Systematic Profiling of Invasion-related Genes Signature Predicts Prognostic Features and Identificates Molecular Subtypes of Lung Adenocarcinoma

Ping Yu
The First Affiliated Hospital of China Medical University
Linlin Tong
The First Affiliated Hospital of China Medical University
Yujia Song
The First Affiliated Hospital of China Medical University
Hui Qu
The First Affiliated Hospital of China Medical University
Ying Chen (dongyechenying@126.com)
The First Affiliated Hospital of China Medical University https://orcid.org/0000-0002-0022-4034

Research article

Keywords: invasion genes, molecular subtype, LUAD, TCGA, multi-gene signature

Posted Date: March 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-351860/v1

Abstract

Background: Due to the high heterogeneity of lung adenocarcinoma (LUAD), molecular subtype based on gene expression profiles is of great significance for diagnosis, and prognosis prediction in patients with LUAD.

Methods: Invasion-related genes were obtained from the CancerSEA database, and LUAD expression profiles were downloaded from The Cancer Genome Atlas. The *ConsensusClusterPlus* was used to obtain molecular subtypes based on invasion-related genes. The *limma* software package was used to identify differentially expressed genes (DEGs). A multi-gene risk model was constructed by Lasso-Cox analysis. A nomogram was also constructed based on risk scores and meaningful clinical features.

Results: 3 subtypes (C1, C2, C3) based on the expression of invasion-related genes were obtained. C3 had the worst prognosis. A total of 669 DEGs were identified among the subtypes. Pathway enrichment analysis results showed that the DEGs were mainly enriched in the cell cycle, DNA replication, the p53 signaling pathway, and other tumor-related pathways. A 5-gene signature (*KRT6A*, *MELTF*, *IRX5*, *MS4A1*, *CRTAC1*) was identified by using Lasso-Cox analysis. The training, validation, and external independent cohorts proved that the model was robust and had better prediction ability than other lung cancer models. The gene expression results showed that the expression levels of *MS4A1* and *KRT6A* in tumor tissues were higher than in normal tissues, while *CRTAC1* expression in tumor tissues was lower than in normal tissues. At the same time, the 5 genes were significantly expressed in pan-cancer immune subtypes. Gene set enrichment analysis showed that MS4A1, KRT6A, and CRAT1 genes were both enriched in the HALLMARK_IL2_STAT5_SIGNALING pathway, and IRX5 and MELTF gene were both enriched in the HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION pathway.

Conclusion: The 5-gene signature prognostic stratification system based on invasion-related genes could be used to assess prognostic risk in patients with LUAD.

Introduction

Lung cancer is the most common cause of cancer-related death worldwide(Ferlay et al.,, 2015). About 85% of lung cancers are non-small cell lung carcinomas(Zheng, 2016), which can be divided into 3 histological subtypes, with lung adenocarcinoma (LUAD) being the most common(Travis, 2011). Major risk factors for LUAD include smoking, genetic factors, diet, alcohol consumption, and exposure to ionizing radiation and environmental pollutants(Malhotra et al.,, 2016, Pesch et al.,, 2012). As the early symptoms of LUAD are not obvious, most patients with LUAD are diagnosed with advanced stages, though metastasis occurs earlier, and the 5-year overall survival (OS) rate is less than 20%(Lin et al.,, 2016). Although progress has been made in terms of early diagnostic methods, chemotherapy, radiotherapy, and surgical diagnosis and treatment options in recent years, the prognoses of patients with LUAD remain poor(Qi et al.,, 2016).

At present, lung cancer treatment mainly depends on histological type and clinical stage, but due to the high heterogeneity of LUAD, even patients with the same histological type and clinical stage of LUAD have different prognoses, so the classification of LUAD based on high-throughput sequencing data is of great significance for individualized and accurate LUAD treatment.

In recent years, the use of high-throughput sequencing technology to detect a large number of gene expression changes, combined with the use of bioinformatics methods to systematically analyze tumor-related genes and their regulatory mechanisms, has become an important research means in functional genomics, and it has been widely used to screen potential tumor biomarkers(Zheng et al.,, 2019, Xing and Zeng, 2016, Cancer Genome Atlas, 2012). In their research of LUAD, Krzystanek et al.(Krzystanek et al.,, 2016) identified a 7-gene signature (*ADAM10*, *DLGAP5*, *RAD51AP1*, *FGFR10P*, *NCGAP*, *KIF15*, *ASPM*) by analyzing microarray data of early LUAD from the Gene Expression Omnibus (GEO) database, and they found significant differences in survival and prognosis among these genes. Li et al.(Li et al.,, 2020a) constructed a 5-gene signature, which was closely related to the tumor microenvironment, by using the GSE103584 dataset. The 13-gene signature constructed by He et al.(He et al.,, 2020) with metabolism-related genes was helpful to predict the prognoses of patients with LUAD. Han et al.(Han et al.,, 2020) constructed a multi-gene signature based on tumor-infiltrating B lymphocyte-specific genes to predict the clinical outcomes of radiotherapy and immunotherapy in patients with LUAD. Li et al.(Li et al.,, 2020b) used a 6-gene signature to predict the prognoses of patients with LUAD.

In this study, we identified molecular subtypes of LUAD based on tumor invasion-related genes by using gene expression data from public databases, such as The Cancer Genome Atlas (TCGA) and GEO, for the first time. We evaluated the relationships between the molecular subtypes and prognosis and clinical features. The prognostic risk model based on differentially expressed genes (DEGs) between the LUAD subtypes could be used to evaluate LUAD prognosis. In addition, the nomogram we constructed could be used to help clinical decision-making and prognosis judgment.

Methods And Materials

Data download and preprocessing

RNA sequencing data and clinical follow-up information for LUAD were downloaded from the TCGA database. The GSE31210 chip dataset containing survival time information was downloaded from the GEO database.

Invasion-related genes were obtained from the *CancerSEA* website(Yuan et al.,, 2019), which contains 97 genes (**Supplementary Table 1**).

The TCGA-LUAD RNA sequencing data were preprocessed as follows: 1) the samples with no clinical follow-up information were removed; 2) the samples with no survival time information were removed; 3) the samples with no status information were removed; 4) the Ensemble IDs was transformed into gene symbols; and 5) the median expression of multiple gene symbols was obtained.

The GEO data were preprocessed as follows: 1) the samples with no clinical follow-up information were removed; 2) the samples with no survival time or status information were removed; 3) the probes were converted into gene symbols; 4) the probes were mapped to multiple genes, and the probes were deleted; and 5) the median expression of multiple gene symbols was obtained.

After preprocessing, there were 500 TCGA-LUAD samples and 126 GSE31210 dataset samples. The clinical statistics of the samples can be found in **Supplementary Table 2**.

Consistent clustering

The expression levels of 97 invasion-related genes were extracted from the TCGA expression profiles, and genes related to LUAD prognosis were obtained by univariate Cox analysis using the coxph function in R ($p < 0.01$). ConsensusClusterPlus (V1.48.0) was used to cluster the samples consistently according to significant genes from the univariate Cox analysis (parameters: reps = 100, pItem = 0.8, pFeature = 1, distance = Minkowski). Pam and Minkowski distances were used as the clustering algorithm and distance measure, respectively.

