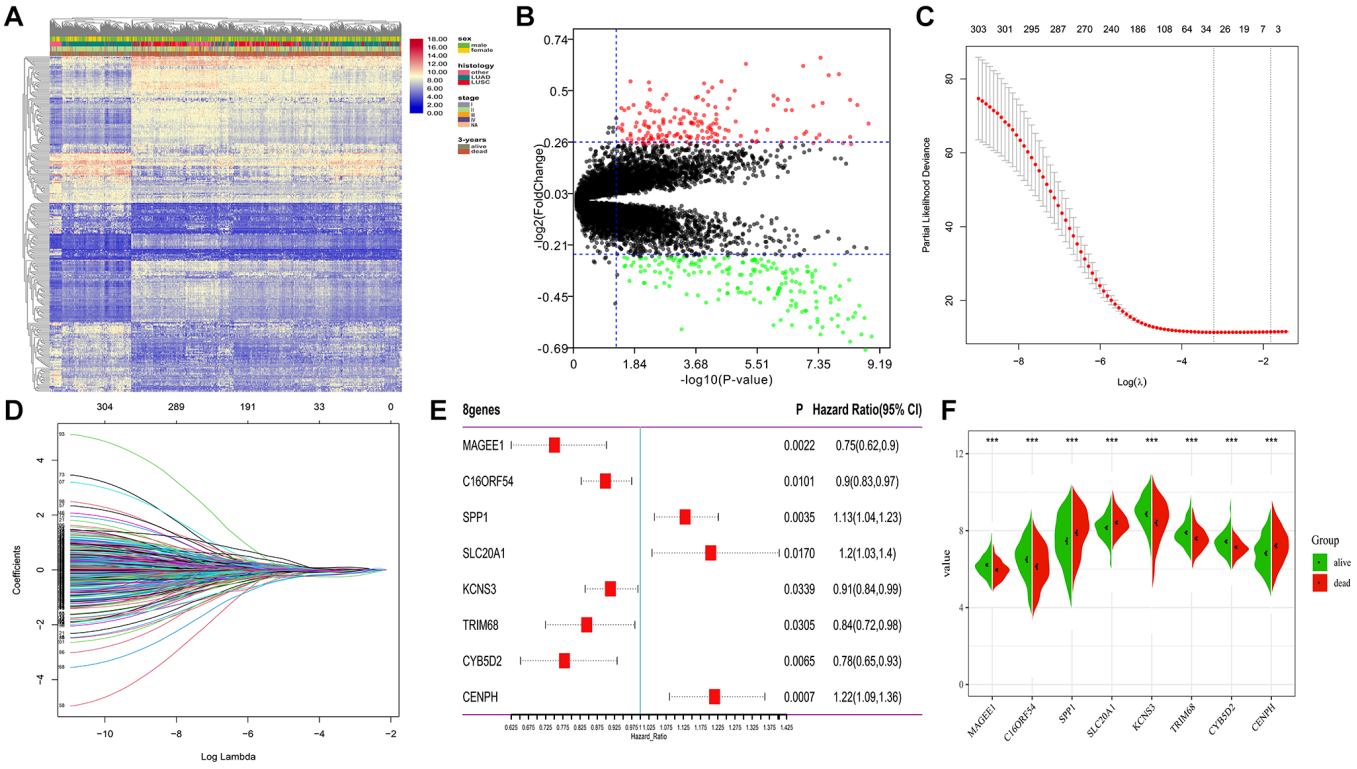


Figure 2. Differential expression and LASSO-Cox regression results of DEGs. (**A**, **B**) Heatmap plot and volcano plot represents the expression of 306 DEGs between samples from alive and deceased patients based on 3-year survival data, satisfying the criteria of adjusted p-value < 0.05, |log2FoldChange| >0.26. (**C**, **D**) 26 genes considered the more correlated with prognosis were identified by LASSO regression method. (**E**) Coefficients of eight genes estimated by multivariate Cox regression. (**F**) Expression profiles of eight hub genes between samples from alive and deceased patients with 3-year survival data.

Table 1. The eight hub markers identified distinguish alive from deceased patients based on 3-year survival data.

Gene symbol	Protein name (UniProt accession)	Log2FC (dead/alive)	Known functions	Relation to cancer
SPP1	secreted phosphoprotein 1	0.43	Functions to control the survival, growth, differentiation and effector function of tissues and cells	Implicated in tumorigenesis in various cancer types [16, 17] Over expressed in lung neoplasms [18, 19]
SLC20A1	Sodium-dependent phosphate transporter 1	0.27	High-affinity inorganic phosphate:sodium symporter activity	High expression of SLC20A1 mRNA inhibits the progress of lung cancer [20]
KCNS3	Potassium voltage-gated channel subfamily S member 3	−0.47	Involved in energy metabolism	The expression is related to breast cancer and lung cancer, promoted metastasis [21, 22]
TRIM68	Tripartite motif containing 68	−0.31	Enhances the transcriptional activity of the AR [23]	Preferentially expressed in prostate cancer cells [23]
CYB5D2	Cytochrome b5 domain containing 2	−0.29	Interacting selectively and non-covalently with heme	Downregulation of CYB5D2 is associated with breast cancer progression [24]
CENPH	Centromere protein H	0.39	Negative regulation of cysteine-type endopeptidase activity	Higher expression levels of CENPH tended to have worse OS in lung cancer [25]
MAGEE1	MAGE family member E1	−0.26	Participating in specific biological processes	The expression of MAGEE1 is correlated with tumor-cell proliferation of NSCLC [26]
C16ORF54	Chromosome 16 open reading frame 54	−0.37	Protein amino acid binding and integral component of membrane	Tobacco Smoke Pollution results in decreased expression of C16ORF54 mRNA in lung cancer [27]

samples from alive or deceased patients with 3-year survival in lung cancer (Figure 2F). The K-M survival curve analysis revealed that the expression of these eight genes is significantly associated with lung cancer prognosis (Supplementary Figure 1).

Construction and validation of the PRS in lung cancer

The predictive model was constructed using the eight hub markers identified using the multiple Cox regression method. The risk score of each patient was calculated based on the cox coefficients: PRS = 0.1223 × expression level of SPP1 + 0.1862 × expression level of SLC20A1 – 0.0909 × expression level of KCNS3 – 0.1693 × expression level of TRIM68 – 0.2490 × expression level of CYB5D2 + 0.1954 × expression level of CENPH – 0.2869 × expression level of MAGEE1 – 0.1069 × expression level of C16ORF54.

The cut-off value was determined automatically based on the median risk score, and lung cancer patients were divided into the low- (n = 390) and high-risk (n = 259) groups using the cut-off value of −2.39. As illustrated in Figure 3A, the distribution of the PRS, OS time, and heatmap for the eight-gene signature in the training set is shown from top to bottom. Furthermore, the tROC analysis revealed that PRS was the most accurate predictor of OS (Supplementary Figure 2). The AUC of this PRS model in 1-year, 3-year, 5-year was 0.72, 0.75, and 0.71, respectively (Figure 3B). Moreover, the K-M survival analysis revealed that the patients had worse OS in the high-risk group than in the low-risk group (HR = 2.72; 95% confidence interval (CI), 2.26 to 3.27, $p < 0.0001$), and the C-index of the PRS for predicting survival was 0.67; 95% CI, 0.65 to 0.70 (Figure 3C).

Validation of PRS signature in the subgroup and independent lung cancer validation sets

To confirm the prognostic robustness of PRS features and complex models across cohorts, we further validated it in the three independent external cohorts described earlier. Similarly, in each of the three validation sets, patients in the high-risk group had poorer outcomes, while those in the low-risk group had a higher survival rate (Figure 4A–4C). K-M analysis confirmed that the predicted high-risk group had a significantly shorter time to death (Figure 4D–4F), indicating good predictive performance of the PRS, with a HR of 2.38; 95% CI, 1.61 to 3.53 in validation set I, 1.35; 95% CI, 1.06 to 1.71 in validation set II, and 2.71; 95% CI, 1.77 to 4.18 in validation set III.

Subsequently, a stratified analysis was performed to assess whether PRS characteristics could predict the probability of patient survival in the same subgroup of clinical factors. Patients in the training cohort were classified clinically by p-stage (I/II/III-IV) (Figure 5A), histological type (glandular/squamous) (Figure 5B), gender (female/male) (Figure 5C, 5D) and age (<65/≥65) (Figure 5E, 5F). The results showed that PRS characteristics could divide patients with the same age, sex, p-stage, and histological type into high-risk and low-

risk groups. In each tier, OS was shorter in patients with high-risk scores than in patients with low-risk scores ($p < 0.001$) (Figure 5). As a result of the above analysis, the HR does not change significantly across subgroups, and the PRS is an independent risk factor of lung cancer prognosis. It has predictive value in different people. In addition, to further prove the specificity of PRS as a prognostic factor of lung cancer in the clinic, we tested it in four other primary global cancers. The results showed that PRS signature was not associated with the prognosis of liver cancer, bowel cancer, gastric cancer, or breast cancer (Supplementary Figure 3), indicating that the PRS signature is only related to the prognosis of lung cancer.

