European Review for Medical and Pharmacological Sciences 2022; 26: 3807-3826

Construction of a hypoxic immune microenvironment associated gene-based model for prognosis prediction of lung adenocarcinoma

G.-Y. LIN, S. WU, Z.-S. GAO, L.-H. WU, J.-J. YAN, X.-Q. GUO, Z.-Y. WANG

Department of Respiratory and Critical Illness Medicine, The First Hospital of Putian, Putian, China

Abstract. – OBJECTIVE: Lung adenocarcinoma (LUAD) accounts for the majority of cancer deaths worldwide, with a high incidence rate and mortality. It is highly important to develop biomarker model to accurately predict the prognosis.

MATERIALS AND METHODS: RNA-Seq data and clinical follow-up data of LUAD were downloaded from The Cancer Genome Atlas (TCGA) database. Hypoxia-related gene sets were collected from the Gene Set Enrichment Analysis (GSEA) website. A gene signature model was established using the Limma package in the R software, univariate and multivariate survival analyses, and least absolute shrinkage and selection operator (LASSO) algorithms.

RESULTS: Two hypoxia subtypes (C1 and C2) were classified according to the expressions of 55 prognostic hypoxic-related genes. Differentially expressed genes (DEGs) between two hypoxia subtypes and immune group were analyzed. Then, 390 DEGs related to hypoxic immune microenvironment were filtered. According to hypoxia type and immune type, the samples were classified into hypoxia-high & immune-low group, hypoxia-low & immune-high group. Based on these differentially expressed genes (DEGs), a 5-genes signature model, which showed a stable prediction performance on datasets of different platforms and immunotherapy datasets, was finally developed. Meanwhile, it demonstrated a better performance compared with other existing models. The AUC of the 5-gene signature was high in both the training dataset and 4 independent validation datasets and was confirmed as a clinical feature-independent prognostic model.

CONCLUSIONS: This study developed a hypoxic immune microenvironment associated gene-based model for prognostic prediction of LUAD, providing clinicians with a reliable prognostic assessment tool and facilitating clinical treatment decision-making.

Key Words:
Lung adenocarcinoma, Immune, Hypoxia, Riskscore, The cancer genome atlas, Nomogram.

Introduction

Lung cancer is a fastest-growing malignancy with high morbidity and mortality. Non-small cell lung cancer (NSCLC) as the most common histologic type of lung cancer accounts for about 85% of all lung cancer cases[1]. NSCLC is divided into lung adenocarcinoma (LUAD) and squamous cell carcinoma[2]. At present, the incidence of LUAD has exceeded squamous cell carcinoma to become the main type of lung cancer[3]. There have been significant advances in the diagnosis and classification of LUAD, and tumor treatment has evolved from initial surgical resection, surgical combined adjuvant chemoradiotherapy, and targeted therapy to immunotherapy, however, only a small number of patients can benefit from it[4,5]. Hypoxia is a physical microenvironment of lung cancer cells that can cause changes in signaling pathways in lung cancer cells, further promoting tumor metastasis and angiogenesis. Therefore, hypoxia is regarded as one of the important factors decisive in the prognosis of malignant tumors[6]. Hence, identifying reliable prognostic tools to predict clinical outcomes and facilitate treatment decision-making is clearly of paramount importance.

At present, computer science and information technology have witnessed a rapid development, and are widely used in many fields, which also lays a solid foundation for the research of life science[7]. Driven by the development of molecular biology technology, the Human Genome Project has produced a large amount of biological data,

Corresponding Author: ZhiYong Wang, MD; e-mail: wzydoctor@163.com

and on such a basis, bioinformatics was produced[8] to offer support for the study of specific drug use and prognosis of patients with a specific disease based on gene expression[9,10]. However, single-parameter prediction of prognosis survival in LUAD patients is challenging because of the influence of extensive variability on prognosis and genetic heterogeneity. Therefore, exploring a prognostic evaluation system with a combination of multiple biomarkers may help improve the accuracy of prediction.

At present, the prognostic model based on multiple genes has been widely studied in predicting overall survival of patients with LUAD and has a greater diagnostic significance than the single-parameter prediction model[11,12]. In the present study, hypoxia-related genes not only participate in the disease process but can also be used to predict the severity of the disease[13-15]. Moreover, immune related genes have also been found to have certain value in the prediction of disease prognosis[16,17]. In this study, we found that through different bioinformatics analysis process more hypoxia and immune-related genes were associated with prognosis. Moreover, according to the low oxygen immune microenvironment influence on cancer cell proliferation expansion, LUAD hypoxia-related genes and immune-related genes were chosen to establish a prognosis of clinical parameters, which demonstrated a higher accuracy than the traditional model.

Materials and Methods

Data Source and Preprocessing

LUAD RNA-seq data and clinical follow-up information were downloaded from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). The expression profile was converted to TPM, with genes with low expression levels being removed (genes with expression levels below 1 accounted for more than 50% of all samples). Ensembl ID was converted to Gene Symbol. If multiple Ensembl IDs correspond to the same gene Symbol, the median was taken as the expression profile of the gene symbol. Log2 conversion was performed on the expression profile data.

The raw data of three datasets, including GSE31210[18], GSE30219[19] and GSE50081[20], were downloaded from Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). All the three datasets were sequenced by GPL570 platform ([HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array). The RMA function (Robust multi-array Average expression measure) of R-packet Affy (V1.66.0) was used for processing spectral data[21]. The expression profile of the dataset was obtained by normalization. The probe was converted to a Gene Symbol chip according to the data set based on the GPL570 annotation file. Specifically, when multiple probes correspond to the same Gene Symbol, the median value was taken as the expression profile of the Gene Symbol. When one probe corresponds to multiple gene symbols, the probe expression was removed. Only LUAD tumor samples with survival time and status were kept in the microarray dataset. The samples with survival information in immunotherapy dataset Imvigor210[22] were retained. Clinical information after data preprocessing was showed in Table I.

A total of 200 hypoxia-related genes were obtained from HALLMARK_HYPOXIA pathways in the document H.A.L. v7.3.symbols. gmt. on the gene set enrichment analysis (GSEA) website[23]. The workflow was showed in Figure 1.

Identification of Hypoxia Subtype and Hypoxic-Related Differentially Expressed Genes

ConsensusClusterPlus was used to identify anoxic molecular subtypes. Expression profile data in TCGA dataset were used to filter genes with low expression (expression level below 1 was in more than 50% in all samples), and then the expression level of hypoxia gene was

Table I. Risk coefficient of risk model and multivariate COX results.

Gene	coef	HR	HR (lower, 0.95)	HR (upper, 0.95)	p
MS4A1	-0.148	0.862	0.748	0.994	0.042
CPA3	-0.148	0.863	0.764	0.975	0.018
FSCN1	0.173	1.190	1.047	1.352	0.008
PTPRH	0.184	1.203	1.053	1.374	0.007
DKK1	0.110	1.116	1.032	1.207	0.006

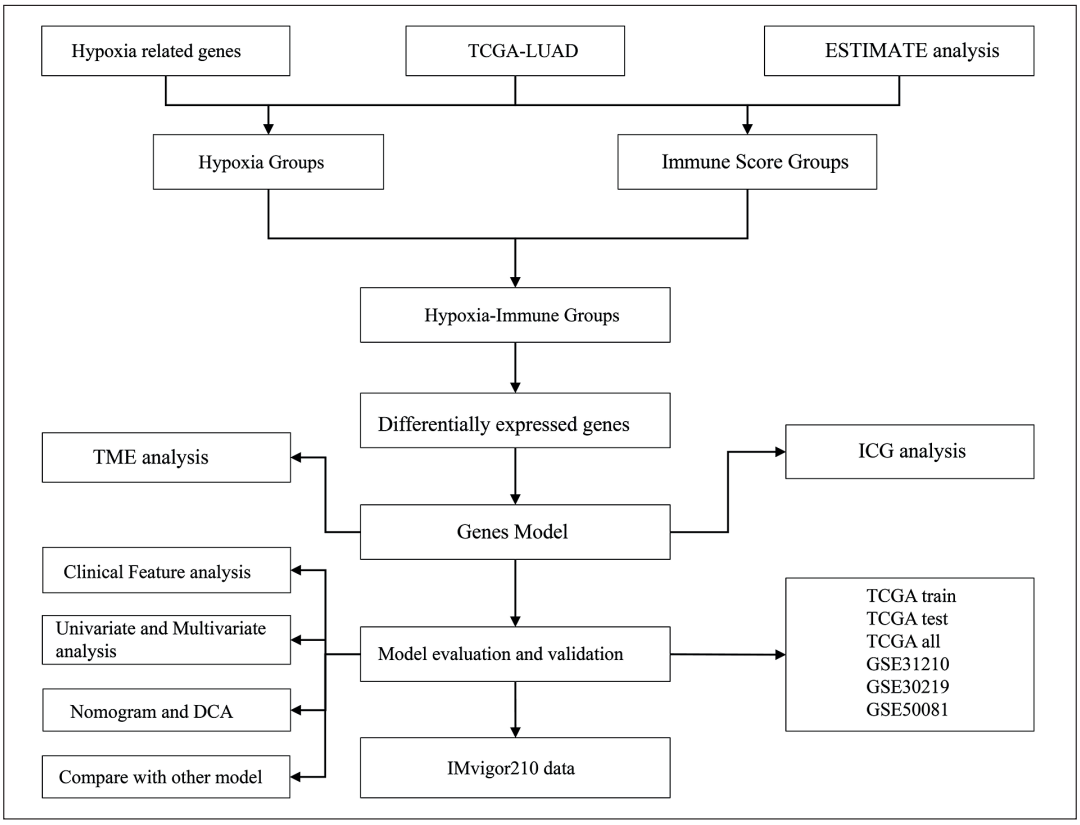

Figure 1. Flowchart of the whole study.

extracted. Based on 200 hypoxia-related genes, univariate Cox analysis was performed using the Coxph function in R to obtain DEGs related to the prognosis of LUAD ($p < 0.05$). ConsensusClusterPlus in R Package (With Volume Parameters: Distance = "Euclidean", clusterAlg = "km", and innerLinkage = 'ward.d2') were applied to classify samples base on these hypoxic-related DEGs. The subtype allocation was then verified using the mRNA expression data of the above hypoxic genes by applying random neighbor embedding (T-SNE) method based on the T-distribution. R software package GSVA was used to perform single-sample GSEA analysis, and the scores of each sample on different functions of Hallmark dataset were calculated.

Identification of Immune Status and Immune-Related DEGs

R software package Estimation of Stromal and Immune cells in MAlignant Tumor tissues using Expression data (ESTIMATE) was used to calculate the ImmuneScore of each sample in the TCGA dataset. The optimal segmentation point was determined by SurvMiner package, and the samples were divided into two groups of high and low immunization.