Identification of differentially expressed genes

The DEGs of different molecular subtypes were calculated by using the limma package in R(Ritchie et al.,, 2015). The DEGs were filtered according to the threshold of FDR < 0.01 and | log2fc | > 1, and then volcano maps were plotted.

GO and KEGG enrichment analyses

The results of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment analyses using the DAVID database showed that $p < 0.05$ was statistically significant. The results were visualized by using the ggplot2 package in R (V3.5.3).

5. Calculation of immune scores

Stromal, immune, and estimate scores were calculated by using the ESTIMATE package in R. Ten immune cells were evaluated by using MCPcounter, and 28 immune cells were evaluated by using single-sample gene set enrichment analysis (ssGSEA) with the GSVA package in R(Charoentong et al.,, 2017).

Construction of prognostic risk model

The 500 samples in the TCGA dataset were divided into training and validation cohorts. To avoid the influence of random assignment bias on the stability of subsequent modeling, all samples were put back into the random grouping 100 times in advance, and the group sampling was carried out according to the training cohort-to-validation cohort ratio of 1:1. The final training cohort contained 250 samples, and the final validation cohort contained 250 samples.

Univariate and multivariate Cox regression analyses

The survival *coxph* function in R was used to analyze the DEGs among the molecular subtypes and the survival data through univariate Cox proportional hazard regression. We selected $p < 0.001$ as the filter threshold, and genes related to prognosis were obtained. The R package *glmnet* was used to perform Lasso regression on the DEGs and prognosis-related genes to compress the risk model to reduce the number of genes(Kukreja, 2006). The step method in the stats package in R starts from the most complex model and reduces the number of parameters by deleting 1 variable in turn. The smaller the step value, the more superior the model. This means that the fitting degree of the model is better with fewer parameters. The number of genes in the risk model was further reduced by using the Akaike information criterion (AIC) algorithm.

Gene set enrichment analysis

To observe the relationships between risk scores and biological functions in different samples, the gene expression profiles of these samples were selected for ssGSEA using the GSVA R software package. The ssGSEA scores of each function corresponding to each sample were obtained by calculating the scores of different functions for each sample. The correlations between these functions and risk scores were further evaluated. Functions with correlations > 0.4 were selected.

Construction and verification of nomogram

Nomograms can show the results of risk models intuitively and effectively, and they are convenient to use to predict outcomes. A nomogram uses line length to indicate the degree of influence of different variables and different values of variables on outcomes. Based on the results of multivariate cox analyses, a nomogram model was constructed (Balachandran et al.,, 2015).

Clinical expression of genes in the Oncomine

Oncomine (http://www.oncomine.org) is a gene chip-based database and integrated data-mining platform. In this study, we set the screening criteria as follows: 1) cancer type: LUAD; 2) analysis type: cancer vs normal analysis; and 3) threshold criteria: $p < 0.05$, fold change > 1.5, and gene rank = top 10%.

Immunohistochemistry and protein level validation

The Human Protein Atlas (HPA) provides information on the tissue and cell distributions of 26,000 human proteins. We explored protein levels relating to the 5 genes in normal lung and tumor tissues.

Pan cancer analysis of hub genes

The transcriptome data of 33 cancers from the UCSC Xena database were downloaded, and normal patients were removed, the ggplot2 package were used to plot the prognostic forest of KRT6A, MELTF, IRX5, MS4A1, CRTAC1 genes in 33 cancer types. For LUAD, we analyzed the expression differences of these 5 genes among different immune subtypes. According to the expression levels of 5 genes, LUAD patients were divided into high and low groups, and the *clusterProfilter* package was used for GSEA analysis.

Results

Study flow chart

To make the research easier for readers to understand, we drew a methodology flow chart (**Supplementary Fig. 1**).

Molecular typing of LUAD based on invasion-related genes

Through univariate Cox analysis of the 97 invasion-related genes in the TCGA expression profile, 19 genes were found to be associated with LUAD prognosis ($p < 0.01$; **Supplementary Table 3**). Consistent cluster analysis showed that the samples could be clustered at $k = 3$ (Fig. 1A). The expression levels of the invasion-related genes in the 3 subtypes are shown in Fig. 1B. These levels were different among the C1, C2, and C3 subtypes. Most of the genes were highly expressed in the C3 subtype and lowly expressed in the C2 subtype. We further analyzed the relationships between the 3 subtypes and prognosis. The results showed that there were significant differences between the 3 subtypes. The prognoses of patients with the C2 subtype were the best, and those of patients with the C3 subtype were the worst (log-rank $p < 0.05$; Fig. 1C, D).

Identification and functional analysis of DEGs among subtypes

The DEGs between C1, C2 and C3 were identified by using the *limma* package in R. The volcano map of the DEGs between C1 and C3 is shown in **Supplementary Fig. 2A**; there were 98 upregulated genes and 123 downregulated genes. The volcano map of the DEGs between C1 and C2 is shown in **Supplementary Fig. 2B**; there was 1 upregulated gene and 4 downregulated genes. The volcano map of the DEGs between C2 and C3 is shown in **Supplementary Fig. 2C**; there were 389 upregulated genes and 267 downregulated genes.

A total of 669 DEGs between C1/C2, C2/C3, and C1/C3 were obtained, and these DEGs were further analyzed by KEGG pathway and GO functional enrichment analyses using the *WebGestaltR* (V0.4.2) software package in R. The biological functions of the top 10 genes enriched in biological processes (**Supplementary Fig. 2D**), cellular components (**Supplementary Fig. 2E**), and molecular functions (**Supplementary Fig. 2F**) were visualized. The KEGG pathway analysis results showed that the DEGs were significantly enriched in the cell cycle, DNA replication, p53 signaling pathway, microRNAs in cancer, small cell lung cancer, and other tumor-related pathways (**Supplementary Fig. 2G**).

Clinical correlations of molecular subtypes and comparison with existing subtypes

The distributions of different clinical features among the C1, C2, and C3 subtypes were compared. The results showed that there were significantly more C2 patients than C1 and C3 patients in the T1, N0, and

Stage I samples, while there were significantly fewer C2 patients than C1 and C3 patients in the T2, N1, and Stage II samples (Fig. 1E-G). The number of survivors in the C2 group was significantly higher than in the C1 and C3 groups (Fig. 1H). These results confirmed that patients with the C2 subtype had the best prognoses.

Previous studies have analyzed 33 cancers in the TCGA database. These studies clustered non-blood tumors into 6 immune subtypes based on the distributions of various features, such as macrophages, immune-infiltrating lymphocytes, transforming growth factor-beta response, interferon-γ response, and wound healing; these subtypes include C1 (wound healing), C2 (INF-γ predominance), C3 (inflammation), C4 (lymphocyte depletion), C5 (immunological silencing), and C6 (transforming growth factor-beta predominance), among which C1 and C6 have been associated with poor prognosis(Thorsson et al.,, 2018). By comparing the molecular subtypes with these immune subtypes, it was found that most LUAD patients in the TCGA dataset belonged to the C1, C2, and C3 immune subtypes (about 89.5%), and there were no patients with the C5 immune subtype in the LUAD TCGA dataset (Fig. 1I). By comparing the distributions of the molecular and immune subtypes, it was found that patients with the C3 molecular subtype showed the highest proportion of the C2 immune subtype, reaching 54% (Fig. 1J). The proportion of the C2 immune subtype among the C2 molecular subtype was lower, and the proportion of the C3 immune subtype was higher than that of the C3 molecular subtype. The survival curve analysis results showed that there were significant differences in OS among the immune subtypes ($p < 0.05$; Fig. 1K). These results suggested that the prognosis of the C3 immune subtype was better than that of the C2 immune subtype.