Prognostic nomogram for OS

A total of 614, 259, 438, and 740 patients with full-scale five clinical annotations including age, sex (male or female), histology (LUAD, LUSC or other), p-stage (I, II, III or IV) and PRS (low or high) were extracted

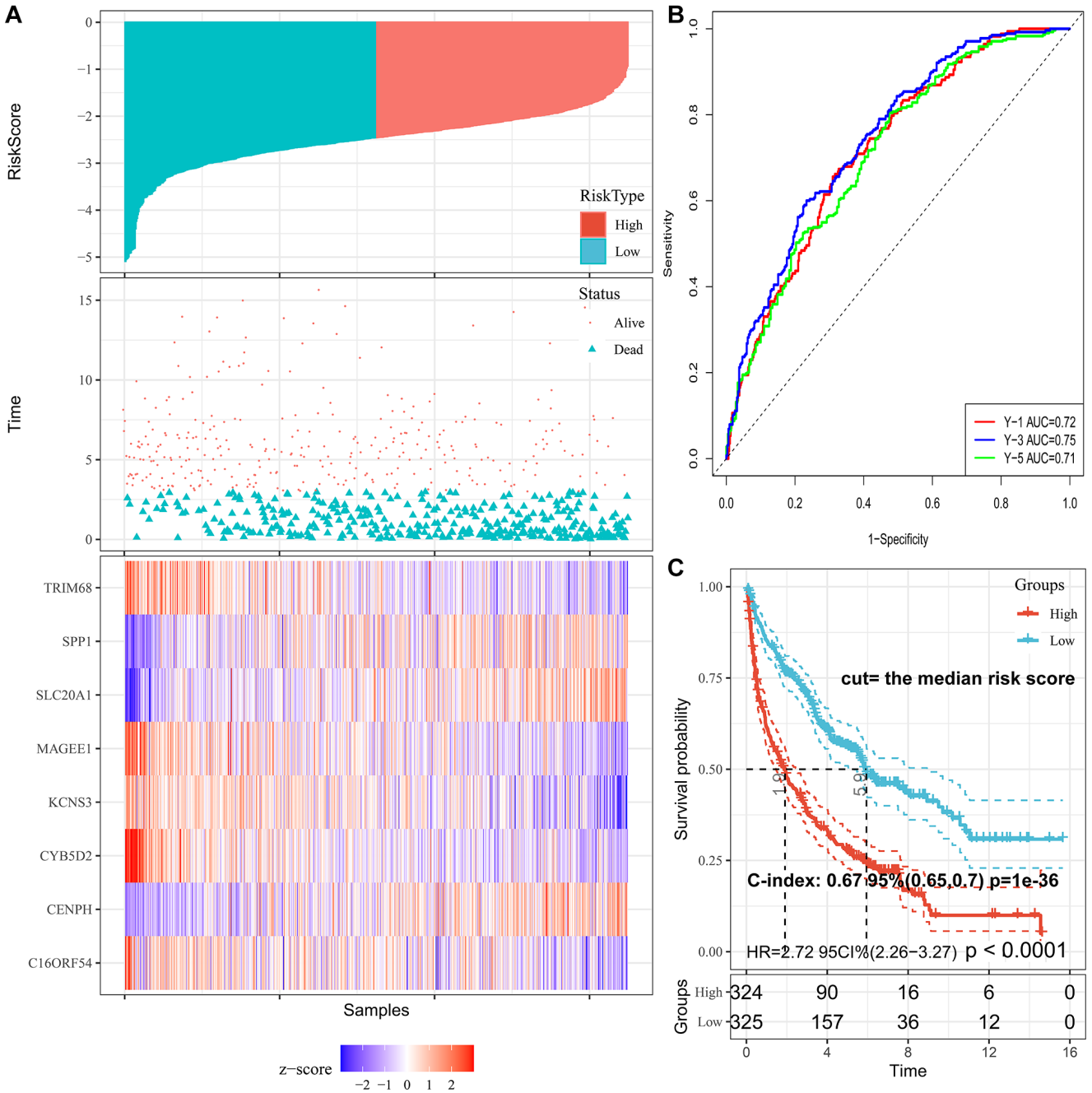


Figure 3. Characteristics of PRS signature in the training cohort. (**A**) Risk scores distribution, survival status, and gene expression patterns of patients in high- and low-risk groups in the training cohort. (**B**) Time-dependent ROC analysis for predicting OS. (**C**) Survival curves and C-index for high- and low-risk groups.

from the training, validation I, II, and III sets, respectively (Table 2). In the training set, multivariate Cox regression analysis revealed that these five variables were correlated with the prognosis of lung cancer, with PRS being the most significant predictor of overall survival in the Cox model (complex model) (Figure 6A). These variables were used to construct a decision tree to improve risk stratification for overall survival. As shown in Figure 6B, only p-stage and PRS remained in the decision tree, with three different risk subgroups identified. To quantify the risk assessment and survival probability for individual patients, a nomogram incorporating PRS and other clinicopathological features was constructed (Figure 6C). Furthermore, we calculated the value of each covariate of patient No. 350 (GSM1213824) and mapped it to the corresponding score, calculated the total score, and its probability at 3-year and 5-year survival. The calculated values were 0.757 and 0.881. For 3- or 5-year survival, the probability calibration plot revealed the best agreement between nomogram prediction and actual observation (45-degree dotted line) (Figure 6D), indicating that the nomogram is highly accurate. When compared to other features, the nomogram exhibited the most powerful and stable ability for survival prediction, with an average AUC greater than 0.7, significantly better than the pathological p-stage (Figure 6E).

At the same time, the complex model combining PRS, and clinical risk factors also had a good predictive performance of 3-year survival, namely 0.788, 0.709 and 0.614, respectively in the three validation sets (Supplementary Figure 4).

Functional analysis of the survival-related DEGs

To further understand the underlying mechanism of the survival-related DEGs, we analyzed 306 DEGs between samples from alive or deceased patients with 3-year survival in the training set. Enrichment analyses involved the KEGG, GO functional enrichment and GSEA of hallmark in MSigDB. The genes were divided into 3 categories: biological process, cellular component, and molecular function according to the GO terms (Figure 7A–7C). The most abundant groups were nuclear division, chromosomal region, and cofactor binding, respectively, in the three categories. We discovered that pathways involving the cell cycle,

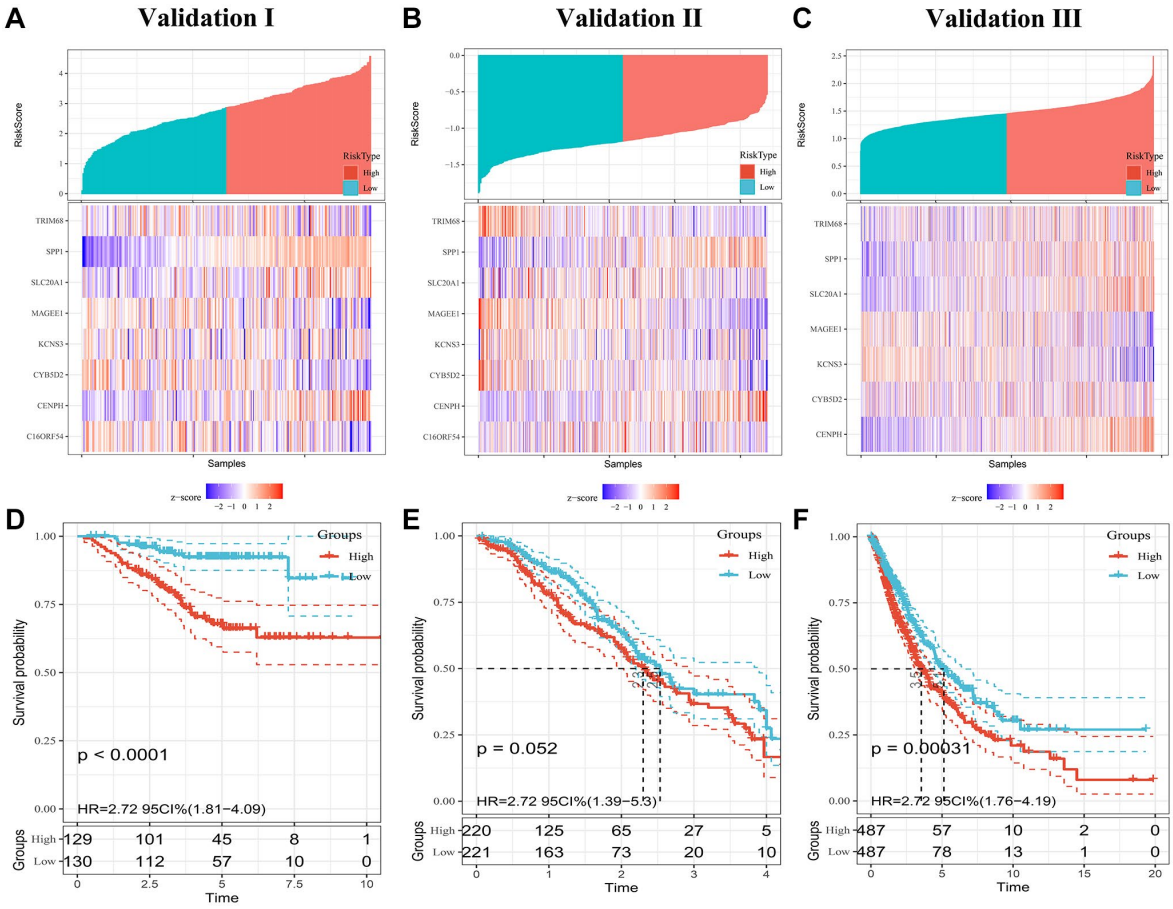


Figure 4. Evaluating PRS signatures in validation sets. (**A–C**) Risk score distribution and survival status of patients in high- and low-risk groups in validation sets. (**D–F**) Survival curves in validation sets.