Identification of Hypoxia-Immune-Related Prognostic DEGs

The hypoxia and immune status identified above was further combined into a two-dimension index, whereby patients were divided into three groups ($Hypoxia^{High}$ & $Immune^{Low}$, $Hypoxia^{Low}$ & $Immune^{High}$, and Mixed groups). Limma packet analysis was conducted to identify DEGs between $Hypoxia^{High}$ & $Immune^{Low}$ and $Hypoxia^{Low}$ & $Immune^{High}$ group according to the threshold FDR < 0.05 and | log2 fold change (FC) | > 1.

Functional Enrichment Analysis of DEGs

With the screened DEGs, gene ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways analysis were performed on R software package Clusterprofiler (v3.16.0). A $p<0.05$ was considered as statistically significant.

Establishment and Validation of a Prognostic Predictive Signature

The univariate Cox regression analysis was conducted to analyze the prognostic value of the DEGs for overall survival (genes of a $p<0.05$ being considered as statistically significant). Subsequently, the range of genes was narrowed by Lasso Cox regression analysis was performed using Glmnet in R package. Finally, multivariate Cox regression analysis was performed to identify closely correlated genes. And the selected genes were recruited into the prognostic model, and the calculation formula of the prognostic risk model was as follows:

$$\text{RiskScore} = \sum_{i=1}^{n} coef\ (i) * expression \qquad (i)$$

Where coef was the coefficient of a gene, i represented the gene, and expression (i) represented the gene expression. RiskScore cutoff was bounded by the median value, and those samples with a value greater than the median value were seen as high-risk samples, while those samples less than or equal to the median value were low-risk samples.

The risk calculation method was adopted to verify the hypoxia data from TCGA training dataset, TCGA validation dataset, entire TCGA data sets, and completely independent GSE30219, GSE31210, and GSE31210 datasets. Imvigor210 acted as a validation immunotherapy dataset.

Comparison of RiskScore Between Clinical Features

Based on the median RiskScore, all the samples were divided into high-risk group and low-risk group. Kaplan-Meier survival analysis using Log-rank test was used to compare survival differences between the two groups. Univariate Cox proportional risk regression analysis was used to analyze the effects of RiskScore, T stage, N stage, M stage, AJCC stage, grade, grouping, gender, age, smoking and other clinical parameters on LUAD prognosis. In addition, whether above features were an independent prognostic factor for LUDA was studied through conducting multivariate Cox proportional risk regressions.

Comparison of Immune Scores Between High and Low Risk Groups

We used immune assessment softwares, ESTIMATE, MCPCounter[24,25], TIMER[26], and ssGSEA of GSVA to evaluate the immune infiltration. Also, immune checkpoints, such as PDCD1 and CTLA4, for immunotherapy[24,27,28] were compared in high- and low-risk groups to analyze the possibility of immunotherapy.

Reliability of Prognostic Models

We performed univariate and multivariate analysis with clinical characteristics and RiskScore to validate the stability of our RiskScore model. Based on the univariate and multivariate results, we constructed a Nomogram and the corrected histogram to provide a basis for clinical diagnosis and prognosis of LUAD. At the same time, the decision curve analysis (DCA) diagram was drawn. In addition, we collected published models for comparison of receiver operating characteristic (ROC) curves.

Statistical Analysis

Heatmap of DEGs was plotted using "pheatmap" R package with zero-mean normalization. Two groups of boxplot were analyzed by Wilcoxon-test. For Kaplan-Meier curves, *p*-values and hazard ratio (HR) with 95% confidence interval (CI) were generated by log-rank tests and univariate Cox proportional hazards regression. All analytical methods mentioned above were performed on R package software version 4.2.2. All statistical tests were two-sided. $p<0.05$ was considered as statistically significant.

Results

Identification of Hypoxic Molecular Subtypes

Firstly, 182 hypoxia genes were obtained from TCGA expression profile data. Univariate COX survival analysis was performed on the 182 hypoxia genes using the COXPH function in R, and 55 hypoxia related DEGs were identified to be related to the prognosis of LUAD. Based on these 55 hypoxia related DEGs, the LUAD samples in TCGA were well classified into C1 subtype (257 samples) and C2 subtype (243 samples) (Figure 2A). Then, the T-distribution-based random neighbor embed (T-SNE) method was used to verify the subtype allocation using the mRNA expressions of 55 hypoxia genes, and the two subtypes could be clearly classified (Figure 2B). In addition, KM survival analysis demonstrated that LUAD samples in C1 group had a higher survival rate in comparison to C2 subtype ($p<0.0001$,

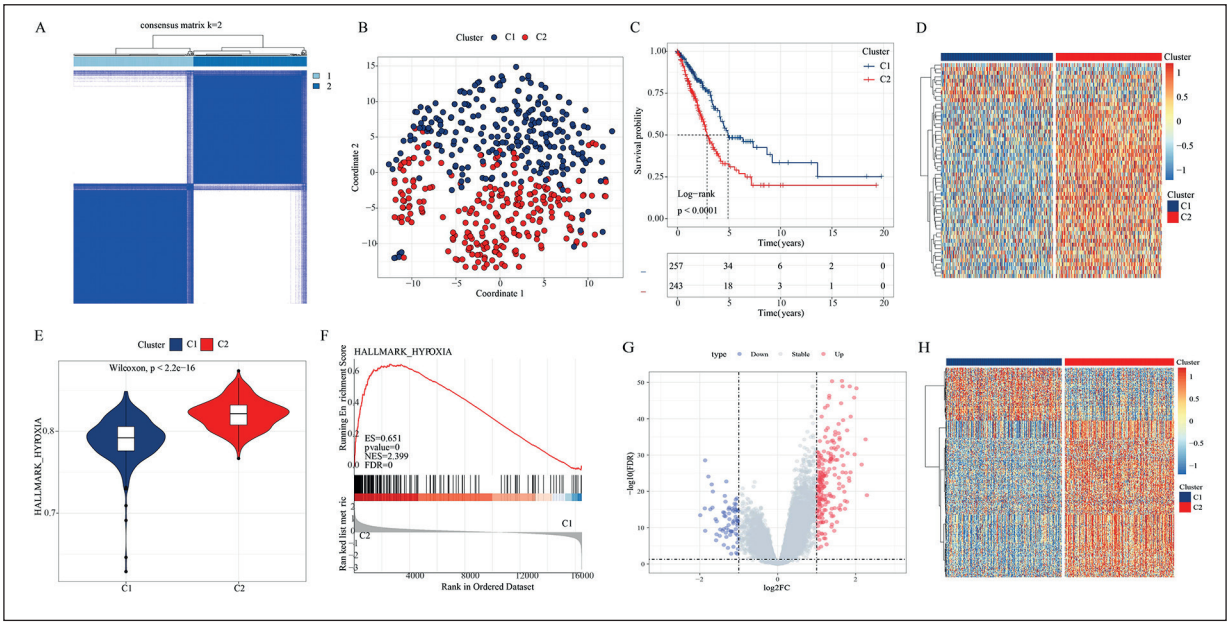

Figure 2. Identification of hypoxia molecular subtypes. **A**, ConsensusClusterPlus identifies hypoxic molecular subtypes. **B**, T-distribution random neighbor embedding (T-SNE) method to distinguish hypoxia subtypes. **C**, Kaplan-Meier survival analysis on the effects of hypoxia subtypes on patient survival. **D**, Volcanic map of differentially expressed genes in hypoxia subtypes. **E**, A single sample GSEA analysis was performed using the R software package GSVA to calculate the scores of each sample on different functions in the Hallmark dataset. **F**, The GSEA input file contains the expression profile data of TCGA, the sample labels of molecular subtype markers, and the sample label marks the sample as group C1 or group C2. **G**, Volcanic map of differential genes between subtypes. (**H**) Heat maps of genes that differ between subtypes.

Figure 2C). A heat map of 55 hypoxic gene expressions showed that most of the genes were higher-expressed in the C2 subtype (Figure 2D).

We compared the difference in HALLMARK_HYPOXIA score using ssGSEA analysis between the C1 and C2 subgroups and found that the Hypoxia score in C2 was significantly higher than that in C1 (Figure 2E). Functional enrichment results showed that HALLMARK_HYPOXIA was significantly enriched in the C2 subtype (Figure 2F). Based on the above analysis, C2 subtype was defined as the hypoxia-high expression group, whereas C1 was the hypoxia-low expression group.

Next, limma packet analysis was conducted to determine DEGs between C1 and C2 subgroups, 285 DEGs including 213 up-regulated genes and 72 down-regulated genes were identified (Figure 2G, H).

Classification of Immune Molecular Subtypes

First, we used R software package ESTIMATE to calculate the ImmuneScore of samples in the TCGA dataset, and the samples were divided into two categories, namely, immune^{High} and immune^{Low} group. KM survival analysis result showed that survival time in the immune^{High} group was significantly longer than that in the immune^{Low} group (Figure 3A). Next, 403 DEGs were obtained between immune^{High} and immune^{Low} group using Limma package, and here 10 genes were upregulated, and 393 genes were down-regulated in immune^{High} group (Figure 3B, C).

Identification of Hypoxic Immune Microenvironment Group

According to the above hypoxic and immune status, LUAD patients were classified into three groups (Hypoxia^{High} & Immune^{Low}, Hypoxia^{Low} & Immune^{High}, and Mixed groups). The samples in Hypoxia^{High} & Immune^{Low} group showed a shorter survival time compared to Hypoxia^{Low} & Immune^{High} group (Figure 4A). 645 DEGs were identified between Hypoxia^{High} & Immune^{Low} group and Hypoxia^{Low} & Immune^{High} group, and here 271 genes were upregulated, and 374 genes were down-regulated in Hypoxia^{High} & Immune^{Low} (Figure 4B). The Venn diagram results showed that there were 133 risk DEGs and 257 protective DEGs (Figure 4C, D).

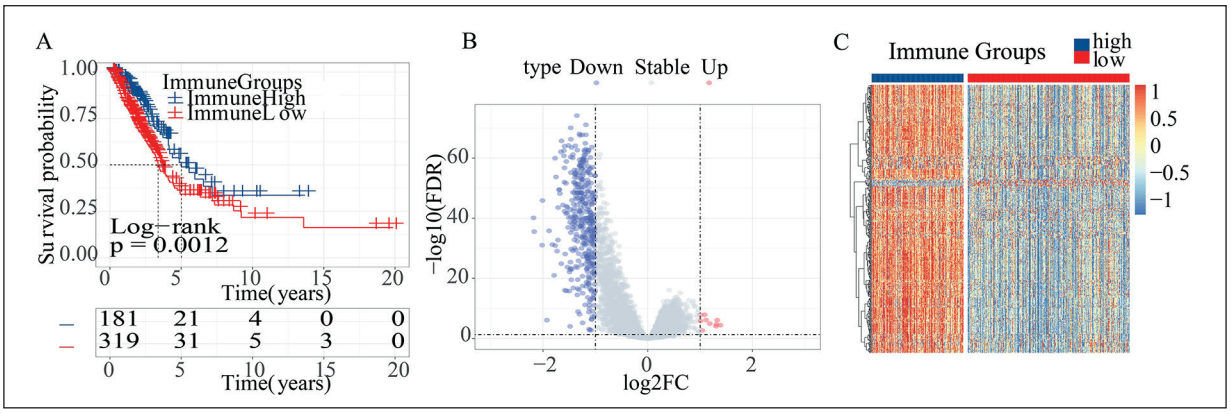

Figure 3. Immune infiltration score group analysis. **A**, SurvMiner package was used to determine the optimal segmentation point (immunescore_cutof. PDF), which was divided into two groups of high and low immunization, and then the survival curves between immunization groups were calculated. **B**, Volcanic map of differential genes in immune grouping. **C**, Heat maps of differential genes in immune groupings.