Comparison of immune scores among subtypes

The relationships between the molecular subtypes of the TCGA dataset and immune scores were identified by using the ESTIMATE software package in R, MCPcounter, and the ssGSEA method in the GSVA package. The results showed that there were significant differences in immune scores among the different subtypes (Fig. 2A-C). The heat map of immune scores among the 3 subtypes is shown in Fig. 2D.

Construction of risk model

The 500 samples in the TCGA dataset were grouped according to the training set-to-validation set ratio of 1:1, and the univariate Cox proportional hazards regression model method was used to evaluate the 669 DEGs between the molecular subtypes. A total of 29 genes were found to be associated with prognosis (**Supplementary Table 4**). Lasso regression was used to further compress the 29 genes. The trajectory of each independent variable is shown in **Supplementary Fig. 3A**. As lambda decreased, the number of independent variable coefficients tending to 0 increased. We used 10-fold cross-validation to build the model and analyzed the confidence interval (CI) under each lambda (**Supplementary Fig. 3B**). When lambda equaled 0.003518527, the model was considered optimal, and 12 genes (*KRT6A*, *MELTF*, *IL20RB*, *PLEK2*, *LOXL2*, *IRX5*, *SLC16A11*, *FAM189A2*, *ITGA6*, *PKP2*, *MS4A1*, *CRTAC1*) were selected as target

genes. The AIC algorithm was used to further compress these 12 target genes, and 5 target genes (*KRT6A*, *MELTF*, *IRX5*, *MS4A1*, *CRTAC1*) were finally obtained.

The Kaplan-Meier curves of the 5 genes are shown in **Supplementary Fig. 3C-G**. The 5 genes could be divided into 2 groups with high and low risk ($p < 0.05$). The final 5-gene signature formula was: RiskScore = 0.08073881 * KRT6A + 0.18237095 * MELTF − 0.17903164 * IRX5−0.26862737 * MS4A1−0.09946249 * CRTAC1.

Risk scores were further converted into Z-scores. Samples with scores > 0 were divided into the high-risk group, and samples with scores < 0 were divided into the low-risk group. A total of 119 samples were divided into the high-risk group, and 131 samples were divided into the low-risk group. The survival curve results showed that there was a significant difference in prognosis between the 2 groups ($p < 0.0001$; Fig. 3A).

The risk score distributions of the samples were calculated according to expression levels and then plotted (Fig. 3B). The survival times of the samples with high risk scores were significantly shorter than those of the samples with low risk scores, suggesting that samples with high risk scores had worse prognoses. The timeROC software package in R was used to analyze the prognostic classification efficiency of risk scores. The model had a large area under the curve (AUC) at 1, 3, and 5 years; the 1-year AUC was 0.72, and the 5-year AUC was 0.74 (Fig. 3C).

Verification of risk model robustness in internal and external datasets

The robustness of the model was verified by the internal dataset (TCGA validation set and all datasets) and external dataset (GSE31210 dataset). In all datasets, the same model and coefficients as those in the training set were used. The survival curve showed significant differences between the high- and low-risk groups in the verification set and all datasets (Fig. 3D and Fig. 3G).The risk score of each sample was calculated according to gene expression, risk score distributions were plotted in TCGA internal validation set and all datasets in Fig. 3E and Fig. 3H. The classification efficiencies of prognosis prediction at 1, 3, and 5 years in the TCGA testing cohort and entire TCGA cohort are shown in Fig. 3F and Fig. 3I, respectively. The 1-year AUC reached 0.73 in both datasets.

Z-score transformation of risk scores was performed in GSE31210 dataset. Samples with risk scores > 0 after Z-score transformation were divided into the high-risk group, and samples with risk scores < 0 after Z-score transformation were divided into the low-risk group. This resulted in 94 samples in the high-risk group and 132 samples in the low-risk group. The survival curve showed a significant difference between the high- and low-risk groups ($p = 0.0028$; Fig. 3J).

The risk score distribution of the samples in the GSE31210 cohort was consistent with that of the training set (Fig. 3K). Receiver operating characteristic (ROC) analysis showed that the 1-year AUC reached 0.79 (Fig. 3L).

Relationships among risk scores, clinical features, and molecular subtypes

Survival analysis of different clinical subgroups was carried out based on risk scores. The results showed that the 5-gene signature could significantly distinguish age, sex, tumor/node/metastasis (TNM) stage, stage, recurrence, chemotherapy, and smoking status (current smoker, never smoked, reformed smoker) samples into high- and low-risk groups ($p < 0.05$; Fig. 4A-T). The 5-gene signature could not divide the M1 samples into 2 groups with significant prognostic difference, which may be due to the small number of M1 samples ($p > 0.05$; Fig. 4J). In general, our model could be used as a prognostic marker for different clinical subgroups if the sample size was appropriate.

Risk score distributions in terms of different clinical features were also assessed. The results showed that there were no significant differences in terms of age or stage ($p > 0.05$; Fig. 5B, E). Risk scores showed significant differences in terms of sex (female, male), T stage (T1, T2, T3, T4), N stage (N0, N1, N2, N3), stage (Stage I, Stage II, Stage III, Stage IV), and smoking status (current smoker, never smoke, reformed smoker) ($p < 0.05$; Fig. 5A, C, D, F, G). We also compared risk scores among the 3 subtypes (C1, C2, C3). The results showed that the risk scores of C3 subtype samples with poor prognosis were significantly higher than those of C2 subtype samples with good prognosis (Fig. 5H), which further suggested that high risk scores were associated with poor survival outcomes.

Relationships between risk scores and pathways

To observe the relationships between the risk scores and biological functions of different samples, GSEA was used to calculate the scores of different functions for each sample as well as correlations between these functions and risk scores. Correlation scores > 0.4 were considered to show positive correlations. Nine pathways were positively correlated with risk scores, and 10 pathways were negatively correlated with risk scores (**Supplementary Fig. 4A**). The 19 most relevant KEGG pathways were selected for cluster analysis (**Supplementary Fig. 4B**) based on their enrichment scores. Tumor-related pathways, such as KEGG_P53_SIGNALING_PATHWAY, KEGG_CELL_CYCLE, KEGG_MISMATCH_REPAIR, and KEGG_DNA_REPLICATION, were activated as risk scores increased, while others, such as KEGG_ARACHIDONIC_ACID_METABOLISM, KEGG_GLYCEROPHOSPHOLIPID_METABOLISM, and KEGG_ETHER_LIPID_METABOLISM, were deactivated as risk scores increased.