cellular senescence, oocyte meiosis, and the p53 signaling pathway were enriched in KEGG (Figure 7D). Additionally, the GSEA results based on the hallmark gene sets in MSigDB indicated that these DEGs were primarily associated with not only HALLMARK G2M CHECKPOINT (normalized enrichment score (NES) = −1.30, $p < 0.001$), but also HALLMARK E2F TARGETS (NES = 2.024, $p < 0.001$) (Figure 7E–7G). The top pathway and hallmark:cell cycle and E2F TARGETS clarified the division of activity in lung cancer cells. We summarized a working model of the activated pathways in Figure 8.

DISCUSSION

Lung cancer, one of the most common malignant tumors worldwide, claims over one million deaths each year and has a dismal 5-year survival rate [28]. Thus, changes in the prognosis of lung cancer patients may occur long before detectable clinicopathological abnormalities, highlighting the correlation between biomarkers such as the expression of specific genes (i.e., hub genes) and lung cancer prognosis [29, 30]. Many studies at the biological and clinical levels have suggested the link between gene mutation sites and disease progression, and the use of high-throughput sequencing data based on omics to make more accurate diagnosis and prognosis predictions for lung cancer patients, so as to formulate individualized treatment plans on this basis to bring greater benefits to the prognosis of patients [31].

The current cancer progression prediction is mainly based on disease manifestations and the

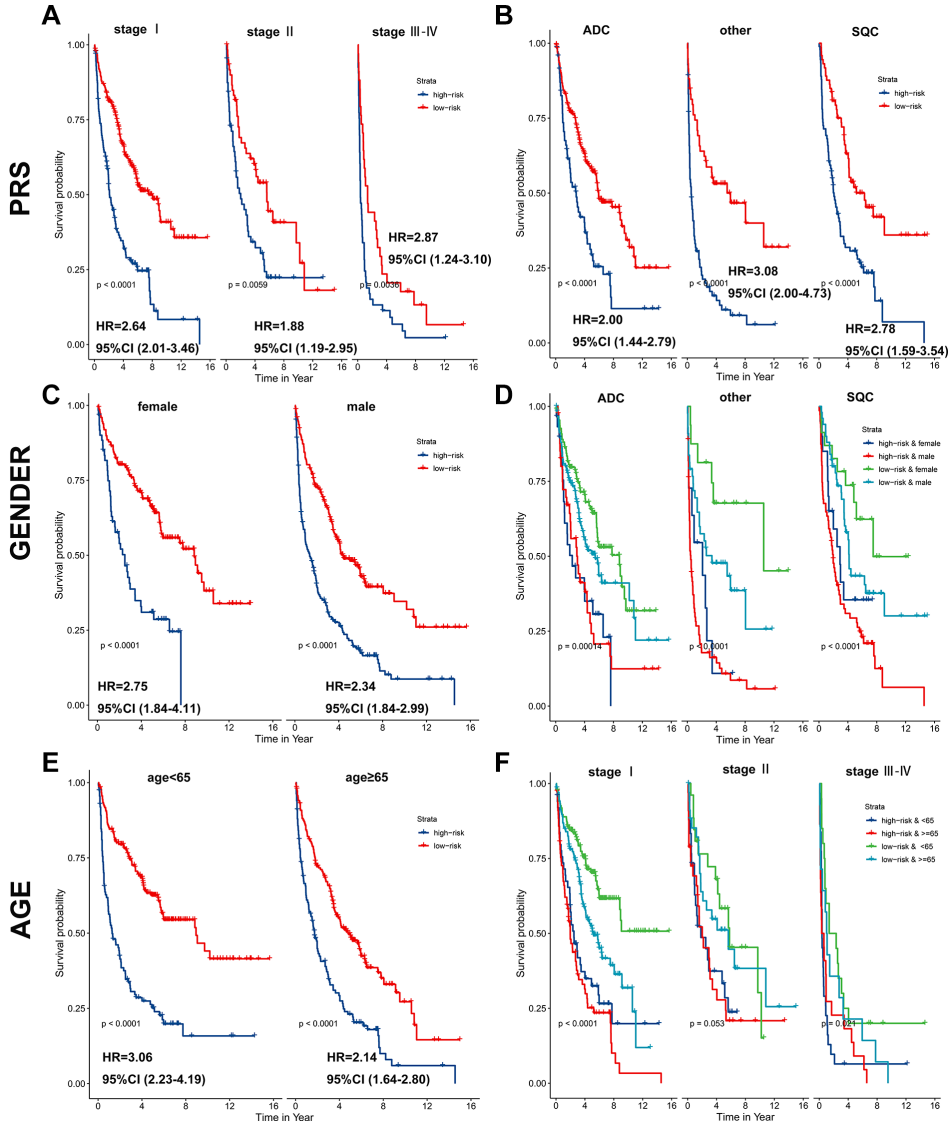


Figure 5. PRS as a valuable predictor for OS in subgroups. PRS discriminated high-risk patients with different clinicopathological characteristics, including (**A**) p-stage, (**B**) histological type, (**C**, **D**) gender, and (**E**, **F**) age.

Table 2. Clinical characteristics and PRS model of lung cancer patients in the training and validation sets.

Exposure	Training set	Validation set I	Validation set II	Validation set III
	N = 614	*N* = 259	*N* = 438	*N* = 740
Age	64.29 ± 10.55	60.4 ± 7.9	65.0 ± 9.9	66.1 ± 9.5
Gender	–	–	–	–
Male	404 (65.8%)	133 (51.4%)	230 (52.5%)	453 (61.2%)
Female	210 (34.2%)	126 (48.6%)	208 (47.5%)	287 (38.8%)
Histology	–	–	–	–
LUAD	308 (50.2%)	234 (90.3%)	309 (70.5%)	476 (64.3%)
LUSC	164 (26.7%)	25 (9.7%)	120 (27.4%)	264 (35.7%)
Other	142 (23.1%)	0	9 (2.1%)	0
P-stage	–	–	–	–
I	402 (65.5%)	186 (71.8%)	237 (54.1%)	383 (51.8%)
II	125 (20.4%)	56 (21.6%)	81 (18.5%)	192 (26.0%)
III	77 (12.5%)	17 (6.6%)	114 (26.0%)	135 (18.2%)
IV	10 (1.6%)	0	6 (1.4%)	30 (4.0%)
PRS	–	–	–	–
Low-risk	369 (60.0%)	129 (49.8%)	220 (49.8%)	487 (50.0%)
High-risk	245 (40.0%)	130 (50.2%)	221 (50.2%)	487 (50.0%)

Tumor-Node-Metastasis staging system of American Joint Commission on Cancer (AJCC). However, both methods' static representations of clinicopathologic factors fail to account for the genetic heterogeneity of cancer, limiting their predictive value [11]. Recent studies have shown that gene mutations and expression disorders are associated with disease progression and therapeutic response in lung cancer [32–34]. However, these current biomarkers, such as epidermal growth factor receptor (EGFR) and Kirsten rat sarcoma virus (KRAS) oncogene homologs, do not fully represent the complex mechanisms of lung cancer progression [35, 36]. Larsen et al. developed a 54-gene signature in LUAD [10], but the combined accuracy in predicting recurrence is only 69% (79% sensitivity, 59% specificity). Sheng et al. reported a new biological marker discovery pathway, which integrates RNA sequencing (RNA-seq) and clinical data to identify progression gene signatures (PGSs) based on survival genes, and discovered 22 LUAD-PGS genes and 23 LUSC-PGS genes that have a high predictive value (area under the curve (AUC) = 0.85, 0.92, respectively) [11]. This model needs to be further optimized to facilitate clinical implementation. Recently, Xie et al. developed a prognostic model for death due to extensive-stage squamous cell lung cancer (SCLC), with an unadjusted concordance (C)-index of 0.714 [12].