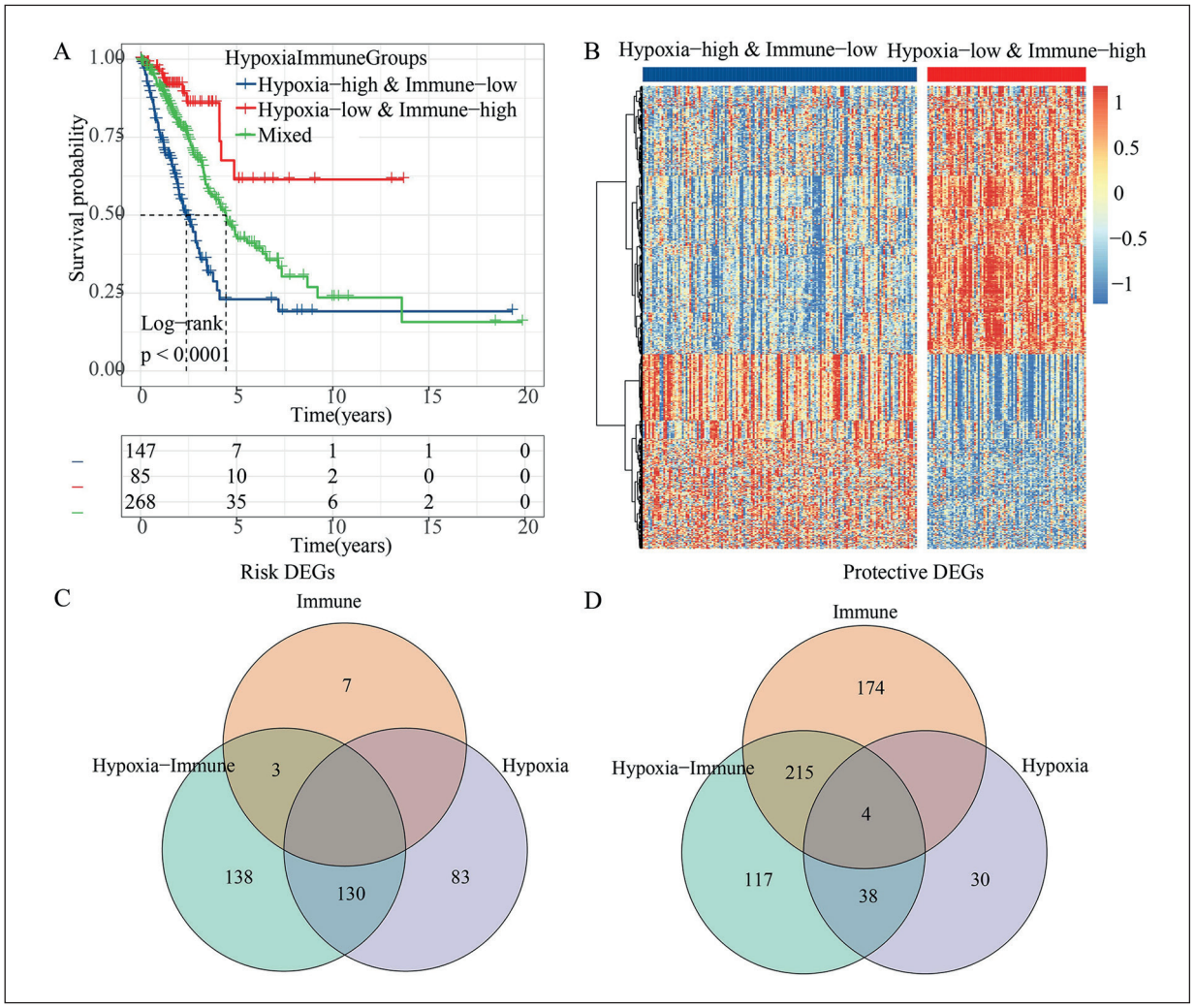

Figure 4. Identification of hypoxic immune microenvironment group. **A**, Survival curve of hypoxic immune microenvironment grouping. **B**, Differential gene heat map of hypoxic immune microenvironment. **C**, Identification of risk differential genes. **D**, Identification of protective differential genes.

Functional Analysis of Protective DEGs and Risk DEGs

KEGG pathway analysis and GO enrichment analysis were performed on the protective DEGs and risk DEGs using R software package Clusterprofiler (v3.16.0). For GO and KEGG analyses of the protective DEGs, the functions annotated were significantly correlated with immunity pathways (**Supplementary Figure 1**). For the GO and KEGG analyses of the risk DEGs, the functions and pathways annotated were significantly associated with p53 signaling pathway and PI3K-Akt signaling pathway (**Supplementary Figure 1**).

Establishment of Prognostic Risk Models

Univariate cox analysis of 390 DEGs (risk DEGs and protective DEGs) on the training dataset revealed 103 prognostic genes correlated with LUAD ($p < 0.05$). Then, R software package glmnet was introduced to carry out LASSO regression analysis[29]. Firstly, the change trajectory of each independent variable was analyzed, as shown in Figure 5A. With the gradual increase of lambda, the number of independent variable coefficients approaching 0 also increased gradually. 10-fold cross validation was applied to build the model and analyzed the confidence interval under each lambda. The model reached the optimal when lambda=0.061 (Figure 5B). Therefore, we selected 12 genes when lambda=0.061 for further analysis. Next, multivariate cox analysis was performed on the 12 genes, and the number of genes was further reduced by stepAIC method. Finally, 5 genes were retained to build prognostic model (Table II). The 5 genes could effectively distinguish samples as high-risk and low risk (**Supplementary Figure 2**).

The RiskScore was constructed based on their cox coefficients of the 5 genes: RiskScore = -0.148*Exp (MS4A1)-0.148*Exp (CPA3)+0.173*-Exp (FSCN1)+0.184 *Exp (PTPRH)+0.110*Exp (DKK1).

The RiskScore of each sample in TCGA training dataset was obtained according to the calculation formula of RiskScore, and then the median was taken as the cutoff point. Samples with RiskScore greater than the median were divided into high-risk group, while those samples with RiskScore lower than or equal to the median value were in low-risk group. The distribution of survival time was plotted from low risk to high-risk group in the TCGA training dataset, with a higher proportion of survivors distributing in the low-risk area (Figure 5C, D). Meanwhile, the expression profile of five genes was developed, with the RiskScore increasing from left to right, moreover, the results showed that MS4A1 and CPA3 were protective factors, while FSCN1, PTRRH and DKK1 were risk factors (Figure 5E). Through survival curve analysis, we found that patients in low-risk group showed a high survival rate than that in high-risk group (Figure 5F). However, at different time points (1, 3 and 5 years), the ROC curves of our RiskScore all had high AUC values above 0.7 (Figure 5G), indicating that our model had a high performance.

At the same time, to verify the reliability of our risk model, we used TCGA validation dataset and entire dataset for validation. The results showed that our risk model also had a strong performance in TCGA validation dataset (Figure 6A-E) and entire dataset (Figure 6F-J).

Meanwhile, in order to further verify our risk model on different platforms, we validated our risk model in three independent datasets of GSE30219 (Figure 7A-E), GSE31210 (Figure 7F-J) and GSE50081 (Figure 7K-O). Similarly, our risk model performed well in LUAD prognostic prediction.

RiskScore Analysis of Clinical Characteristics Samples

We compared the distribution of RiskScore on clinical features in TCGA dataset. The results showed that RiskScore distribution of samples in AJCC-stage, subtype, gender, age, smoking, immune group and immune hypoxia group were significantly different. Specifically, higher RiskScore was positively correlated with greater clinical characteristics indicative of a worse prognosis (Figure 8). Besides, we performed risk stratification on patients with AJCC-stage, subtype, gender, age, Smoking, immune group and immune hypoxia group, and conducted the KM survival analysis. The high-risk scores had worse OS than the low-risk scores in stage T1-T2, stage T3-T4, stage N0, stage N1-N3, M0, M1, stage I-II, stage III-IV, Smoking 1, Smoking 2-4, Age>65, Age=65, male, female, indicating the a strong performance of RiskScore (**Supplementary Figure 3**).

RiskScore Was an Independent Prognostic Factor

To identify the independence of the RiskScore model in clinical application, univariate and multivariate COX analyses were used to analyze the relevant HR, 95%CI of HR, and P