Construction of nomogram

Univariate and multivariate Cox regression analyses were used to analyze the independence of the 5-gene signature model in terms of clinical applications. Univariate analysis results showed that TNM stage, stage, and risk scores were significantly correlated with survival time; multivariate Cox regression analysis results showed that risk scores (HR = 1.63, 95% CI = 1.34–2.96, $p < 1e\text{-}5$) and N stage (HR = 1.99, 95% CI = 1.34–2.96, $p < 0.001$) were independent prognostic risk factors (**Supplementary Fig. 5A, B**). A nomogram was constructed based on the significant variables of multiple factors (Fig. 6A), and the

results showed that risk scores had the greatest effect on survival prediction, suggesting that the 5-gene signature was a good predictor of survival. Furthermore, by using calibration curves to evaluate the accuracy of the model (Fig. 6B), it was observed that the calibration curves at 1, 3, and 5 years were close to the standard curve, suggesting that the model had good prediction performance. Moreover, decision curve analysis was used to evaluate the model's reliability (Fig. 6C), and the results showed that the benefits of risk scores and the nomogram were significantly higher than those of the extreme curve, and the effect of the nomogram was higher than the effects of T stage, N stage, and risk scores, which were close to the extreme curve, suggesting that risk scores and the nomogram had good clinical applicability.

Comparison of risk model with other models

To prove the superiority of our model, 3 risk models, including an 8-gene signature (Li) (Li et al.,, 2018), a 3-gene signature (Yue) (Yue et al.,, 2019), and a 3-gene signature (Liu) (Liu et al.,, 2018), were chosen to compare with our 5-gene signature. To make the models comparable, we calculated the risk score of each LUAD sample in the TCGA dataset by the same method, and we evaluated the ROC curve of each model. Z-score transformation of risk scores was performed. The samples with risk scores > 0 after Z-score transformation were divided into the high-risk group, and samples with risk scores < 0 after Z-score transformation were divided into the low-risk group. The survival curves were plotted. The results showed that all 3 models could significantly classify the high- and low-risk groups into prognostic categories (Fig. 6E, G, I). However, the AUCs of the ROC curves of the 3 models were lower than those of the 5-gene signature at 1, 3, and 5 years in the TCGA dataset (Fig. 6D, F, H). These results showed that our model had good clinical predictive power.

Expression of 5 genes in 33 pan-cancers

The box diagram showed that *MS4A1* was significantly highly expressed in LUAD, HNSC, and kidney renal clear cell carcinoma, while in bladder carcinoma, colon adenocarcinoma, KICH, and READ tumors, *MS4A1* was significantly lowly expressed (Fig. 7A). Compared with normal samples, *KRT6A* and *MELTF* showed significantly high expression in most cancer types, including LUAD (Fig. 7B, **C**), while *CRTAC1* was expressed lowly in most cancer types, including LUAD (Fig. 7D, *IRX5* was significantly highly expressed in breast cancer, CHOL, colon adenocarcinoma, kidney renal papillary cell carcinoma, liver hepatocellular carcinoma, and READ tumors, while in KICH, kidney renal clear cell carcinoma, lung squamous cell carcinoma, and PRAD, *IRX5* was significantly lowly expressed (Fig. 7E).

Analysis of prognosis of genes in pan-cancer

We found that the MS4A1 gene is a protective gene in more than half of the tumors, which means, patients with high expression of MS4A1 have a better prognosis. KRT6A and MELTF genes are risk genes in most tumors including LUAD. CRTAC1 and IRX5 genes are significant protective genes in LUAD, while CRTAC1 gene is a high-risk gene in LUSC (Fig. 8A-E).

Clinical validation and gene set enrichment analysis of 5 genes

The results showed that in the Oncomine database, *CRAT1* was lowly expressed in 12 LUAD studies, *MS4A1* and *MELTF* were highly expressed in 1 LUAD study, *KRT6A* was highly expressed in 2 LUAD studies, and *IRX5* showed no significant expression in any LUAD study (Fig. 9A-E). In The Human Protein Atlas database, the immunochemistry results of the 5 genes were analyzed, but only 4 genes (*MS4A1*, *KRT6A*, *MELTF*, *CRTAC1*) had protein expression data. The results showed that the expression levels of *MS4A1*, *KRT6A*, and *MELTF* in tumor tissues were higher than in normal tissues, while *CRTAC1* expression in tumor tissues was lower than in normal tissues (Fig. 9F-I). At the same time, we found that the expression levels of 5 genes were significantly expressed in pan-cancer immune subtypes (Fig. 9J). Gene set enrichment analysis showed that MS4A1, KRT6A, and CRAT1 genes were both enriched in the HALLMARK_IL2_STAT5_SIGNALING pathway, and IRX5 and MELTF gene were both enriched in the HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION pathway. In addition, the KRT6A gene and CRAT1 gene were also enriched in the HALLMARK_KRAS_HALLMARK_UPRAS pathway (Fig. 9K-O).

Discussion

In this study, we first genotyped the 500 LUAD samples of the TCGA dataset based on 97 invasion-related genes, and we divided these samples into 3 subtypes, among which there were significant differences in prognosis. The C3 subtype had poor prognosis, and this was closely related to the pathways of tumorigenesis and development. A total of 669 DEGs were identified, and 5 target genes, including *KRT6A*, *MELTF*, *IRX5*, *MS4A1*, and *CRTAC1*, were obtained by using Lasso regression and the AIC algorithm. A 5-gene prognostic risk model was constructed. The KRT6A protein is a type II cytokeratin, and the *KRT6A* gene is highly expressed in different types of cancer(Chen and Shan, 2019, Ricciardelli et al.,, 2017). Some studies have shown that *KRT6A* is overexpressed in LUAD, and the overexpression of *KRT6A* is positively correlated with positive lymph nodes and invasive tumors. High expression of *KRT6A* in LUAD may promote the proliferation and metastasis of lung cancer through epithelial-mesenchymal transformation and cancer cell transformation(Yang et al.,, 2020). The KRT6A protein is a potential biomarker for distinguishing LUAD from squamous cell carcinoma(Xiao et al.,, 2017). The IRX5 is a transcription factor that is closely associated with a variety of malignancies(Holmquist Mengelbier et al.,, 2019, Zhu et al.,, 2020). IRX5 can promote the invasion and migration of colorectal cancer cells by inhibiting the RHOA-ROCK1-LIMK1 axis(Zhu et al.,, 2019). *IRX5* expression has been shown to be positively correlated with OS in smokers and negatively correlated with OS in non-smokers with LUAD(Zhang et al.,, 2018). *MS4A1* can be used as an immune-related gene to predict the prognoses of patients with LUAD(Song et al.,, 2020), and the dysregulation of the MS4A1 protein in interstitial lymphocytes may be involved in the progression of asbestos-related squamous cell carcinoma(Wright et al.,, 2012). Tissue and serum MELTF levels can be used as biomarkers of gastric cancer progression, and inhibition of *MELTF* expression can inhibit the invasive ability of gastric cancer cells(Sawaki et al.,, 2019). Cartilage acidic protein 1 (CRTAC1) is the extracellular matrix protein of human cartilage. CRTAC1

secreted by chondrocytes is the glycosylated extracellular matrix molecule of human articular cartilage(Steck et al.,, 2007). At present, there have been no studies of MELTF and CRTAC1 in LUAD, but such studies may provide new findings for prognostic markers of LUAD. We plan to further verify the mechanism of MELTF and CRTAC1 in LUAD.