In this study, through more rigorous identification, identified eight independent prognostic genes to establish a risk score for assessing the survival probability in the training set. As shown in Table 1, several of the gene signatures we identified have been investigated in various types of tumors. For example, MAGEE1 was associated with important clinical and molecular features in glioma [37, 38], and can be considered an important marker in determining the prognosis of glioblastoma [39]. More importantly, the expression of MAGEE1 is correlated with tumor-cell proliferation of NSCLC [26]. As a gene that is overexpressed in breast, bladder, colorectal, head and neck, liver, lung, and esophageal cancers [40], SPP1 has the potential to influence not only the occurrence and progression of LUAD, but also to serve as an independent prognostic marker and a novel therapeutic target [41–43]. High SLC20A1 expression is associated with poor prognoses in basal-like breast cancers, longue cancer and esophageal adenocarcinoma [44–46]. CENPH was found to drive the molecular changes during the pathologic stages of LUAD [47], and patients with a higher expression level of CENPH tended to have a poorer OS [25]. We constructed PRS signatures containing these eight genes under a novel pipeline to support prognosis and OS prediction of lung cancer. The 8-gene PRS and the complex model both had predictive effects in three large cohorts, with AUCs exceeding 70% or even 80%. Taken together, these results suggest that the variable expression of 8-gene model is associated with different prognosis in lung cancer, and may serve as a

prognostic biomarker as well as a treatment target for lung cancer patients.

The prognostic value of PRS features was further validated in another two independent sets. PRS was able to identify high-risk patients in both validation groups, implying that it can be used as a reliable risk factor for the overall population. Patients with higher PRS had poorer survival compared to those with lower PRS. In addition, when combined with clinical risk factors, a

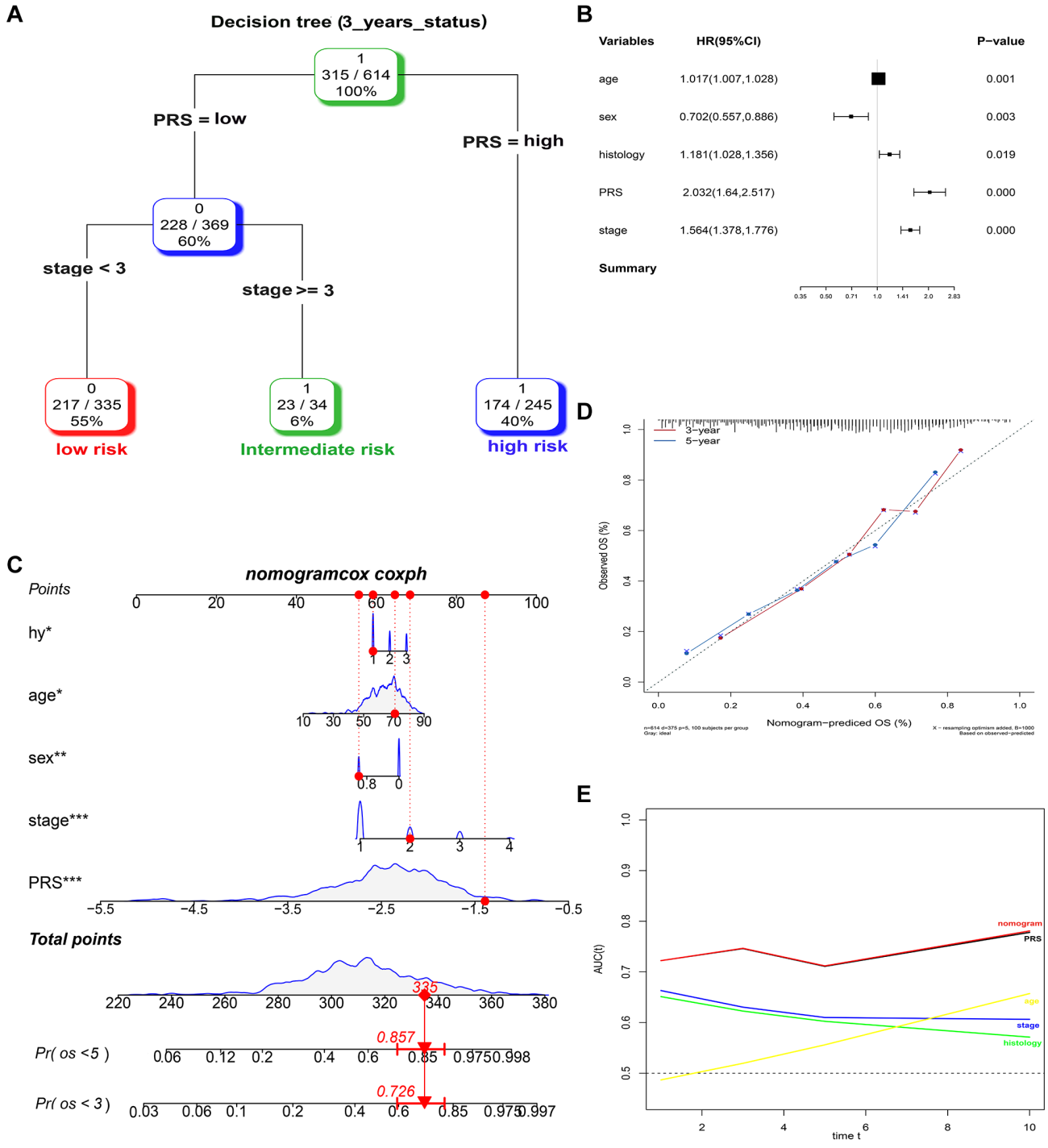


Figure 6. Combination of the PRS signature and clinical features improves survival prediction in training sets. (**A**) A decision tree was constructed to improve risk stratification. (**B**) Multivariate Cox regression model (complex model). (**C**) Survival nomogram for quantifying risk assessment for individual patients. (**D**) Calibration analysis revealed a high degree of accuracy in predicting survival at 3 or 5 years. (**E**) Among all clinical variables, tROC analysis demonstrated that the nomogram was the most stable and powerful predictor of OS.