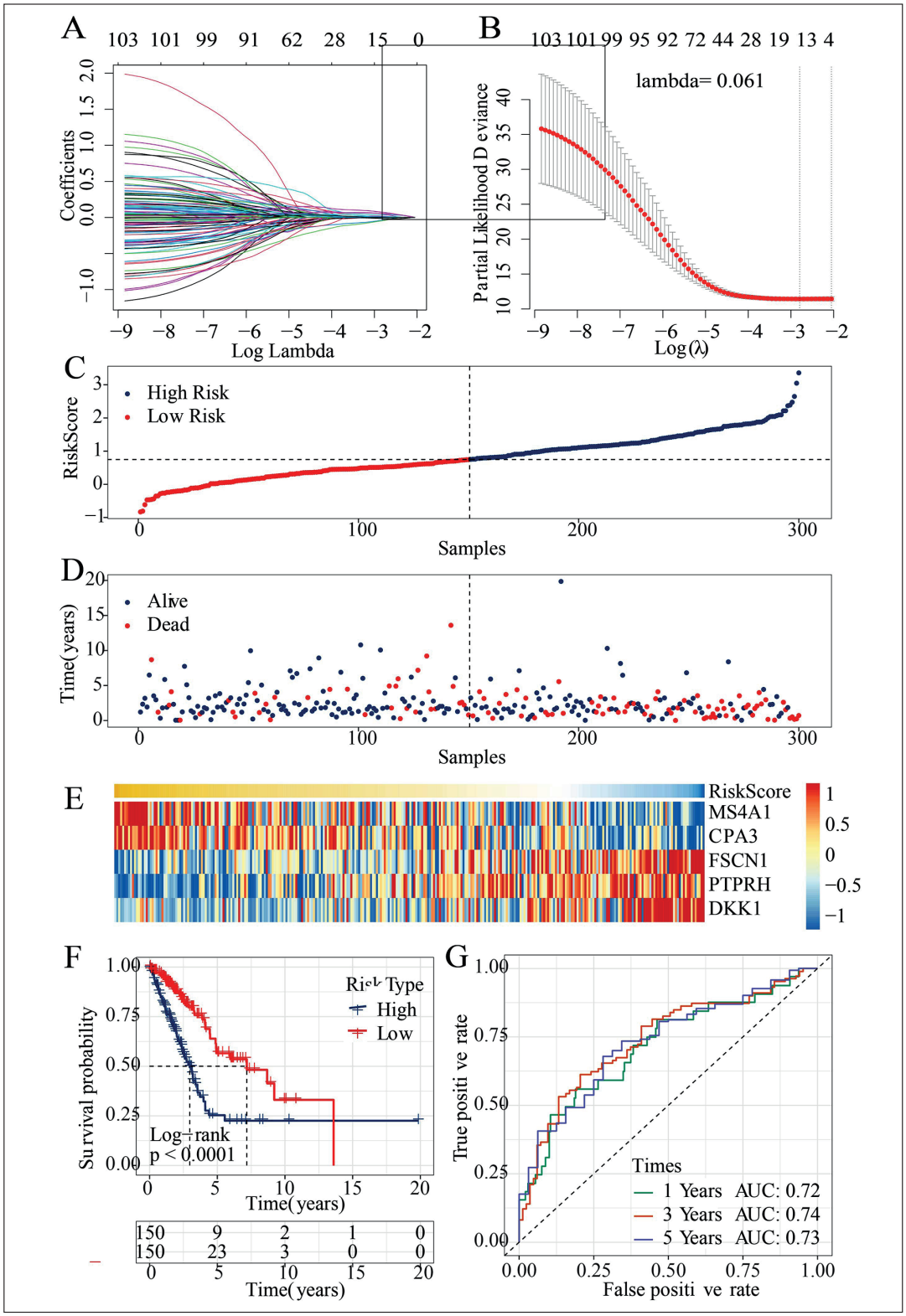

Figure 5. Survival time distribution of TCGA training datasets from low risk to high risk. **A**, LASSO coefficients profiles of 390 protein-coding genes. **B**, LASSO regression with 10-fold cross-validation obtained 12 prognostic genes using minimum lambda value. **C**, The curve of risk score. **D**, Survival status of the patients. More deceased patients corresponding to the higher risk score. **E**, Heatmap of the expression profiles of the five prognostic genes in low- and high-risk group. **F-G**, Kaplan-Meier survival analysis curve and ROC curve of TCGA training dataset.

Table II. Clinical information of datasets.

Features	TCGA-LUAD	GSE31210	GSE30219	GSE50081	IMvigor210
EVENT					
Alive	318	191	40	76	109
Dead	182	35	43	51	189
T Stage					
T1	167				
T2	267				
T3	45				
T4	18				
TX	3				
N Stage					
N0	324				
N1	94				
N2	69				
N3	2				
NX	11				
M Stage					
M0	332				
M1	24				
MX	144				
Stage					
I	268				
II	119				
III	80				
IV	25				
X	8				
Smoking					
1	71				
2	119				
3	129				
4	163				
5	4				
7	14				
Gender					
Male	230				
Female	270				
Age					
≤ 65	237				
> 65	253				
NA	10				
Response					
CR					25
PR					43
PD					167
SD					63
IC.Level					
IC0					83
IC1					112
IC2+					102
NA					1
TC.Level					
TC0					238
TC1					17
TC2+					42
NA					1
Immune phenotype					
Desert					69
Excluded					113
Inflamed					62
NA					54

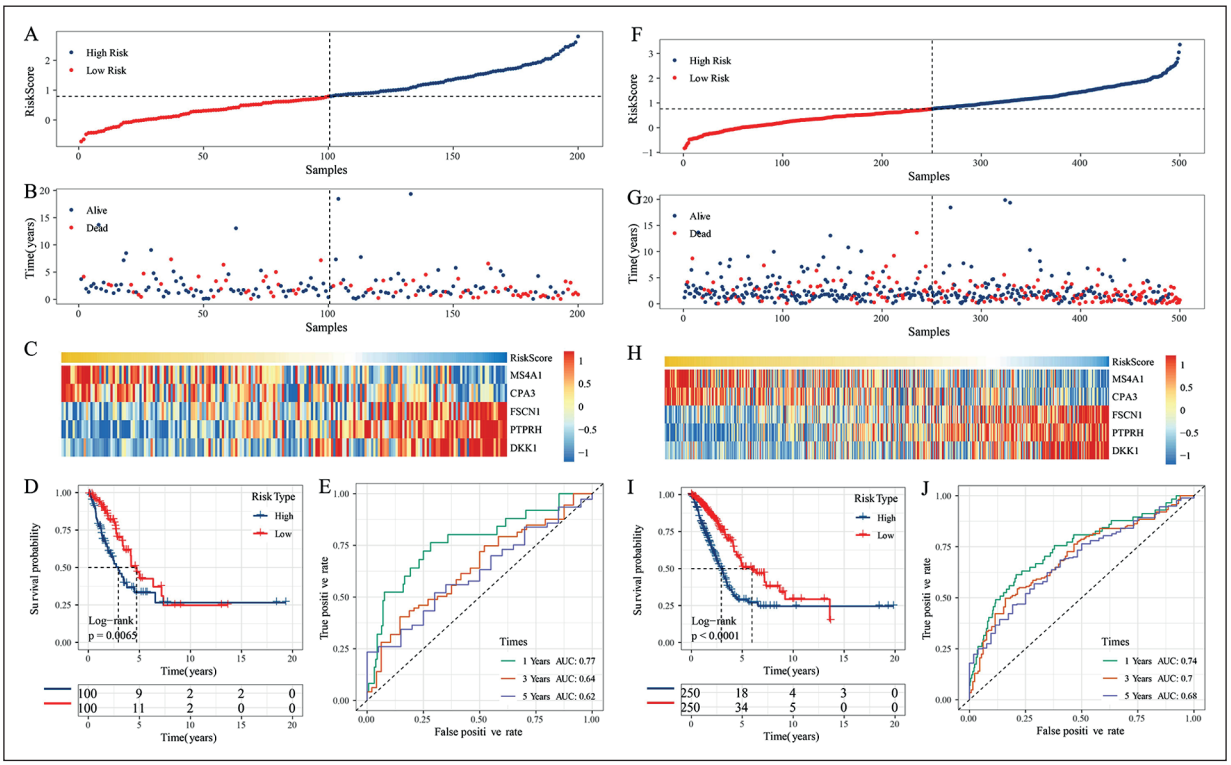

Figure 6. Validation of risk model in TCGA validation dataset and TCGA all dataset. **A-B**, TCGA verifies the distribution of RiskScore and the corresponding distribution of survival states in the dataset. **C**, Heat map of RiskScore model gene expression. **D-E**, TCGA verifies the KM curve and ROC curve of the dataset. **F-G**, The distribution of RiskScore and the corresponding distribution of survival states in all TCGA datasets. **H**, Heat map of RiskScore model gene expression. **I-J**, KM curve and ROC curve of all TCGA data sets.

values in clinical information from the entire TCGA dataset. We systematically analyzed clinical information from TCGA patient records, including age, gender, T stage, N stage, M stage, smoking, stage, and our Risktype grouping information, and univariate analysis showed that RiskType, TNM stage, Stage were significant factors associated with patients' survival ($p< 0.05$). Multivariate analysis demonstrated that RiskType, T stage and N stage were an independent prognostic factors ($p < 0.001$). Those data suggested that RiskType was also an independent prognostic (Figure 9).

RiskScore and Clinical Characteristics Were Used to Construct the Nomogram

Based on above analysis, clinical features T stage, N stage and RiskScore in entire TCGA dataset were used to construct a nomogram (Figure 10A). According to the results of the model, RiskScore feature had the greatest influence on the prediction of survival rate, indicating that the risk model could better predict LUAD prognosis. At the same time, the histogram (1, 3 and 5 years) was corrected to visualize its performance. Compared with the ideal model, the calibration curves showed a strong performance (Figure 10B). The clinical prediction model was evaluated by DCA, and the results showed that the nomogram had the optimal prediction effect than the RiskScore (Figure 10C).

Immune Status Analysis of High- and Low-Risk Population

We compared the expression differences of the immune checkpoint genes in high-risk group and low-risk group in the TCGA dataset. It was found that the expressions of CD27, CTLA4 and TIGIT in the low-risk group were higher than the high-risk group (Figure 11A). In addition, immune scores on the TCGA dataset samples showed that the immune scores of the low-risk group were higher than the high-risk group (Figure 11B-F).

Predictive Analysis of Immunotherapy Response

An immunotherapy dataset (Imvigor210) recording expression data of human mUC samples