We Z-scored the risk scores and divided the samples whose risk scores were > 0 into the high-risk group and those whose risk scores were < 0 into the low-risk group. The results showed that the high-risk score samples had significantly shorter survival times than the low-risk score samples. By analyzing the relationships between risk scores and pathways, we found that the tumor-related pathways of KEGG_P53_SIGNALING_PATHWAY, KEGG_CELL_CYCLE, KEGG_MISMATCH_REPAIR, and KEGG_DNA_REPLICATION increased with increased risk scores. The main ways to repair DNA include base excision repair, mismatch repair, nucleotide excision repair, and homologous recombination repair. DNA mismatch repair defects are important biomarkers for predicting the efficacy of immune checkpoint inhibitors in the treatment of many malignant tumors(Takamochi et al.,, 2017). Some genes, such as *MCM4*, *MCM5*, and *MCM8*, may affect LUAD prognosis by regulating the cell cycle, DNA replication, and other biological processes and pathways(Liu et al.,, 2019). However, the relationships between *KRT6A*, *MELTF*, *IRX5*, *MS4A1*, and *CRTAC1* and the p53 signaling pathway, the cell cycle, DNA mismatch repair, and DNA replication are still unclear. Our study may provide new ideas for the study of the mechanism of LUAD progression and metastasis.

According to the significant clinical characteristics in the univariate and multivariate regression analyses, T stage, N stage, and risk scores were used to construct the nomogram. Calibration and decision curve analysis curves suggested that the model had good prediction performance. Both the internal and external datasets also confirmed that the 5-gene signature was robust, and it could perform well in the independent dataset (GSE31210). Our model performed better than other models of LUAD. One advantage of our model is that targeted sequencing based on particular genes reduces health care costs significantly compared to whole-genome sequencing. Second, we selected invasion-related genes as the target genes, which is very important for the early diagnosis and prognosis prediction of LUAD. More importantly, in the routine clinical diagnosis and treatment process, patients' treatment plans and prognoses are largely dependent on pathological stage, the determination of which currently depends on the anatomic location of LUAD, so the biological heterogeneity of patients with LUAD is not currently being fully reflected. The nomogram we constructed can make up for this deficiency and provide a basis for the individualized treatment of patients with LUAD.

Gene expression was explored by using the Oncomine, GEO, and HPA databases. The results showed that the expression levels of *MS4A1* and *KRT6A* in tumor tissues were higher than in normal tissues, while *CRTAC1* expression in tumor tissues was lower than in normal tissues.

Our study has some limitations. First, the population in the TCGA database is predominantly white and Black, and our results need to be validated in other racial groups. Second, the construction of the alignment map was done retrospectively, so our results need to be further validated in multicenter clinical

trials and prospective studies. In the future, we will explore whether other regression modeling methods can further improve the prediction accuracy of the model.

Conclusion

In conclusion, we identified molecular subtypes of LUAD based on tumor invasion-related genes, and we developed a 5-gene signature prognostic hierarchical system. We recommend the use of this classifier as a molecular diagnostic test to assess the prognostic risk of LUAD.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Publicly available datasets were analyzed in this study. This data can be found here: TCGA.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

Not applicable.

Authors' contributions

All authors contributed to the literature investigation, data collection, writing the manuscript, providing useful discussion of its content, and undertaking reviews or revising the manuscript before submission.

Acknowledgements

Not applicable.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer.

136:E359-86.

2. Zheng M. (2016) Classification and Pathology of Lung Cancer. Surg Oncol Clin N Am. **25**:447-68.
3. Travis WD. (2011) Pathology of lung cancer. Clin Chest Med. **32**:669-92.
4. Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. (2016) Risk factors for lung cancer worldwide. Eur Respir J. **48**:889-902.
5. Pesch B, Kendzia B, Gustavsson P, Jockel KH, Johnen G, Pohlabeln H, et al. (2012) Cigarette smoking and lung cancer--relative risk estimates for the major histological types from a pooled analysis of case-control studies. Int J Cancer. **131**:1210-9.
6. Lin JJ, Cardarella S, Lydon CA, Dahlberg SE, Jackman DM, Janne PA, et al. (2016) Five-Year Survival in EGFR-Mutant Metastatic Lung Adenocarcinoma Treated with EGFR-TKIs. J Thorac Oncol. **11**:556-65.
7. Qi L, Li Y, Qin Y, Shi G, Li T, Wang J, et al. (2016) An individualised signature for predicting response with concordant survival benefit for lung adenocarcinoma patients receiving platinum-based chemotherapy. Br J Cancer. **115**:1513-9.
8. Zheng M, Hu Y, Gou R, Wang J, Nie X, Li X, et al. (2019) Integrated multi-omics analysis of genomics, epigenomics, and transcriptomics in ovarian carcinoma. Aging (Albany NY). **11**:4198-215.
9. Xing W, Zeng CJTB. (2016) An integrated transcriptomic and computational analysis for biomarker identification in human glioma.
10. Cancer Genome Atlas N. (2012) Comprehensive molecular portraits of human breast tumours. Nature. **490**:61-70.
11. Krzystanek M, Moldvay J, Szuts D, Szallasi Z, Eklund AC. (2016) A robust prognostic gene expression signature for early stage lung adenocarcinoma. Biomark Res. **4**:4.
12. Li J, Wang H, Li Z, Zhang C, Zhang C, Li C, et al. (2020a) A 5-Gene Signature Is Closely Related to Tumor Immune Microenvironment and Predicts the Prognosis of Patients with Non-Small Cell Lung Cancer. Biomed Res Int. **2020**:2147397.
13. He L, Chen J, Xu F, Li J, Li J. (2020) Prognostic Implication of a Metabolism-Associated Gene Signature in Lung Adenocarcinoma. Mol Ther Oncolytics. **19**:265-77.
14. Han L, Shi H, Luo Y, Sun W, Li S, Zhang N, et al. (2020) Gene signature based on B cell predicts clinical outcome of radiotherapy and immunotherapy for patients with lung adenocarcinoma. Cancer Med.
15. Li J, Li Q, Su Z, Sun Q, Zhao Y, Feng T, et al. (2020b) Lipid metabolism gene-wide profile and survival signature of lung adenocarcinoma. Lipids Health Dis. **19**:222.
16. Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, et al. (2019) CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res. **47**:D900-D8.
17. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. **43**:e47.

18. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. (2017) Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Rep. **18**:248-62.
19. Kukreja SL, Löfberg, J., & Brenner, M. J. . (2006) A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. IFAC proceedings volumes. **39**:814-9.
20. Balachandran VP, Gonen M, Smith JJ, Dematteo RPJLO. (2015) Nomograms in oncology: more than meets the eye. **16**:e173-e80.
21. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. (2018) The Immune Landscape of Cancer. Immunity. **48**:812-30 e14.
22. Li S, Xuan Y, Gao B, Sun X, Miao S, Lu T, et al. (2018) Identification of an eight-gene prognostic signature for lung adenocarcinoma. Cancer Manag Res. **10**:3383-92.
23. Yue C, Ma H, Zhou Y. (2019) Identification of prognostic gene signature associated with microenvironment of lung adenocarcinoma. PeerJ. **7**:e8128.
24. Liu WT, Wang Y, Zhang J, Ye F, Huang XH, Li B, et al. (2018) A novel strategy of integrated microarray analysis identifies CENPA, CDK1 and CDC20 as a cluster of diagnostic biomarkers in lung adenocarcinoma. Cancer Lett. **425**:43-53.
25. Chen C, Shan H. (2019) Keratin 6A gene silencing suppresses cell invasion and metastasis of nasopharyngeal carcinoma via the betacatenin cascade. Mol Med Rep. **19**:3477-84.
26. Ricciardelli C, Lokman NA, Pyragius CE, Ween MP, Macpherson AM, Ruszkiewicz A, et al. (2017) Keratin 5 overexpression is associated with serous ovarian cancer recurrence and chemotherapy resistance. Oncotarget. **8**:17819-32.
27. Yang B, Zhang W, Zhang M, Wang X, Peng S, Zhang R. (2020) KRT6A Promotes EMT and Cancer Stem Cell Transformation in Lung Adenocarcinoma. Technol Cancer Res Treat. **19**:1533033820921248.
28. Xiao J, Lu X, Chen X, Zou Y, Liu A, Li W, et al. (2017) Eight potential biomarkers for distinguishing between lung adenocarcinoma and squamous cell carcinoma. Oncotarget. **8**:71759-71.
29. Holmquist Mengelbier L, Lindell-Munther S, Yasui H, Jansson C, Esfandyari J, Karlsson J, et al. (2019) The Iroquois homeobox proteins IRX3 and IRX5 have distinct roles in Wilms tumour development and human nephrogenesis. J Pathol. **247**:86-98.
30. Zhu L, Dai L, Yang N, Liu M, Ma S, Li C, et al. (2020) Transcription factorIRX5 promotes hepatocellular carcinoma proliferation and inhibits apoptosis by regulating the p53 signalling pathway. Cell Biochem Funct. **38**:621-9.
31. Zhu Q, Wu Y, Yang M, Wang Z, Zhang H, Jiang X, et al. (2019) IRX5 promotes colorectal cancer metastasis by negatively regulating the core components of the RHOA pathway. Mol Carcinog. **58**:2065-76.
32. Zhang DL, Qu LW, Ma L, Zhou YC, Wang GZ, Zhao XC, et al. (2018) Genome-wide identification of transcription factors that are critical to non-small cell lung cancer. Cancer Lett. **434**:132-43.

33. Song C, Guo Z, Yu D, Wang Y, Wang Q, Dong Z, et al. (2020) A Prognostic Nomogram Combining Immune-Related Gene Signature and Clinical Factors Predicts Survival in Patients With Lung Adenocarcinoma. Front Oncol. **10**:1300.
34. Wright CM, Savarimuthu Francis SM, Tan ME, Martins MU, Winterford C, Davidson MR, et al. (2012) MS4A1 dysregulation in asbestos-related lung squamous cell carcinoma is due to CD20 stromal lymphocyte expression. PLoS One. **7**:e34943.
35. Sawaki K, Kanda M, Umeda S, Miwa T, Tanaka C, Kobayashi D, et al. (2019) Level of Melanotransferrin in Tissue and Sera Serves as a Prognostic Marker of Gastric Cancer. Anticancer Res. **39**:6125-33.
36. Steck E, Braun J, Pelttari K, Kadel S, Kalbacher H, Richter W. (2007) Chondrocyte secreted CRTAC1: a glycosylated extracellular matrix molecule of human articular cartilage. Matrix Biol. **26**:30-41.
37. Takamochi K, Takahashi F, Suehara Y, Sato E, Kohsaka S, Hayashi T, et al. (2017) DNA mismatch repair deficiency in surgically resected lung adenocarcinoma: Microsatellite instability analysis using the Promega panel. Lung Cancer. **110**:26-31.
38. Liu K, Kang M, Liao X, Wang R. (2019) Genome-wide investigation of the clinical significance and prospective molecular mechanism of minichromosome maintenance protein family genes in patients with Lung Adenocarcinoma. PLoS One. **14**:e0219467.

Table

Table 1Sample information

Clinical Features	TCGA-LUAD	GSE31210
OS		
0	318	191
1	182	35
Gender		
Female	270	121
Male	230	105
Age		
≤60	157	0
>60	343	226
T Stage		
T1	167	
T2	267	
T3	45	
T4	18	
N Stage		
N0	324	
N1	94	
N2	69	
N3	2	
M Stage		
M0	332	
M1	24	
Stage		
I	268	168
II	119	58
III	80	0
IV	25	0
Smoking history		