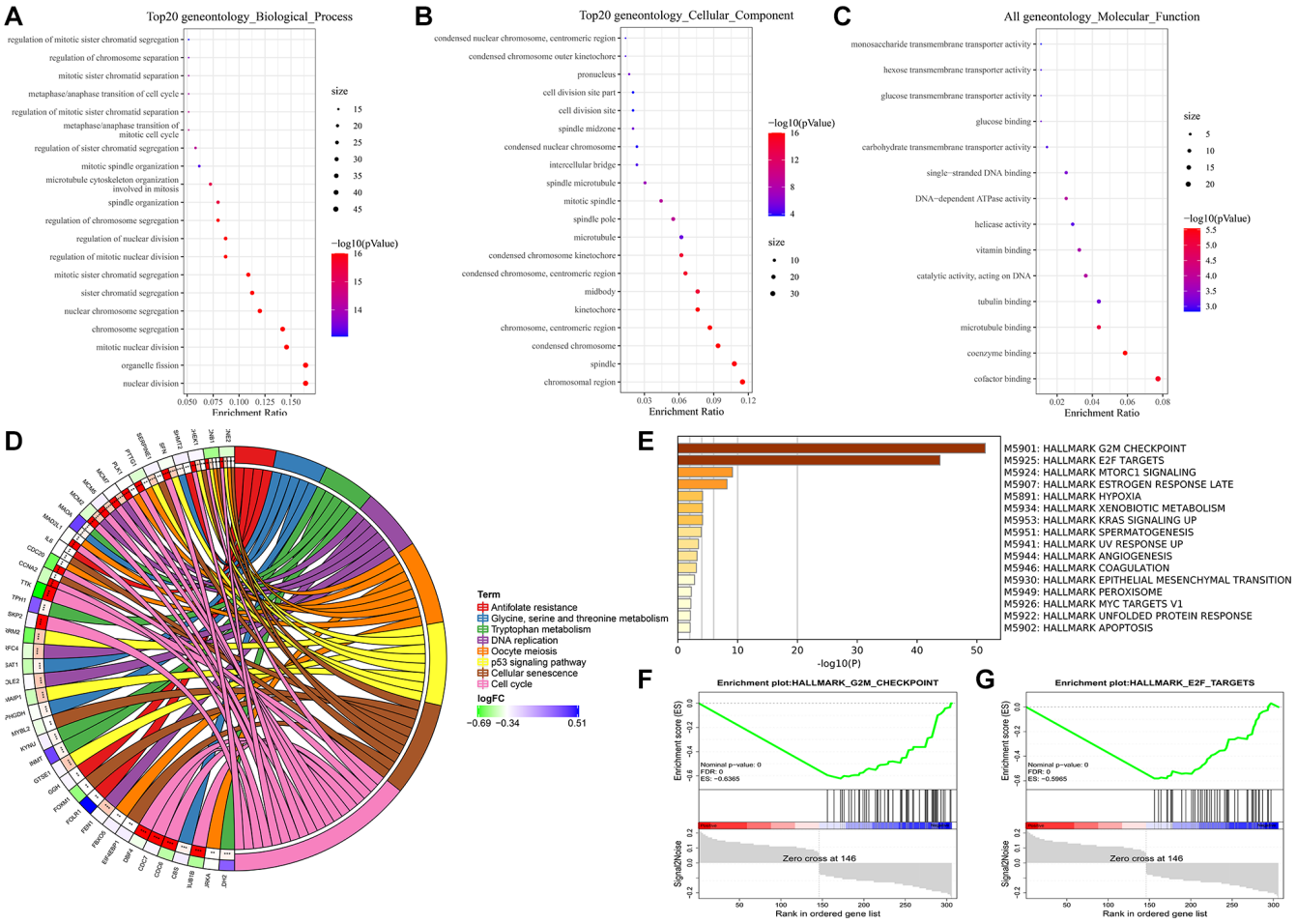


Figure 7. Enrichment analyses of DEGs. (**A**) Biological process. (**B**) Cellular component. (**C**) Molecular function. (**D**) KEGG pathway analysis. (**E–G**) GSEA analysis using hallmark gene sets from MSigDB.

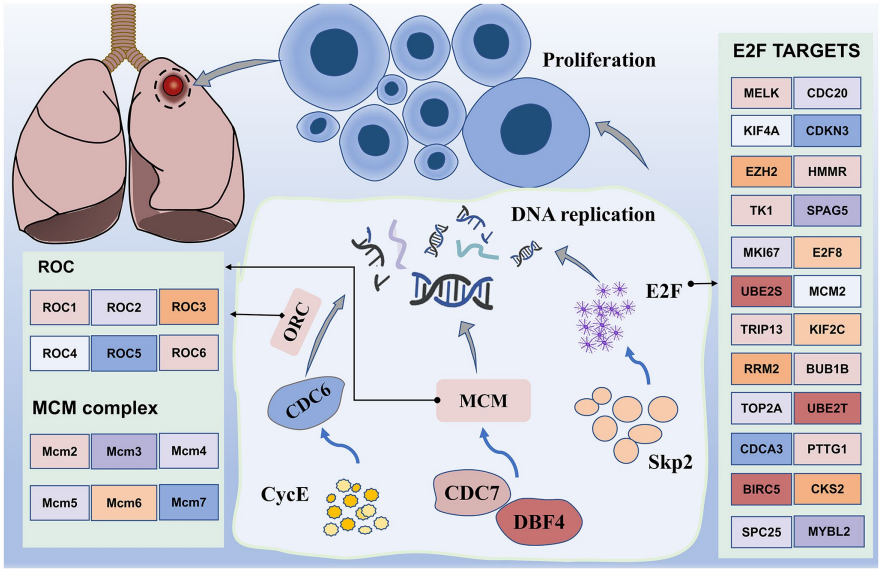


Figure 8. Working model of major enrichment pathways in lung cancer. Proteins activated by CDC7, DBF4, CYCE, MCM, and CDC6 promote DNA replication, in addition to promoting cell amplification, particularly transcription factor E2F, which is regulated by numerous genes. In addition, upregulation of the transcription activity of E2F promotes Skp2 regulation of the PI3K/AKT pathway, thereby potentially promoting the occurrence of lung cancer.

column line plot was built for the risk quantification in individual patients. ROC analysis showed that PRS had considerable risk predictive power for OS, and calibration analysis showed that nomogram survival prediction results were extremely close to actual survival.

In addition, we note the enrichment results, which mainly include the cell cycle. Interestingly, the GSEA results from the hallmark gene sets were also enriched for genes encoding cell cycle-related targets of the E2F transcription factor. Our data uncovered differential expression of multiple genes, including CDC6, CDC7, DBF4, MCM and CycE (Figure 8). Most of them may induce SQLC through DNA replication and cell cycle pathway [48]. Minichromosome maintenance complex component 4 (MCM4), a highly expressed gene in NSCLC, is required for the proliferation of NSCLC cells [49]. MCM proteins, including MCM2-7, are also required for replication initiation and elongation [50]. In addition, Skp2 activated PI3K/AKT pathway activities by upregulating the transcription activity of E2F, thereby potentially promoting the occurrence of lung cancer [51]. These results suggest that the transcriptional regulation through E2F may be a novel therapeutic target in lung cancer.

Although our study reveals the feasibility of a new approach to biomarker discovery that integrates cancer survival and overall genetic profiling data, important questions remain to be addressed in order to facilitate the clinical implementation of PRS in clinical prognostic testing. Our data demonstrate the feasibility of using PRS as a clinical test; however, large-scale clinical studies are needed to statistically validate the ability of PRS to define patients at high risk for poor prognosis. Future studies will also aim to develop new companion therapies for PRS and other biomarker discovery pipelines.

CONCLUSIONS

To predict OS in patients with lung cancer, we constructed an eight-gene based PRS that was further validated in another three validation sets as well as other cancer sets. Because PRS was found to be associated with independent and specific risk factor for lung cancer, patients with higher PRS had poorer survival outcome. By combining genes signature with clinical features, we developed a nomogram model to quantify the risk for individual patients. This model can also be used to identify patients who may benefit from adjuvant therapy, allowing for more personalized treatment in lung cancer. In addition, enrichment analysis revealed that the key genes were associated with the cell cycle and E2F targets.

Abbreviations

LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; PGSs: progression gene signatures; AUC: area under the curve; SCLC: squamous cell lung cancer; GEO: gene expression omnibus; OS: overall survival; PRS: prognosis-related risk score; TCGA: The Cancer Genome Atlas; DEGs: Differentially expressed genes; GSEA: Gene Set Enrichment Analysis; K-M: Kaplan-Meier; CI: confidence interval; NSCLC: non-small cell lung carcinoma; tROC: time-dependent ROC.

AUTHOR CONTRIBUTIONS

JOY, SXH, and XW conceived and designed the present study. TL and SQY drafted the initial manuscript. ZJH, JLT and YY analyzed the data. All authors have read and approved the final manuscript. JW and XC confirm the authenticity of all the raw data.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest related to this study.

FUNDING

This work was supported by the China Postdoctoral Science Foundation (2019M652334), the Natural Science Foundation of Jiangxi (2020BAB206067), the National Natural Science Foundation of China (NO.81360447) and the Natural Science Foundation of Jiangxi (20192ACB20019).

REFERENCES

1. Nasim F, Sabath BF, Eapen GA. Lung Cancer. Med Clin North Am. 2019; 103:463–73. https://doi.org/10.1016/j.mcna.2018.12.006 PMID:30955514

2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA Cancer J Clin. 2021; 71:7–33. https://doi.org/10.3322/caac.21654 PMID:33433946

3. Thakur MK, Gadgeel SM. Predictive and Prognostic Biomarkers in Non-Small Cell Lung Cancer. Semin Respir Crit Care Med. 2016; 37:760–70. https://doi.org/10.1055/s-0036-1592337 PMID:27732997

4. Brundage MD, Davies D, Mackillop WJ. Prognostic factors in non-small cell lung cancer: a decade of progress. Chest. 2002; 122:1037–57. https://doi.org/10.1378/chest.122.3.1037 PMID:12226051

5. Li S, Xuan Y, Gao B, Sun X, Miao S, Lu T, Wang Y, Jiao W. Identification of an eight-gene prognostic signature for lung adenocarcinoma. Cancer Manag Res. 2018; 10:3383–92. https://doi.org/10.2147/CMAR.S173941 PMID:30237740

6. Zengin T, Önal-Süzek T. Analysis of genomic and transcriptomic variations as prognostic signature for lung adenocarcinoma. BMC Bioinformatics. 2020 (Suppl 14); 21:368. https://doi.org/10.1186/s12859-020-03691-3 PMID:32998690

7. Zuo S, Wei M, Zhang H, Chen A, Wu J, Wei J, Dong J. A robust six-gene prognostic signature for prediction of both disease-free and overall survival in non-small cell lung cancer. J Transl Med. 2019; 17:152. https://doi.org/10.1186/s12967-019-1899-y PMID:31088477

8. Seijo LM, Peled N, Ajona D, Boeri M, Field JK, Sozzi G, Pio R, Zulueta JJ, Spira A, Massion PP, Mazzone PJ, Montuenga LM. Biomarkers in Lung Cancer Screening: Achievements, Promises, and Challenges. J Thorac Oncol. 2019; 14:343–57. https://doi.org/10.1016/j.jtho.2018.11.023 PMID:30529598