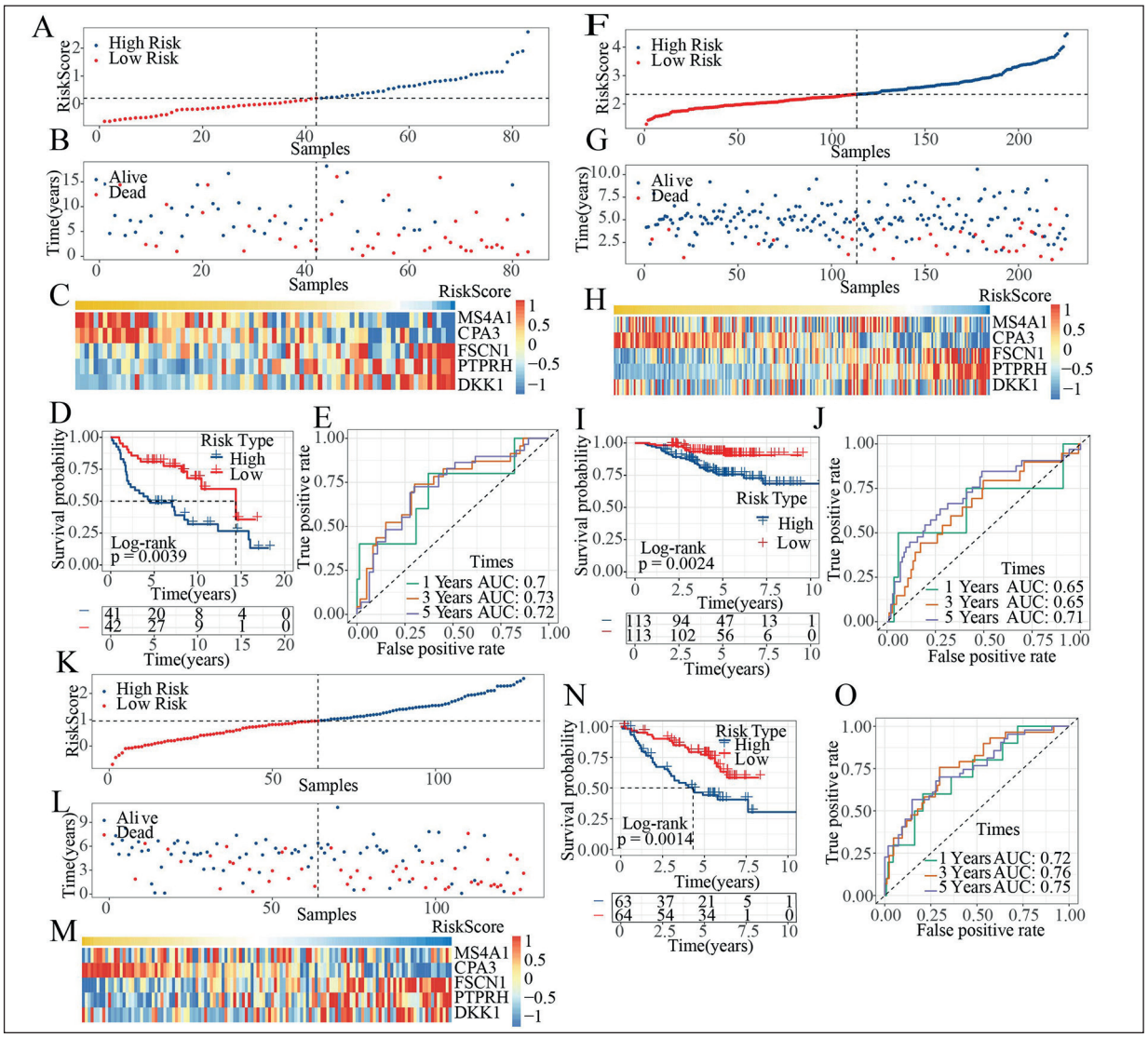

Figure 7. Validation of the risk model in GSEA independent datasets. **A-B**, The distribution of RiskScore and the corresponding distribution of survival states in the independent dataset GSE30219. **C**, Heat map of RiskScore model gene expression. **D-E**, KM curve and ROC curve of independent dataset GSE30219. **F-G**, The distribution of RiskScore and the corresponding distribution of survival states in the independent dataset GSE31210. **H**, Heat map of RiskScore model gene expression. **I-J**, KM curve and ROC curve of independent dataset GSE31210. **K-L**, The distribution of RiskScore and the corresponding distribution of survival states in the independent dataset GSE50081. (**M**) Heat map of RiskScore model gene expression. **N-O**, KM curve and ROC curve of independent dataset GSE50081.

from patients responsive or unresponsive to antiPD-L1 immunotherapy was used to explore whether the 5-gene model could predict patients' benefit from immunotherapy. The Kaplan-Meier curve showed that higher RiskScore was associated with worse survival in mUC patients receiving immunotherapy (Figure 12A). The ROC curve demonstrated that RiskScore had a higher AUC (0.71) value compared to NEO (AUC=0.67) and TMB (AUC=0.64) features (Figure 12B). The mUC patients receiving immunotherapy were divided into high- and low-groups according to their RiskScore. The number of patients with disease progression was more in the high-risk group but fewer in stable patients in the low-risk group (Figure 12C). We used MpCounter to calculate the immune cell scores of Imvigor210 samples and the correlation among Riskscore and TMB, NEO and immune cell scores. The results showed that Riskscore was negatively correlated with CD8 T cells, T cells, and NK cells (Figure 12D). At the same time, comparison of the difference

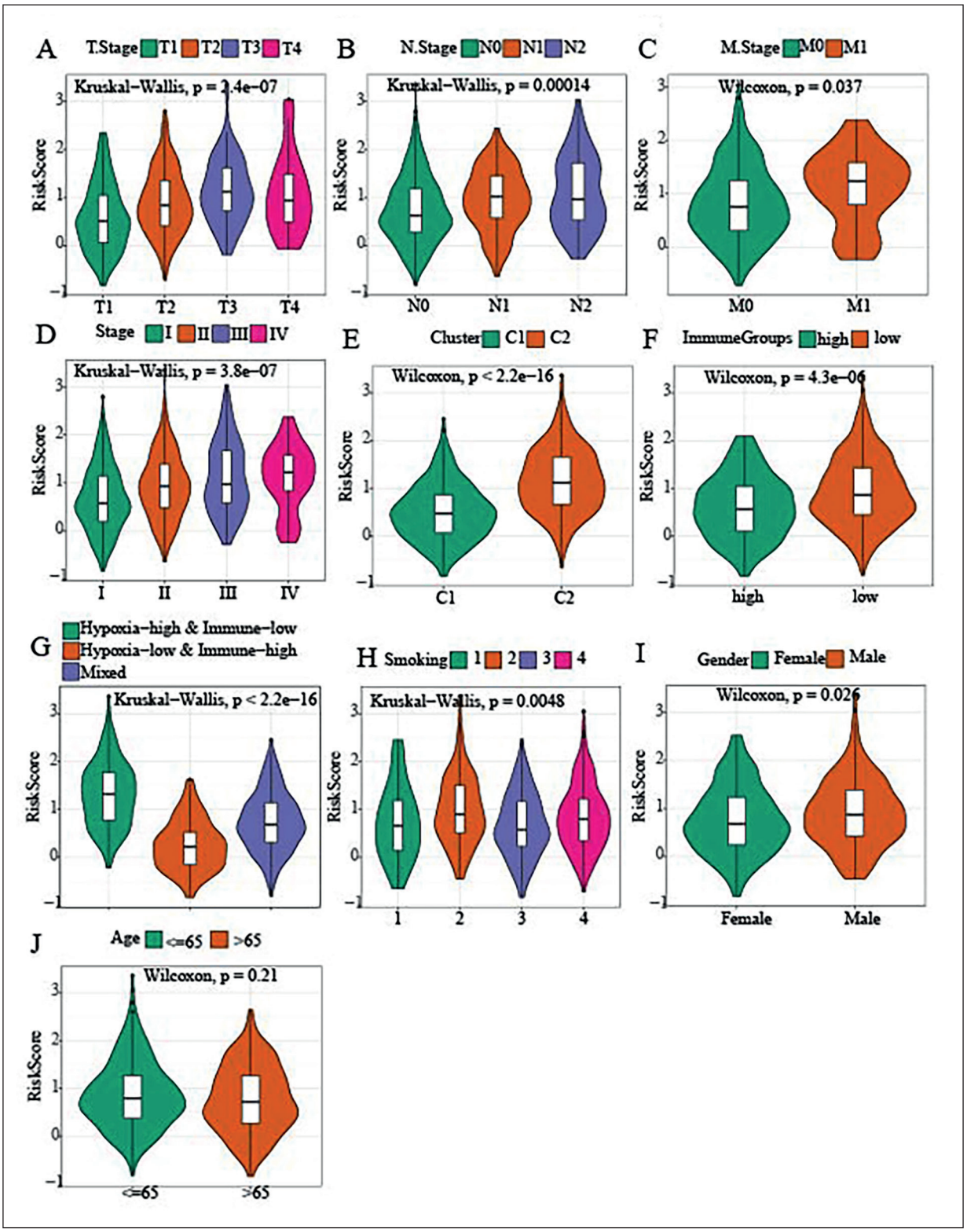

Figure 8. The distribution of risk scores on different clinical features in the TCGA dataset. **A**, Distribution of T. Stage risk scores. **B**, Distribution of N. Stage risk scores. **C**, Distribution of M. Stage risk scores. **D**, Distribution of I-IV. Stage risk scores. **E**, Distribution of Cluster risk scores. **F**, Distribution of ImmuneGroups risk scores. **G**, Distribution of Immunohypoxia grouping risk scores. **H**, Distribution of smoking risk scores. **I**, Distribution of Gender risk scores. **J**, Distribution of age risk scores.

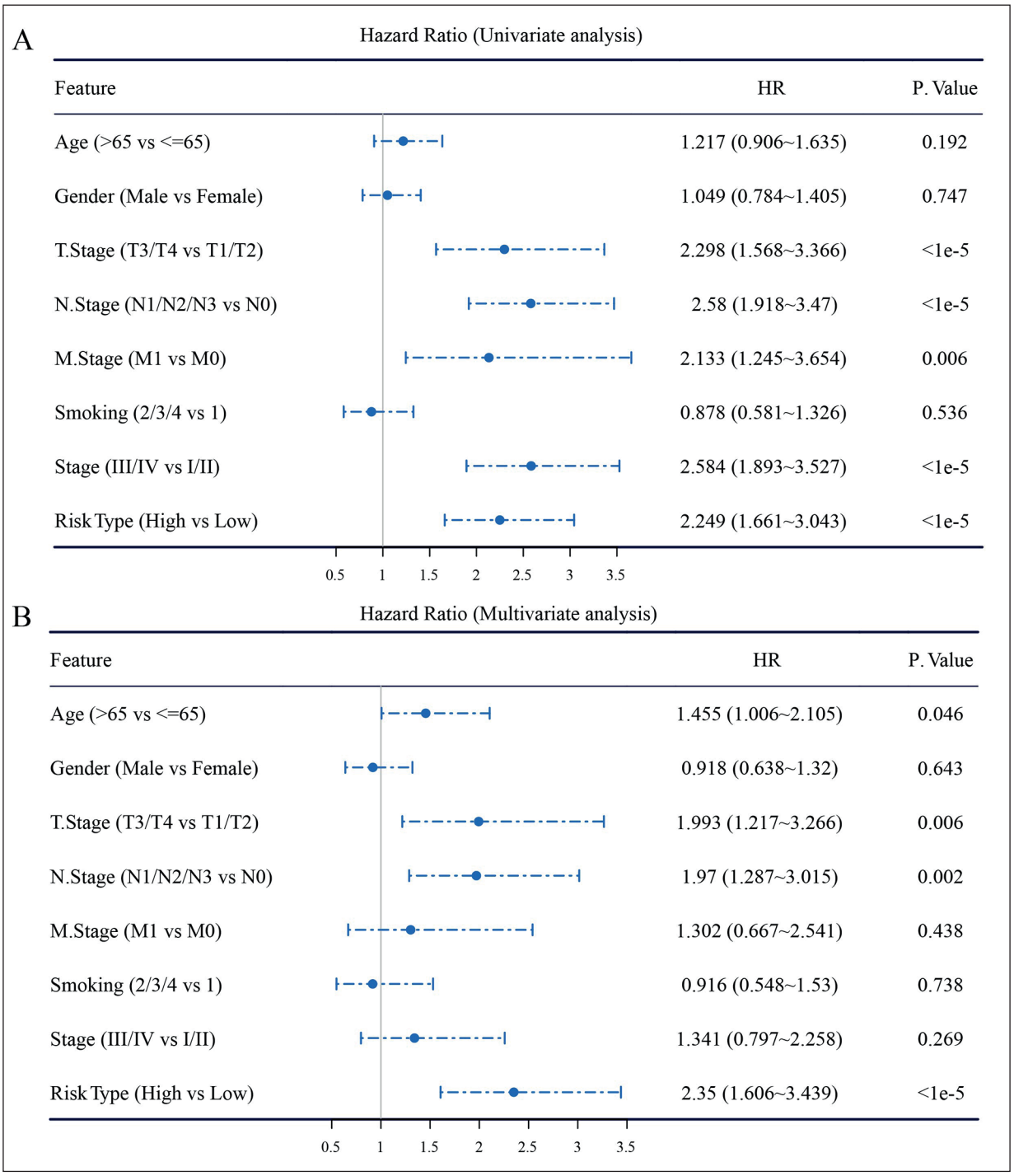

Figure 9. Clinical characteristics and univariate and multivariate RiskScore analyses. **A**, Univariate analysis of clinical characteristics of TCGA dataset. **B**, Multivariate analysis of clinical features of the TCGA dataset.

of RiskScore among four responses (CR, PD, PR and SD) showed significant differences in the effectiveness of RiskScore and immunotherapy that samples with worse immunotherapy outcomes were associated with higher RiskScore (Figure 12E). There was no significant difference among RiskScore and immune cell grouping, tumor cells and immunophenotype (Figure 12F-H).