1	71
2	119
3	129
4	163
5	4

Figures

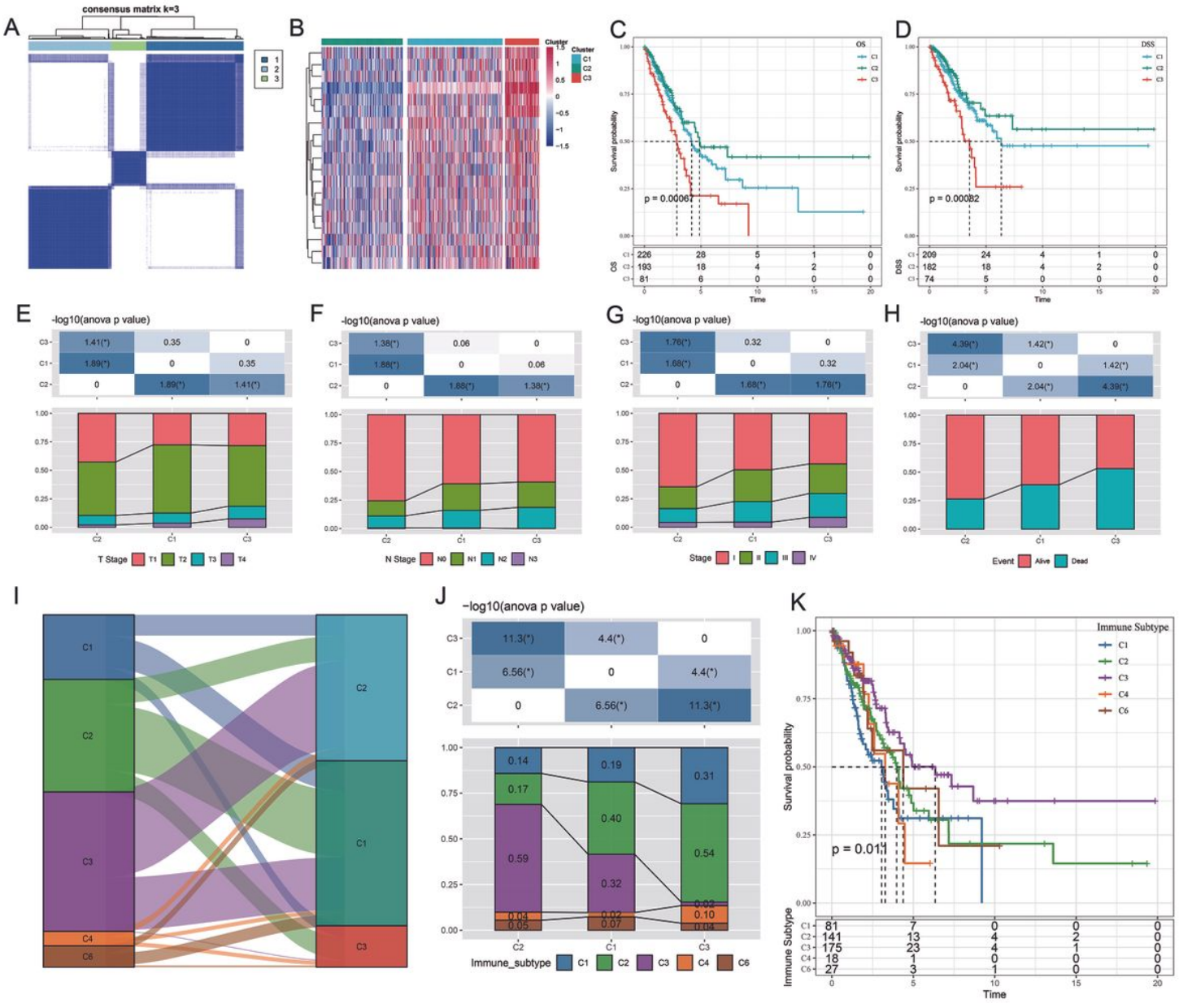

Figure 1

A: Cluster thermograms of samples with consistent clusters of k = 3. B: Cluster thermograms of prognosis-related invasion genes. C: Survival curves of TCGA lung adenocarcinoma samples with different molecular subtypes. D: TCGA lung adenocarcinoma samples according to different molecular subtypes. E-H: Distribution comparison of clinical features among the 3 subtypes of the TCGA dataset. I: Sanki map of molecular subtypes compared with existing molecular immune subtypes. J: Distribution of molecular subtypes compared with existing immune subtypes. K: Survival curves of the molecular immune subtypes.

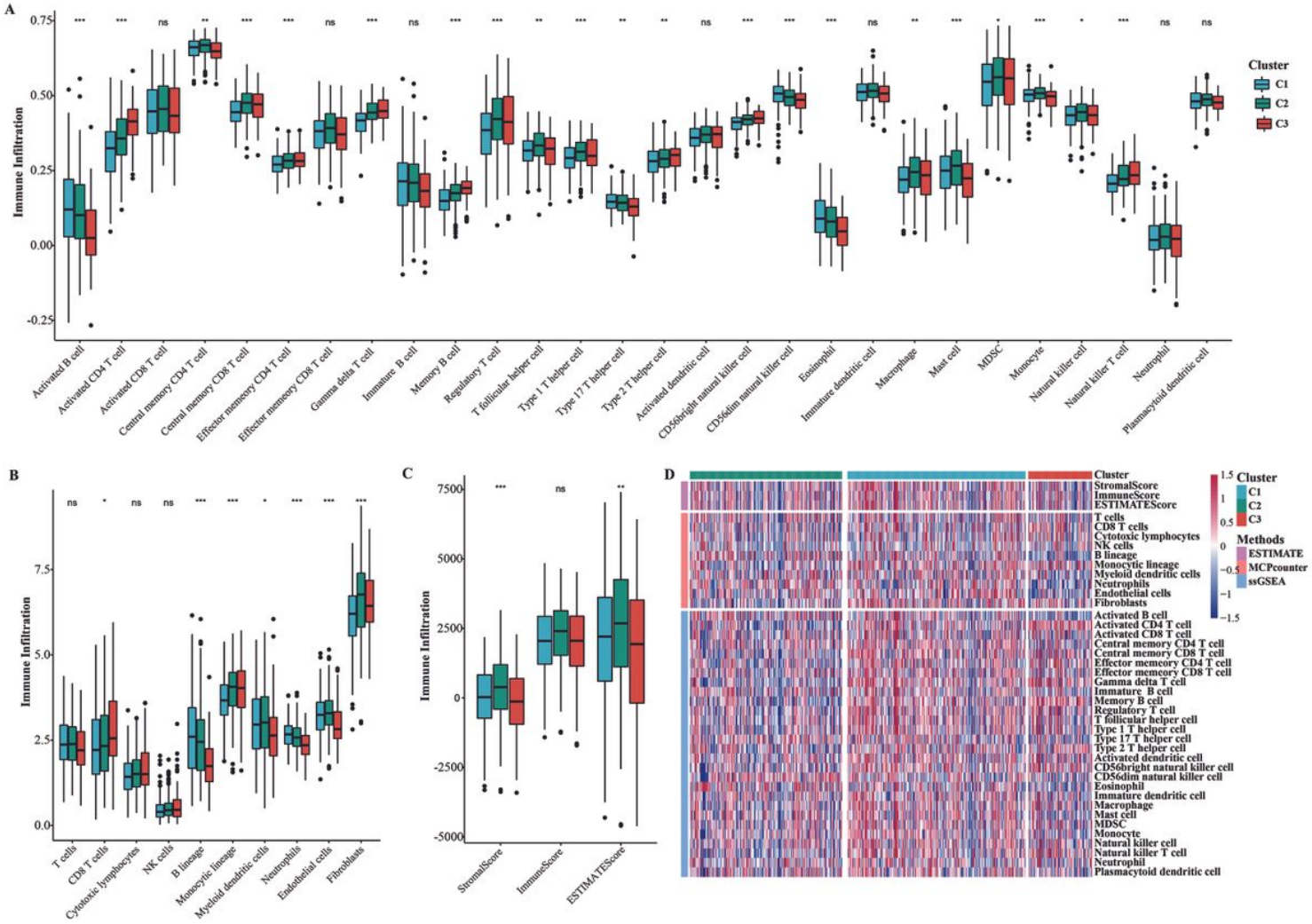

Figure 2

A: Comparison of the ssGSEA immune scores among the subtypes of the TCGA dataset. B: Comparison of the MCPcounter immune scores among the subtypes of the TCGA dataset. C: Comparison of the estimate immune scores among the subtypes of the TCGA dataset. D: Comparison of all 3 immune score types among the molecular subtypes of the TCGA dataset.