9. Zhang MY, Liu XX, Li H, Li R, Liu X, Qu YQ. Elevated mRNA Levels of AURKA, CDC20 and TPX2 are associated with poor prognosis of smoking related lung adenocarcinoma using bioinformatics analysis. Int J Med Sci. 2018; 15:1676–85. https://doi.org/10.7150/ijms.28728 PMID:30588191

10. Larsen JE, Pavey SJ, Passmore LH, Bowman RV, Hayward NK, Fong KM. Gene expression signature predicts recurrence in lung adenocarcinoma. Clin Cancer Res. 2007; 13:2946–54. https://doi.org/10.1158/1078-0432.CCR-06-2525 PMID:17504995

11. Sheng KL, Kang L, Pridham KJ, Dunkenberger LE, Sheng Z, Varghese RT. An integrated approach to biomarker discovery reveals gene signatures highly predictive of cancer progression. Sci Rep. 2020; 10:21246. https://doi.org/10.1038/s41598-020-78126-3 PMID:33277589

12. Zhong J, Zheng Q, An T, Zhao J, Wu M, Wang Y, Zhuo M, Li J, Zhao X, Yang X, Jia B, Chen H, Dong Z, et al. Nomogram to predict cause-specific mortality in extensive-stage small cell lung cancer: A competing risk analysis. Thorac Cancer. 2019; 10:1788–97. https://doi.org/10.1111/1759-7714.13148 PMID:31318178

13. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011; 27:1739–40. https://doi.org/10.1093/bioinformatics/btr260 PMID:21546393

14. Zhang J, Zhao X, Zhao Y, Zhang J, Zhang Z, Wang J, Wang Y, Dai M, Han J. Value of pre-therapy ^{18}F-FDG PET/CT radiomics in predicting EGFR mutation status in patients with non-small cell lung cancer. Eur J Nucl Med Mol Imaging. 2020; 47:1137–46. https://doi.org/10.1007/s00259-019-04592-1 PMID:31728587

15. Wang Y, Li J, Xia Y, Gong R, Wang K, Yan Z, Wan X, Liu G, Wu D, Shi L, Lau W, Wu M, Shen F. Prognostic nomogram for intrahepatic cholangiocarcinoma after partial hepatectomy. J Clin Oncol. 2013; 31:1188–95. https://doi.org/10.1200/JCO.2012.41.5984 PMID:23358969

16. Nakamura M, Eguchi A, Inohana M, Nagahara R, Murayama H, Kawashima M, Mizukami S, Koyanagi M, Hayashi SM, Maronpot RR, Shibutani M, Yoshida T. Differential impacts of mineralocorticoid receptor antagonist potassium canrenoate on liver and renal changes in high fat diet-mediated early hepatocarcinogenesis model rats. J Toxicol Sci. 2018; 43:611–21. https://doi.org/10.2131/jts.43.611 PMID:30298849

17. Daaboul HE, Dagher C, Taleb RI, Bodman-Smith K, Shebaby WN, El-Sibai M, Mroueh MA, Daher CF. β-2-Himachalen-6-ol inhibits 4T1 cells-induced metastatic triple negative breast carcinoma in murine model. Chem Biol Interact. 2019; 309:108703. https://doi.org/10.1016/j.cbi.2019.06.016 PMID:31194954

18. Chiou YH, Liou SH, Wong RH, Chen CY, Lee H. Nickel may contribute to EGFR mutation and synergistically promotes tumor invasion in EGFR-mutated lung cancer via nickel-induced microRNA-21 expression. Toxicol Lett. 2015; 237:46–54. https://doi.org/10.1016/j.toxlet.2015.05.019 PMID:26026961

19. Salem ML, El-Ashmawy NE, Abd El-Fattah EE, Khedr EG. Immunosuppressive role of Benzo[a]pyrene in induction of lung cancer in mice. Chem Biol Interact. 2021; 333:109330. https://doi.org/10.1016/j.cbi.2020.109330 PMID:33245929

20. Cho HY, Miller-DeGraff L, Blankenship-Paris T, Wang X, Bell DA, Lih F, Deterding L, Panduri V, Morgan DL, Yamamoto M, Reddy AJ, Talalay P, Kleeberger SR. Sulforaphane enriched transcriptome of lung

mitochondrial energy metabolism and provided pulmonary injury protection via Nrf2 in mice. Toxicol Appl Pharmacol. 2019; 364:29–44. https://doi.org/10.1016/j.taap.2018.12.004 PMID:30529165

21. Vidya Priyadarsini R, Kumar N, Khan I, Thiyagarajan P, Kondaiah P, Nagini S. Gene expression signature of DMBA-induced hamster buccal pouch carcinomas: modulation by chlorophyllin and ellagic acid. PLoS One. 2012; 7:e34628. https://doi.org/10.1371/journal.pone.0034628 PMID:22485181

22. Labib S, Williams A, Guo CH, Leingartner K, Arlt VM, Schmeiser HH, Yauk CL, White PA, Halappanavar S. Comparative transcriptomic analyses to scrutinize the assumption that genotoxic PAHs exert effects via a common mode of action. Arch Toxicol. 2016; 90:2461–80. https://doi.org/10.1007/s00204-015-1595-5 PMID:26377693

23. Miyajima N, Maruyama S, Bohgaki M, Kano S, Shigemura M, Shinohara N, Nonomura K, Hatakeyama S. TRIM68 regulates ligand-dependent transcription of androgen receptor in prostate cancer cells. Cancer Res. 2008; 68:3486–94. https://doi.org/10.1158/0008-5472.CAN-07-6059 PMID:18451177

24. Ojo D, Rodriguez D, Wei F, Bane A, Tang D. Downregulation of CYB5D2 is associated with breast cancer progression. Sci Rep. 2019; 9:6624. https://doi.org/10.1038/s41598-019-43006-y PMID:31036830

25. Liao WT, Wang X, Xu LH, Kong QL, Yu CP, Li MZ, Shi L, Zeng MS, Song LB. Centromere protein H is a novel prognostic marker for human nonsmall cell lung cancer progression and overall patient survival. Cancer. 2009; 115:1507–17. https://doi.org/10.1002/cncr.24128 PMID:19170237

26. Ito S, Kawano Y, Katakura H, Takenaka K, Adachi M, Sasaki M, Shimizu K, Ikenaka K, Wada H, Tanaka F. Expression of MAGE-D4, a novel MAGE family antigen, is correlated with tumor-cell proliferation of non-small cell lung cancer. Lung Cancer. 2006; 51:79–88. https://doi.org/10.1016/j.lungcan.2005.08.012 PMID:16225959

27. Xiong R, Wu Y, Wu Q, Muskhelishvili L, Davis K, Tripathi P, Chen Y, Chen T, Bryant M, Rosenfeldt H, Healy SM, Cao X. Integration of transcriptome analysis with pathophysiological endpoints to evaluate cigarette smoke toxicity in an in vitro human airway tissue model. Arch Toxicol. 2021; 95:1739–61. https://doi.org/10.1007/s00204-021-03008-0 PMID:33660061

28. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014; 511:543–50. https://doi.org/10.1038/nature13385 PMID:25079552

29. Sears CR, Mazzone PJ. Biomarkers in Lung Cancer. Clin Chest Med. 2020; 41:115–27. https://doi.org/10.1016/j.ccm.2019.10.004 PMID:32008624

30. Sun J, Zhao T, Zhao D, Qi X, Bao X, Shi R, Su C. Development and validation of a hypoxia-related gene signature to predict overall survival in early-stage lung adenocarcinoma patients. Ther Adv Med Oncol. 2020; 12:1758835920937904. https://doi.org/10.1177/1758835920937904 PMID:32655701

31. Yu L, Xiao Z, Tu H, Tong B, Chen S. The expression and prognostic significance of Drp1 in lung cancer: A bioinformatics analysis and immunohistochemistry. Medicine (Baltimore). 2019; 98:e18228. https://doi.org/10.1097/MD.0000000000018228 PMID:31770286

32. Wu N, Liu S, Li J, Hu Z, Yan S, Duan H, Wu D, Ma Y, Li S, Wang X, Wang Y, Li X, Lu X. Deep sequencing reveals the genomic characteristics of lung adenocarcinoma presenting as ground-glass nodules (GGNs). Transl Lung Cancer Res. 2021; 10:1239–55. https://doi.org/10.21037/tlcr-20-1086 PMID:33889506

33. Duma N, Santana-Davila R, Molina JR. Non-Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment. Mayo Clin Proc. 2019; 94:1623–40. https://doi.org/10.1016/j.mayocp.2019.01.013 PMID:31378236

34. Shi R, Bao X, Unger K, Sun J, Lu S, Manapov F, Wang X, Belka C, Li M. Identification and validation of hypoxia-derived gene signatures to predict clinical outcomes and therapeutic responses in stage I lung adenocarcinoma patients. Theranostics. 2021; 11:5061–76. https://doi.org/10.7150/thno.56202 PMID:33754044