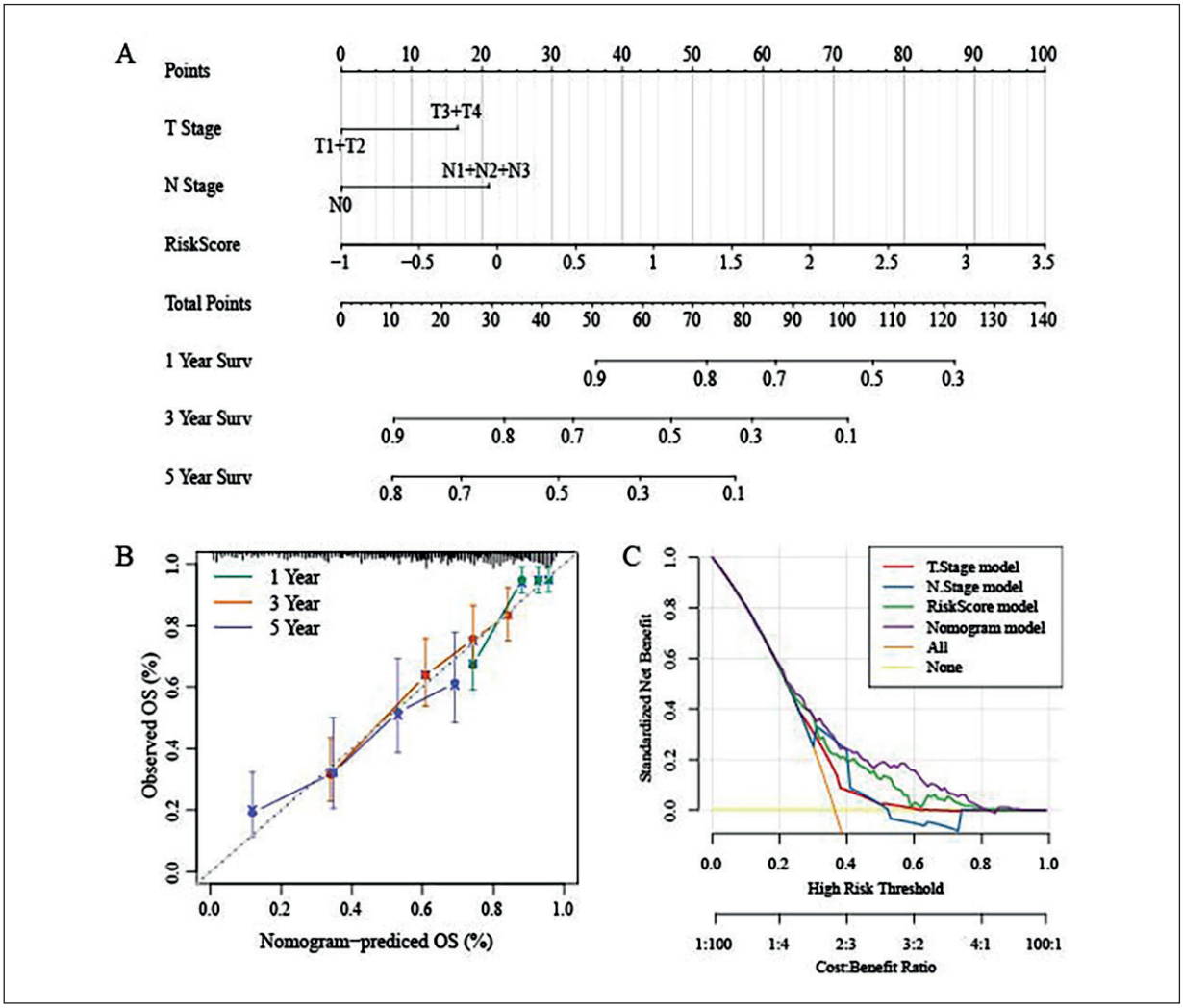

Figure 10. Riskscore and clinical characteristics were used to construct the histogram and forest map. **A**, Riskscore and clinical characteristics were used to construct the histogram. **B**, Riskscore and clinical characteristics were corrected to construct a histogram. **C**, Clinical characteristics and Riskscore Decision Curve Analysis.

Comparison of Risk Models With Other Models

By reviewing previous papers, we selected four published prognostic risk models [30-33] for comparison with our 5-gene model. To make the model comparable, we calculated the RiskScore of each LUAD sample in TCGA dataset using the same method according to the corresponding genes in the four models, and then compared the ROC curve of the RiskScore of these models with our model. Compared with the four published prognostic risk models, our risk model had higher AUC values at 1-, 3- and 5 years, which were 0.74, 0.7 and 0.68, respectively (**Supplementary Figure 4**). The data indicated that our model outperformed in predicting prognosis.

Discussion

At present, pathological staging is still a commonly used method to evaluate the prognosis of LUAD, though it is not often accurate[34]. In recent years, rapid development of high-throughput technology enables a large amount of bioinformatics data to play an important role in LUAD diagnosis, treatment efficacy evaluation and prognosis prediction[35]. At present, with the development and application of genomics, functional genomics, transcriptomics and proteomics, multiple genetic markers have been explored and studied in the prognostic diagnosis of LUAD[36,37]. Increasing evidence showed that some genes have critical

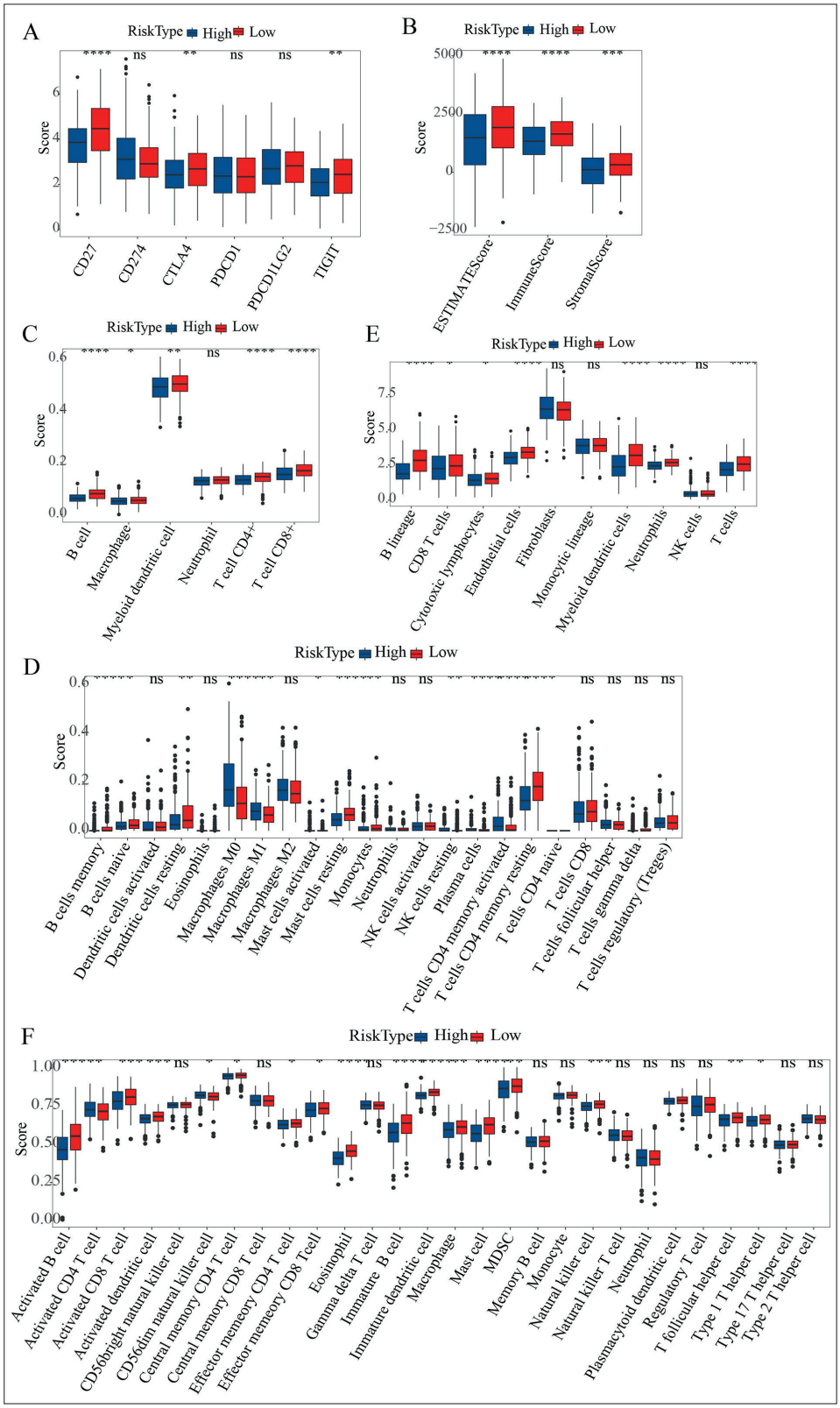

Figure 11. Comparative analysis of immune checkpoints and immune infiltrate scores for high- and low-risk groups. **A**, The differences in the expression of immune checkpoint genes (CD27, CTLA4, PDCD1, PDCDILG2 and TIGIT) found in the existing studies in the TCGA dataset were compared by assessment software. **B-F**, TCGA data set was scored by ESTIMATE, Timer, MCPCounter, Cibersort, ssGSEA invasion scoring software.