35. Villalobos P, Wistuba II. Lung Cancer Biomarkers. Hematol Oncol Clin North Am. 2017; 31:13–29. https://doi.org/10.1016/j.hoc.2016.08.006 PMID:27912828

36. Westcott PM, To MD. The genetics and biology of KRAS in lung cancer. Chin J Cancer. 2013; 32:63–70. https://doi.org/10.5732/cjc.012.10098 PMID:22776234

37. Arora M, Kumari S, Singh J, Chopra A, Chauhan SS. Downregulation of Brain Enriched Type 2 MAGEs Is Associated With Immune Infiltration and Poor Prognosis in Glioma. Front Oncol. 2020; 10:573378. https://doi.org/10.3389/fonc.2020.573378 PMID:33425727

38. Sasaki M, Nakahira K, Kawano Y, Katakura H, Yoshimine T, Shimizu K, Kim SU, Ikenaka K. MAGE-E1, a new member of the melanoma-associated antigen gene family and its expression in human glioma. Cancer Res. 2001; 61:4809–14. PMID:11406556

39. Tabatabaei Yazdi SA, Safaei M, Gholamin M, Abdollahi A, Nili F, Jabbari Nooghabi M, Anvari K, Mojarrad M. Expression and Prognostic Significance of Cancer/Testis Antigens, MAGE-E1, GAGE, and SOX-6, in Glioblastoma: An Immunohistochemistry Evaluation. Iran J Pathol. 2021; 16:128–36. https://doi.org/10.30699/IJP.2020.125038.2368 PMID:33936223

40. Tu Y, Chen C, Fan G. Association between the expression of secreted phosphoprotein - related genes and prognosis of human cancer. BMC Cancer. 2019; 19:1230. https://doi.org/10.1186/s12885-019-6441-3 PMID:31849319

41. Guo Z, Huang J, Wang Y, Liu XP, Li W, Yao J, Li S, Hu W. Analysis of Expression and Its Clinical Significance of the Secreted Phosphoprotein 1 in Lung Adenocarcinoma. Front Genet. 2020; 11:547. https://doi.org/10.3389/fgene.2020.00547 PMID:32595698

42. Luo X, Feng L, Xu W, Bai X, Wu M. Weighted gene co-expression network analysis of hub genes in lung adenocarcinoma. Evol Bioinform Online. 2021; 17:11769343211009898. https://doi.org/10.1177/11769343211009898 PMID:33911849

43. Zhang Y, Du W, Chen Z, Xiang C. Upregulation of PD-L1 by SPP1 mediates macrophage polarization and facilitates immune escape in lung adenocarcinoma. Exp Cell Res. 2017; 359:449–57. https://doi.org/10.1016/j.yexcr.2017.08.028 PMID:28830685

44. Onaga C, Tamori S, Motomura H, Ozaki A, Matsuda C, Matsuoka I, Fujita T, Nozaki Y, Hara Y, Kawano Y, Harada Y, Sato T, Mano Y, et al. High *SLC20A1* Expression Is Associated With Poor Prognoses in Claudin-low and Basal-like Breast Cancers. Anticancer Res. 2021; 41:43–54. https://doi.org/10.21873/anticanres.14750 PMID:33419798

45. Shen T, Wang M, Wang X. Identification of Prognosis-related Hub RNA Binding Proteins Function through Regulating Metabolic Processes in Tongue Cancer. J Cancer. 2021; 12:2230–42. https://doi.org/10.7150/jca.52156 PMID:33758601

46. Dong Z, Wang J, Zhan T, Xu S. Identification of prognostic risk factors for esophageal adenocarcinoma using bioinformatics analysis. Onco Targets Ther. 2018; 11:4327–37. https://doi.org/10.2147/OTT.S156716 PMID:30100738

47. Sun GZ, Zhao TW. Lung adenocarcinoma pathology stages related gene identification. Math Biosci Eng. 2019; 17:737–46. https://doi.org/10.3934/mbe.2020038 PMID:31731374

48. Qian L, Luo Q, Zhao X, Huang J. Pathways enrichment analysis for differentially expressed genes in squamous lung cancer. Pathol Oncol Res. 2014; 20:197–202. https://doi.org/10.1007/s12253-013-9685-2 PMID:24114512

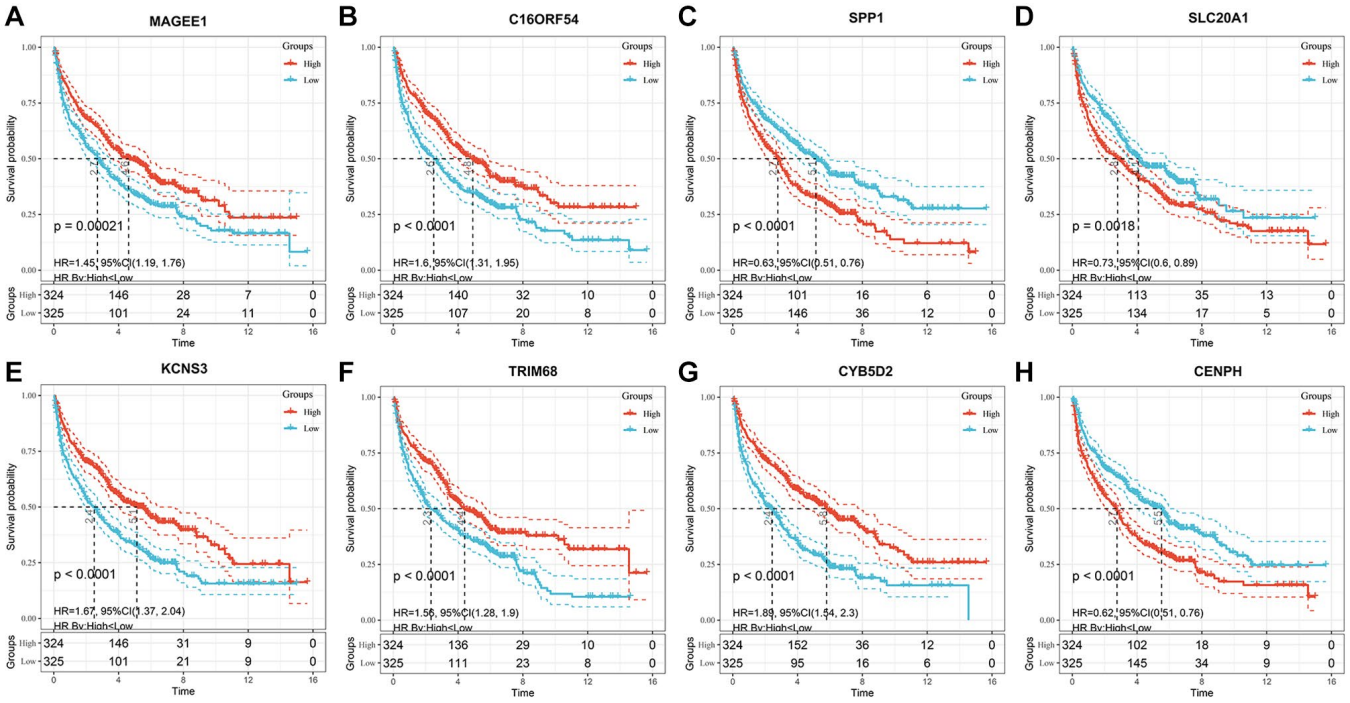
49. Kikuchi J, Kinoshita I, Shimizu Y, Kikuchi E, Takeda K, Aburatani H, Oizumi S, Konishi J, Kaga K, Matsuno Y, Birrer MJ, Nishimura M, Dosaka-Akita H. Minichromosome maintenance (MCM) protein 4 as a marker for proliferation and its clinical and clinicopathological significance in non-small cell lung cancer. Lung Cancer. 2011; 72:229–37. https://doi.org/10.1016/j.lungcan.2010.08.020 PMID:20884074

50. Li Z, Xu X. Post-Translational Modifications of the Mini-Chromosome Maintenance Proteins in DNA Replication. Genes (Basel). 2019; 10:331. https://doi.org/10.3390/genes10050331 PMID:31052337

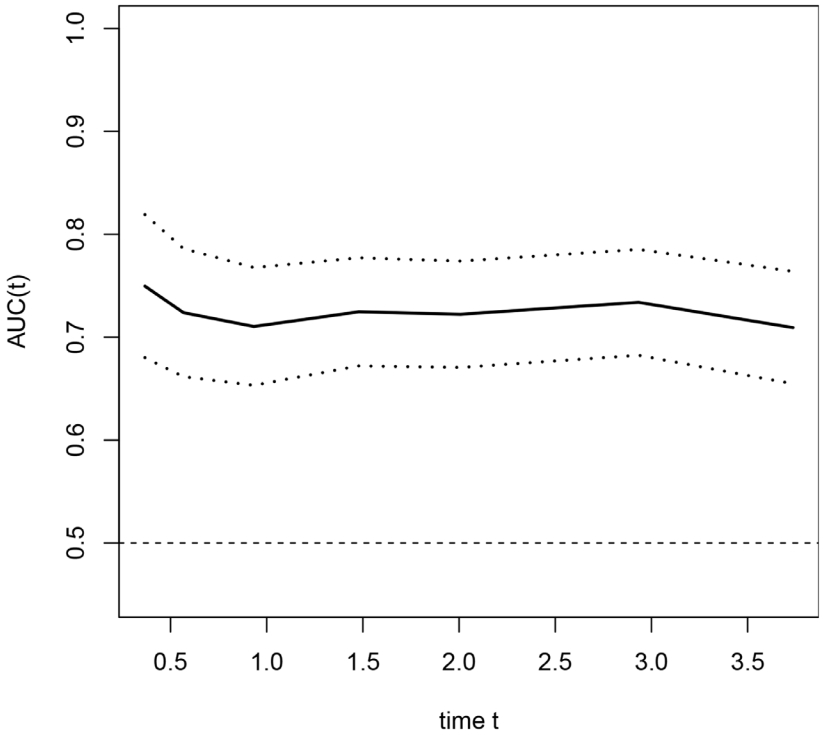
51. Zhang H, Tulahong A, Wang W, Nuerrula Y, Zhang Y, Wu G, Mahati S, Zhu H. Downregulation of microRNA-519 enhances development of lung cancer by mediating the E2F2/PI3K/AKT axis. Int J Clin Exp Pathol. 2020; 13:711–20. PMID:32355519

SUPPLEMENTARY MATERIALS

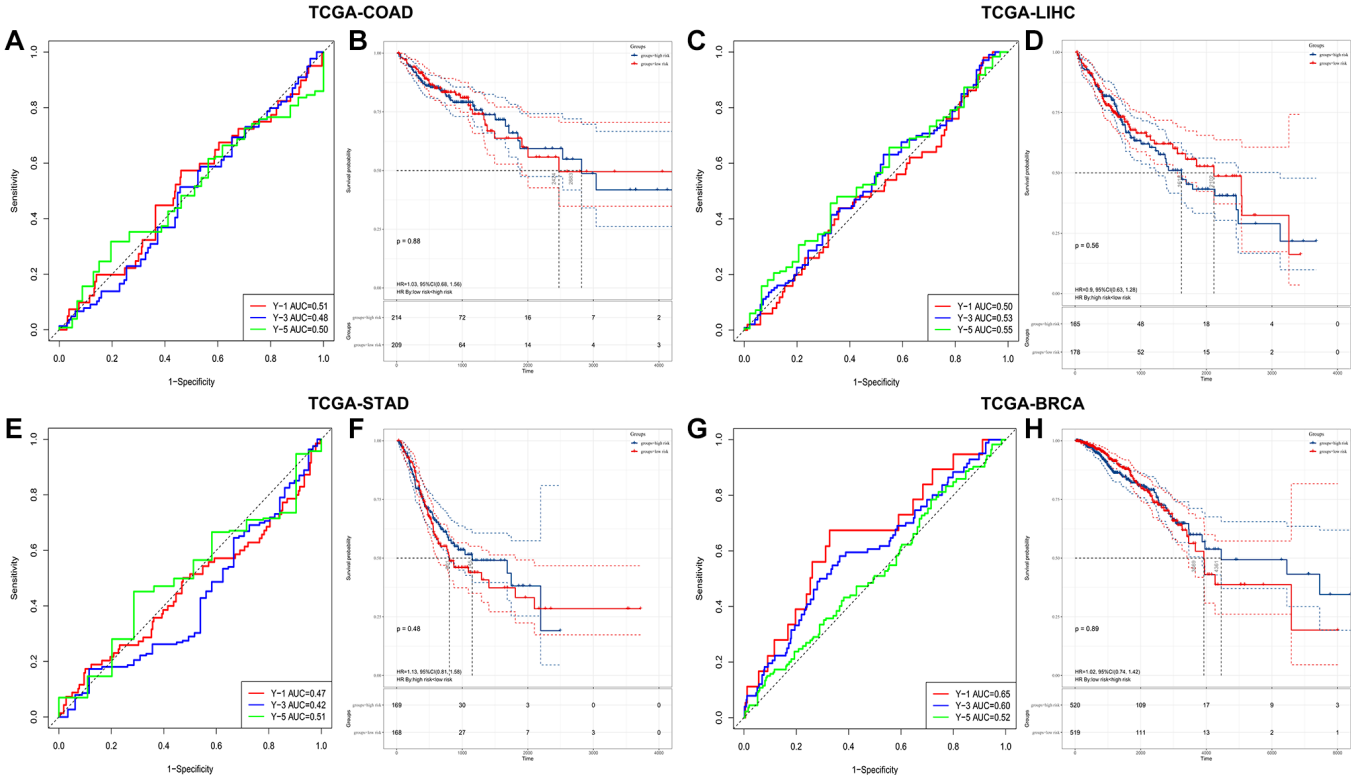
Supplementary Figures



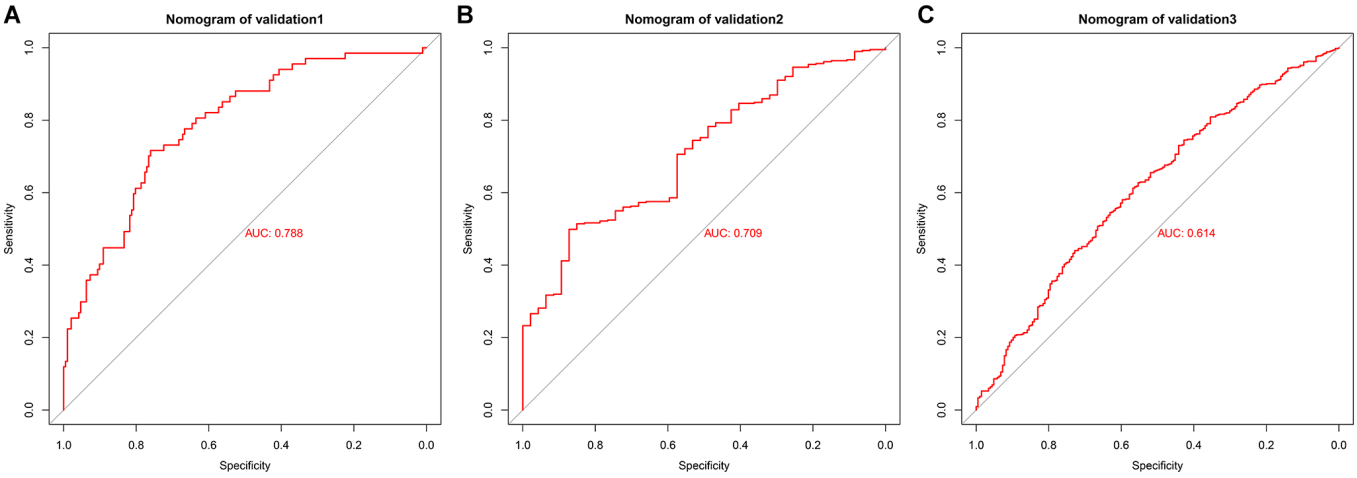
Supplementary Figure 1. The K-M survival curve analyses of 8-genes. (**A**) MAGEE1; (**B**) C16ORF54; (**C**) SPP1; (**D**) SLC20A1; (**E**) KCNS3; (**F**) TRIM68; (**G**) CYB5D2; (**H**) CENPH.



Supplementary Figure 2. The tROC analysis of the PRS.



Supplementary Figure 3. The PRS as a valuable predictor for OS in other cancers. (**A**, **B**) TCGA-COAD; (**C**, **D**) TCGA-LIHC; (**E**, **F**) TCGA-STAD; (**G**, **H**) TCGA-BRCA.



Supplementary Figure 4. ROC analysis of the complex model in validation sets. (**A**) the validation set I; (**B**) the validation set II; (**C**) the validation set III.

Supplementary Table

Supplementary Table 1. Information on the datasets in the training and validation sets.

Cohorts	Datasets	Year	Country	Sample	*N*
Training set	GSE30219	2011	France	lung cancer	274
	GSE37745	2012	Sweden	NSCLC	196
	GSE50081	2013	Canada	NSCLC	181
Validation set I	GSE29013	2011	USA	NSCLC	55
	GSE31210	2011	Japan	LUAD	204
Validation set II	GSE41271	2012	USA	lung cancer	268
	GSE42127	2012	USA	NSCLC	173
Validation set III	TCGA-LUAD	2014	USA	LUAD	494
	TCGA-LUSC	2014	USA	LUSC	480
Specificity validation set	TCGA-LIHC	2015	USA	LIHC	343
	TCGA-COAD	2015	USA	COAD	423
	TCGA-STAD	2015	USA	STAD	337
	TCGA-BRCA	2015	USA	BRCA	1039