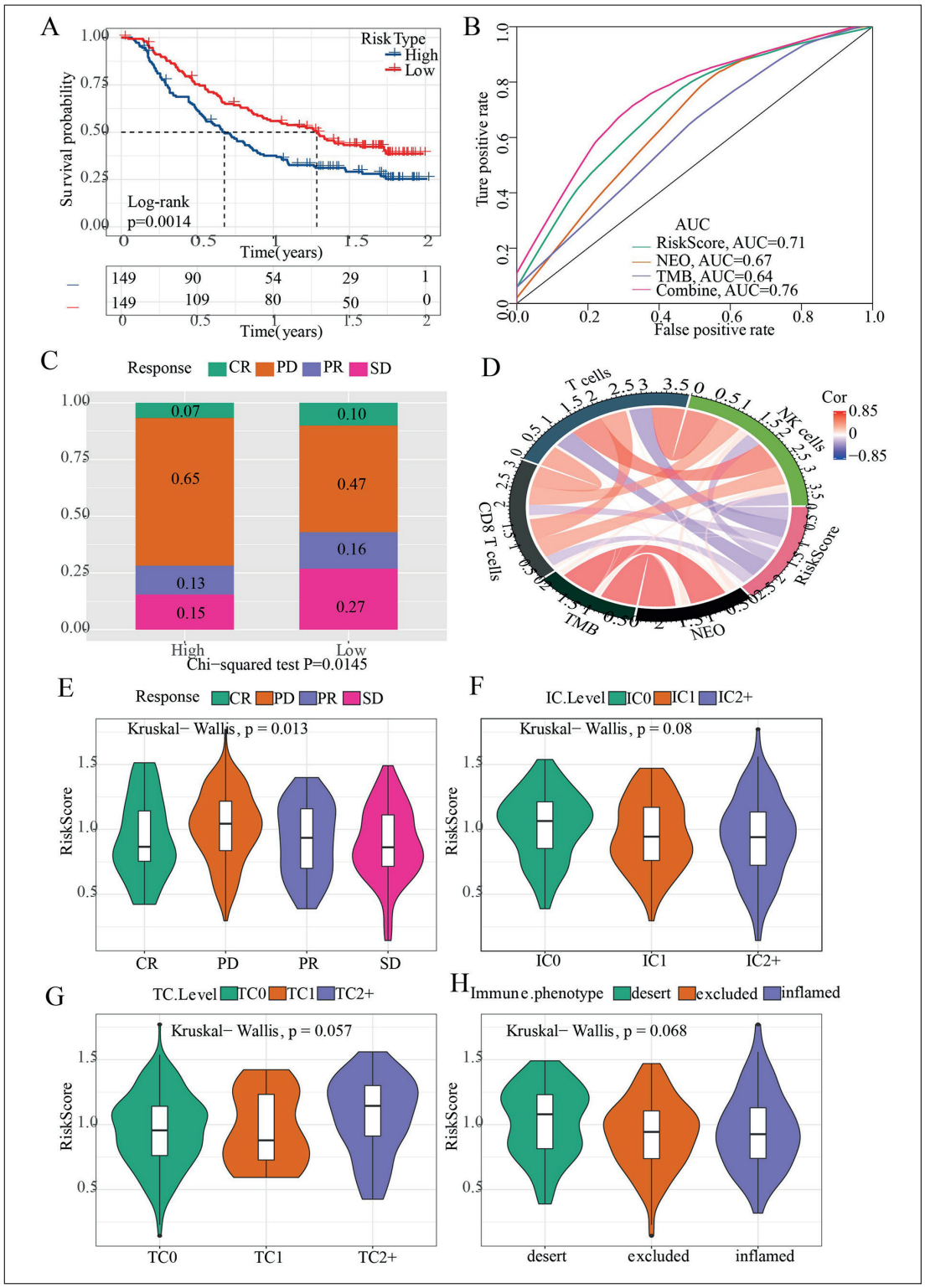

Figure 12. Risk model prediction for immunotherapy. **A**, The Kaplan-Meier curve shows the Imvigor210 dataset. **B**, The Imvigor210 dataset displayed by the area under the ROC curve AUC. **C**, Imvigor210 dataset between different subsets corresponding to immunotherapy stack. **D**, The correlation between Riskscore and immunization score, TMB and NEO in Imvigor210 dataset. **E**, Riskscore differences between groups of immunotherapy effectiveness. **F**, Riskscore differences between immune cell groups. **G**, Riskscore differences between tumor cell groups. **H**, Riskscore differences between immunophenotypic groups. (IC: immune cell; TC: tumour cell; IP: Immune phenotype).

an important regulatory functions in the occurrence and development of tumors, and could be reliable in the prognostic prediction of cancer patients[38,39]. For example, Li et al[37] found a total of 8 RNA binding proteins (IGF2BP1, IFIT1B, PABPC1, TLR8, GAPDH, PIWIL4, RNPC3, ZC3H12C) in the RNA binding proteins that are predictive of the prognosis of LUAD, and have been identified as pivotal genes related to the prognosis. Moreover, these 8 RNA binding proteins were used to construct a prognostic model. A total of 4 genes (FUT4, SLC25A42, IGFBP1 and KLHDC8B) with expression differences in LUAD were obtained through TCGA database combined with GEO dataset, and these 4 genes can predict the prognostic characteristics of GEO dataset LUAD[40].

In addition, the biological characteristics of cells can intuitively reflect whether the cell life process is proceeding normally[41]. Studies[42,43] have found that hypoxia-related genes may help overcome hypoxia-related drug resistance in tumor treatment. Hypoxia is a characteristic of tumor microenvironment, and it is related to the malignant degree of cancer[42]. Current research reports show that hypoxia can regulate the activity of important carcinogenic pathways through changing gene expression and histone modification status[44], such as the overexpression of related genes (HIF-1 alpha, CA IX, VEGF) in nasopharyngeal carcinoma, and this is common and may be related to the up-regulation of HIF-dependent pathway expression caused by hypoxia and a poor prognosis[45]. The hypoxia-related gene YTHDF 1 plays a key role in hypoxia adaptation and pathogenesis of non-small cell lung cancer[46]. As a response to tumor hypoxia, hypoxia-inducible factor-1 is activated and its genetic polymorphisms are found to be closely related to the risk of lung cancer[47]. In this study, we typed 55 LUDA samples of TCGA based on the prognostic-related genes of hypoxia. These samples can be divided into two subtypes, and we found that the C2 subtype was more related to hypoxia and could predict LUAD prognosis.

LUAD as the most common type of non-small cell lung cancer[3] is characterized by dense lymphocyte infiltration, and this indicates that immune system plays an active role in the occurrence and growth of LUAD[48]. Luo et al[49] developed a robust model for predicting the survival outcome of LUAD patients from the perspective of tumor immunology, and proposed an immune prognostic model for LUAD, which can independently identify high-risk patients with poor survival. This study suggested a close relationship between local immune status and prognosis. At present, immune checkpoint blockade (Icb) has achieved significant clinical effects on the treatment of non-small cell lung cancer[50]. In the research of immune-related gene pairs, some studies have been conducted based on the practical and robustness of immune gene data sets with cross-platform compatibility. A four-gene prediction model demonstrated potential predictive value in improving the TNM staging of patients with LUAD[51]. Therefore, in this study, we further selected genes related to the immune microenvironment, and through the immune infiltration score analysis of the TCGA dataset, we divided them into two groups of high immunity and low immunity according to the optimal segmentation point of ImmuneScore and found highly immune groups showed a favorable prognosis.

At present, there is no research reporting on the genes related to hypoxia and immunity in the prognosis of LUAD. We found that the hypoxia-high and immune-low group had the worst prognosis, and the hypoxia-low & immune-high group had the most favorable prognosis. Finally, 390 differential genes related to the hypoxic immune microenvironment were identified from differential genes between the groups. Based on these DEGs, a 5-gene signature model was finally developed. At the same time, we compared the current model with previously established ones[30-33], and finally confirmed that 5-gene signature had a strong robustness and could stably predict LUAD samples in the datasets of different platforms (TCGA, GSE30219, GSE31210, GSE50081 and the immunotherapy data set IMvigor210).

Conclusions

This work developed a 5-gene signature prognostic hierarchical system with a high AUC in both the training set and the independent validation set. Based on hypoxia and immune microenvironment, we were the first to report a 5-gene risk signature that could help to predict overall survival of LUAD patients. The current finding provided more gene targets and will help future researchers to develop new treatments for LUAD patients.

Conflict of Interest
The Authors declare that they have no conflict of interests.

Availability of Data and Materials
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contribution
GuoYong Lin conceived and designed the research; JingJing Yan and XiangQiong Guo acquired the data; Shun Wu and LiHua Wu analyzed the data; ZhiSen Gao did the statistical analysis; GuoYong Lin wrote the manuscript; ZhiYong Wang identified the research and editorial manuscript. All authors red and approved the manuscript.

ORCID ID
ZhiYong Wang: 0000-0003-1728-9106.

References

1) Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc 2008; 83: 584-594.
2) Liu J, Yang XY, Shi WJ. Identifying differentially expressed genes and pathways in two types of non-small cell lung cancer: adenocarcinoma and squamous cell carcinoma. Genet Mol Res 2014; 13: 95-102.
3) Balzer BWR, Loo C, Lewis CR, Trahair TN, Anazodo AC. Adenocarcinoma of the Lung in Childhood and Adolescence: A Systematic Review. J Thorac Oncol 2018; 13: 1832-1841.
4) Xu Y, Zheng M, Wang N, Wang R. Comprehensive Study of Surgical Treated Lung Adenocarcinoma with Ground Glass Nodule Component. Med Sci Monit 2019; 25: 8492-8498.
5) Arakawa S, Yoshida T, Nakayama Y, Motoi N, Ohe Y. Small Cell Cancer Transformation of Lung Adenocarcinoma During Durvalumab Treatment After Chemoradiotherapy. J Thorac Oncol 2020; 15: e145-e146.
6) Chen Y, Zhang Y, Song W, Zhang Y, Dong X, Tan M. Ginsenoside Rh2 Inhibits Migration of Lung Cancer Cells under Hypoxia via mir-491. Anticancer Agents Med Chem 2019; 19: 1633-1641.
7) Afshin A, Babalola D, McLean M, Yu Z, Ma W, Chen CY, Arabi M, Mozaffarian D. Information Technology and Lifestyle: A Systematic Evaluation of Internet and Mobile Interventions for Improving Diet, Physical Activity, Obesity, Tobacco, and Alcohol Use. J Am Heart Assoc 2016; 5: e003058.
8) Chowdhury B, Garai G. A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics 2017; 109: 419-431.
9) Khurshid Z, Zafar MS, Khan RS, Najeeb S, Slowey PD, Rehman IU. Role of Salivary Biomarkers in Oral Cancer Detection. Adv Clin Chem 2018; 86: 23-70.
10) Malas TB, Leonhard WN, Bange H, Granchi Z, Hettne KM, Van Westen GJP, Price LS, t Hoen PAC, Peters DJM. Prioritization of novel ADPKD drug candidates from disease-stage specific gene expression profiles. EBioMedicine 2020; 51: 102585.
11) Wen DG, Zhao XP, You Y, Liu ZJ. Systematic analysis of immune-related genes based on a combination of multiple databases to build a diagnostic and a prognostic risk model for hepatocellular carcinoma. Cancer Immunol Immunother 2021; 70: 773-786.
12) Michaelsen SR, Christensen IJ, Grunnet K, Stockhausen MT, Broholm H, Kosteljanetz M, Poulsen HS. Clinical variables serve as prognostic factors in a model for survival from glioblastoma multiforme: an observational study of a cohort of consecutive non-selected patients from a single institution. BMC Cancer 2013; 13: 402.
13) Lee JH, Jung S, Park WS, Choe EK, Kim E, Shin R, Heo SC, Lee JH, Kim K, Chai YJ. Prognostic nomogram of hypoxia-related genes predicting overall survival of colorectal cancer-Analysis of TCGA database. Sci Rep 2019; 9: 1803.
14) Yang L, Roberts D, Takhar M, Erho N, Bibby BAS, Thiruthaneeswaran N, Bhandari V, Cheng WC, Haider S, McCorry AMB, McArt D, Jain S, Alshalalfa M, Ross A, Schaffer E, Den RB, Jeffrey Karnes R, Klein E, Hoskin PJ, Freedland SJ, Lamb AD, Neal DE, Buffa FM, Bristow RG, Boutros PC, Davicioni E, Choudhury A, West CML. Development and Validation of a 28-gene Hypoxia-related Prognostic Signature for Localized Prostate Cancer. EBioMedicine 2018; 31: 182-189.
15) Chen Q, Hu L, Chen K. Construction of a Nomogram Based on a Hypoxia-Related lncRNA Signature to Improve the Prediction of Gastric Cancer Prognosis. Front Genet 2020; 11: 570325.
16) Huang R, Mao M, Lu Y, Yu Q, Liao L. A novel immune-related genes prognosis biomarker for melanoma: associated with tumor microenvironment. Aging (Albany NY) 2020; 12: 6966-6980.
17) Ge P, Wang W, Li L, Zhang G, Gao Z, Tang Z, Dang X, Wu Y. Profiles of immune cell infiltration and immune-related genes in the tumor microenvironment of colorectal cancer. Biomed Pharmacother 2019; 118: 109228.
18) Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R, Furuta K, Tsuta K, Shibata T, Yamamoto S, Watanabe S, Sakamoto H, Kumamoto K, Takenoshita S, Gotoh N, Mizuno H, Sarai A, Kawano S, Yamaguchi R, Miyano S, Yokota J. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res 2012; 72: 100-111.
19) Rousseaux S, Debernardi A, Jacquiau B, Vitte AL, Vesin A, Nagy-Mignotte H, Moro-Sibilot D,

Brichon PY, Lantuejoul S, Hainaut P, Laffaire J, de Reyniès A, Beer DG, Timsit JF, Brambilla C, Brambilla E, Khochbin S. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Sci Transl Med 2013; 5: 186ra166.

20) Der SD, Sykes J, Pintilie M, Zhu CQ, Strumpf D, Liu N, Jurisica I, Shepherd FA, Tsao MS. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. J Thorac Oncol 2014; 9: 59-64.

21) Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004; 20: 307-315.

22) Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, Kadel EE, III, Koeppen H, Astarita JL, Cubas R, Jhunjhunwala S, Banchereau R, Yang Y, Guan Y, Chalouni C, Ziai J, Şenbabaoğlu Y, Santoro S, Sheinson D, Hung J, Giltnane JM, Pierce AA, Mesh K, Lianoglou S, Riegler J, Carano RAD, Eriksson P, Höglund M, Somarriba L, Halligan DL, van der Heijden MS, Loriot Y, Rosenberg JE, Fong L, Mellman I, Chen DS, Green M, Derleth C, Fine GD, Hegde PS, Bourgon R, Powles T. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature 2018; 554: 544-548.

23) Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005; 102: 15545-15550.

24) Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautès-Fridman C, Fridman WH, de Reyniès A. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol 2016; 17: 218.

25) Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 2012; 16: 284-287.

26) Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, Li B, Liu XS. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer Res 2017; 77: e108-e110.

27) Li W, Wang H, Ma Z, Zhang J, Ou-Yang W, Qi Y, Liu J. Multi-omics Analysis of Microenvironment Characteristics and Immune Escape Mechanisms of Hepatocellular Carcinoma. Front Oncol 2019; 9: 1019.

28) Zheng M, Hu Y, Gou R, Liu O, Nie X, Li X, Liu Q, Hao Y, Liu J, Lin B. Identification of immune-enhanced molecular subtype associated with BRCA1 mutations, immune checkpoints and clinical outcome in ovarian carcinoma. J Cell Mol Med 2020; 24: 2819-2831.

29) Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010; 33: 1-22.

30) 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: 3383-3392.

31) Sun S, Guo W, Wang Z, Wang X, Zhang G, Zhang H, Li R, Gao Y, Qiu B, Tan F, Gao Y, Xue Q, Gao S, He J. Development and validation of an immune-related prognostic signature in lung adenocarcinoma. Cancer Med 2020; 9: 5960-5975.

32) Xue L, Bi G, Zhan C, Zhang Y, Yuan Y, Fan H. Development and Validation of a 12-Gene Immune Relevant Prognostic Signature for Lung Adenocarcinoma Through Machine Learning Strategies. Front Oncol 2020; 10: 835.

33) Yu X, Zhang X, Zhang Y. Identification of a 5-Gene Metabolic Signature for Predicting Prognosis Based on an Integrated Analysis of Tumor Microenvironment in Lung Adenocarcinoma. J Oncol 2020; 2020: 5310793.

34) Travis WD, Asamura H, Bankier AA, Beasley MB, Detterbeck F, Flieder DB, Goo JM, MacMahon H, Naidich D, Nicholson AG, Powell CA, Prokop M, Rami-Porta R, Rusch V, van Schil P, Yatabe Y. The IASLC Lung Cancer Staging Project: Proposals for Coding T Categories for Subsolid Nodules and Assessment of Tumor Size in Part-Solid Tumors in the Forthcoming Eighth Edition of the TNM Classification of Lung Cancer. J Thorac Oncol 2016; 11: 1204-1223.

35) Chen H, Carrot-Zhang J, Zhao Y, Hu H, Freeman SS, Yu S, Ha G, Taylor AM, Berger AC, Westlake L, Zheng Y, Zhang J, Ramachandran A, Zheng Q, Pan Y, Zheng D, Zheng S, Cheng C, Kuang M, Zhou X, Zhang Y, Li H, Ye T, Ma Y, Gao Z, Tao X, Han H, Shang J, Yu Y, Bao D, Huang Y, Li X, Zhang Y, Xiang J, Sun Y, Li Y, Cherniack AD, Campbell JD, Shi L, Meyerson M. Genomic and immune profiling of pre-invasive lung adenocarcinoma. Nat Commun 2019; 10: 5472.

36) Qu Y, Cheng B, Shao N, Jia Y, Song Q, Tan B, Wang J. Prognostic value of immune-related genes in the tumor microenvironment of lung adenocarcinoma and lung squamous cell carcinoma. Aging (Albany NY) 2020; 12: 4757-4777.

37) Wu Q, Wang L, Wei H, Li B, Yang J, Wang Z, Xu J, Zhou YL, Zhang B. Integration of multiple key molecules in lung adenocarcinoma identifies prognostic and immunotherapeutic relevant gene signatures. Int Immunopharmacol 2020; 83: 106477.

38) Liu Y, Yang Y, Luo Y, Wang J, Lu X, Yang Z, Yang J. Prognostic potential of PRPF3 in hepatocellular carcinoma. Aging (Albany NY) 2020; 12: 912-930.

39) Xu F, Zhang Z, Yuan M, Zhao Y, Zhou Y, Pei H, Bai L. M6A Regulatory Genes Play an Important Role in the Prognosis, Progression and Immune Microenvironment of Pancreatic Adenocarcinoma. Cancer Invest 2021; 39: 39-54.

40) Zhao K, Li Z, Tian H. Twenty-gene-based prognostic model predicts lung adenocarcinoma survival. Onco Targets Ther 2018; 11: 3415-3424.

41) Chung S, Kim SH, Seo Y, Kim SK, Lee JY. Quantitative analysis of cell proliferation by a dye di-

lution assay: Application to cell lines and cocultures. Cytometry A 2017; 91: 704-712.

42) Jing X, Yang F, Shao C, Wei K, Xie M, Shen H, Shu Y. Role of hypoxia in cancer therapy by regulating the tumor microenvironment. Mol Cancer 2019; 18: 157.

43) Zhang B, Tang B, Gao J, Li J, Kong L, Qin L. A hypoxia-related signature for clinically predicting diagnosis, prognosis and immune microenvironment of hepatocellular carcinoma patients. J Transl Med 2020; 18: 342.

44) Li H, Peng C, Zhu C, Nie S, Qian X, Shi Z, Shi M, Liang Y, Ding X, Zhang S, Zhang B, Li X, Xu G, Lv Y, Wang L, Friess H, Kong B, Zou X, Shen S. Hypoxia promotes the metastasis of pancreatic cancer through regulating NOX4/KDM5A-mediated histone methylation modification changes in a HIF1A-independent manner. Clin Epigenetics 2021; 13: 18.

45) Hui EP, Chan AT, Pezzella F, Turley H, To KF, Poon TC, Zee B, Mo F, Teo PM, Huang DP, Gatter KC, Johnson PJ, Harris AL. Coexpression of hypoxia-inducible factors 1alpha and 2alpha, carbonic anhydrase IX, and vascular endothelial growth factor in nasopharyngeal carcinoma and relationship to survival. Clin Cancer Res 2002; 8: 2595-2604.

46) Shi Y, Fan S, Wu M, Zuo Z, Li X, Jiang L, Shen Q, Xu P, Zeng L, Zhou Y, Huang Y, Yang Z, Zhou J, Gao J, Zhou H, Xu S, Ji H, Shi P, Wu DD, Yang C, Chen Y. YTHDF1 links hypoxia adaptation and non-small cell lung cancer progression. Nat Commun 2019; 10: 4892.

47) Xu S, Ying K. Association between HIF-1α gene polymorphisms and lung cancer: A meta-analysis. Medicine (Baltimore) 2020; 99: e20610.

48) Tsuta K, Ishii G, Kim E, Shiono S, Nishiwaki Y, Endoh Y, Kodama T, Nagai K, Nagai K. Primary lung adenocarcinoma with massive lymphocyte infiltration. Am J Clin Pathol 2005; 123: 547-552.

49) Luo C, Lei M, Zhang Y, Zhang Q, Li L, Lian J, Liu S, Wang L, Pi G, Zhang Y. Systematic construction and validation of an immune prognostic model for lung adenocarcinoma. J Cell Mol Med 2020; 24: 1233-1244.

50) Chen H, Chong W, Teng C, Yao Y, Wang X, Li X. The immune response-related mutational signatures and driver genes in non-small-cell lung cancer. Cancer Sci 2019; 110: 2348-2356.

51) Shi X, Li R, Dong X, Chen AM, Liu X, Lu D, Feng S, Wang H, Cai K. IRGS: an immune-related gene classifier for lung adenocarcinoma prognosis. J Transl Med 2020; 18: 55.