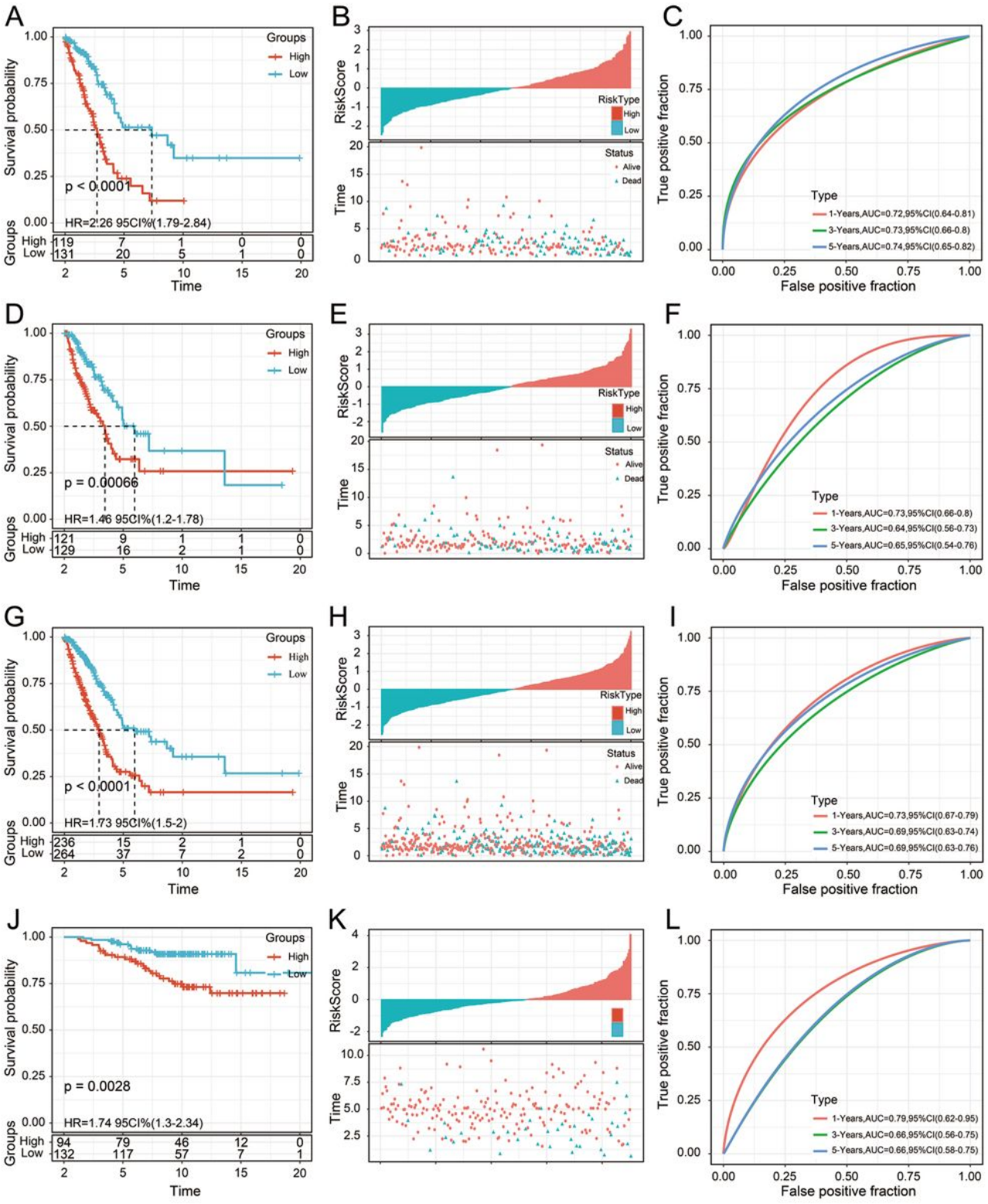

Figure 3

A: Survival curves between the 2 risk groups based on the 5-gene signature classification. B: Distributions of risk scores and survival status based on the 5-gene signature in the TCGA training cohort. C: ROC curve of the 5-gene signature classification in the TCGA training cohort. D-F: Survival curves between the 2 risk groups, distributions of risk scores and survival status, and the ROC curve of the 5-gene signature in the TCGA testing cohort. G-I: Survival curves between the 2 risk groups, distributions of risk scores and

survival status, and the ROC curve of the 5-gene signature in the entire TCGA cohort. J-L: Survival curves between the 2 risk groups, distributions of risk scores and survival status, and the ROC curve of the 5-gene signature in the GSE31210 cohort.

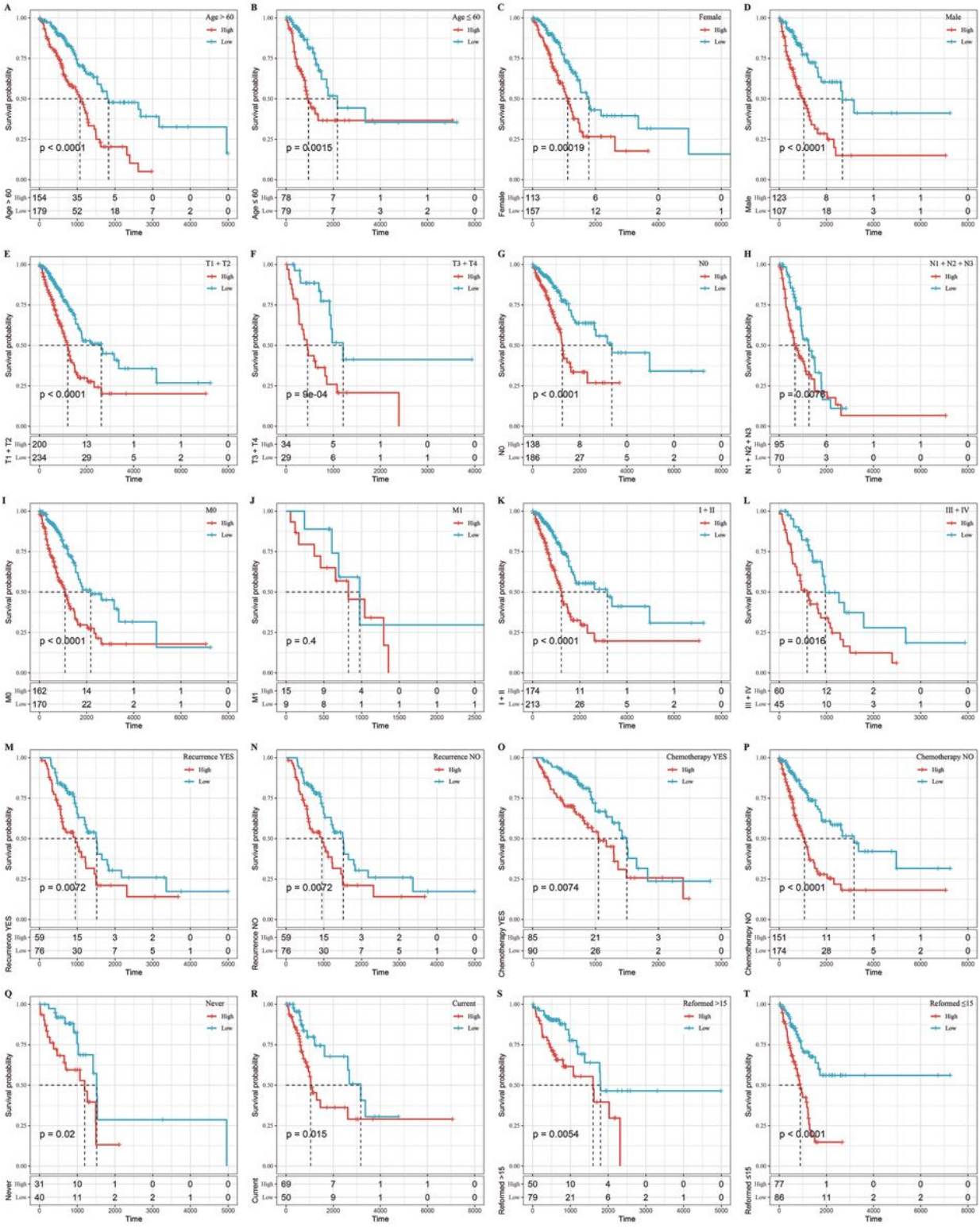

Figure 4

Prognostic performance of the 5-gene signature in terms of different clinical features.

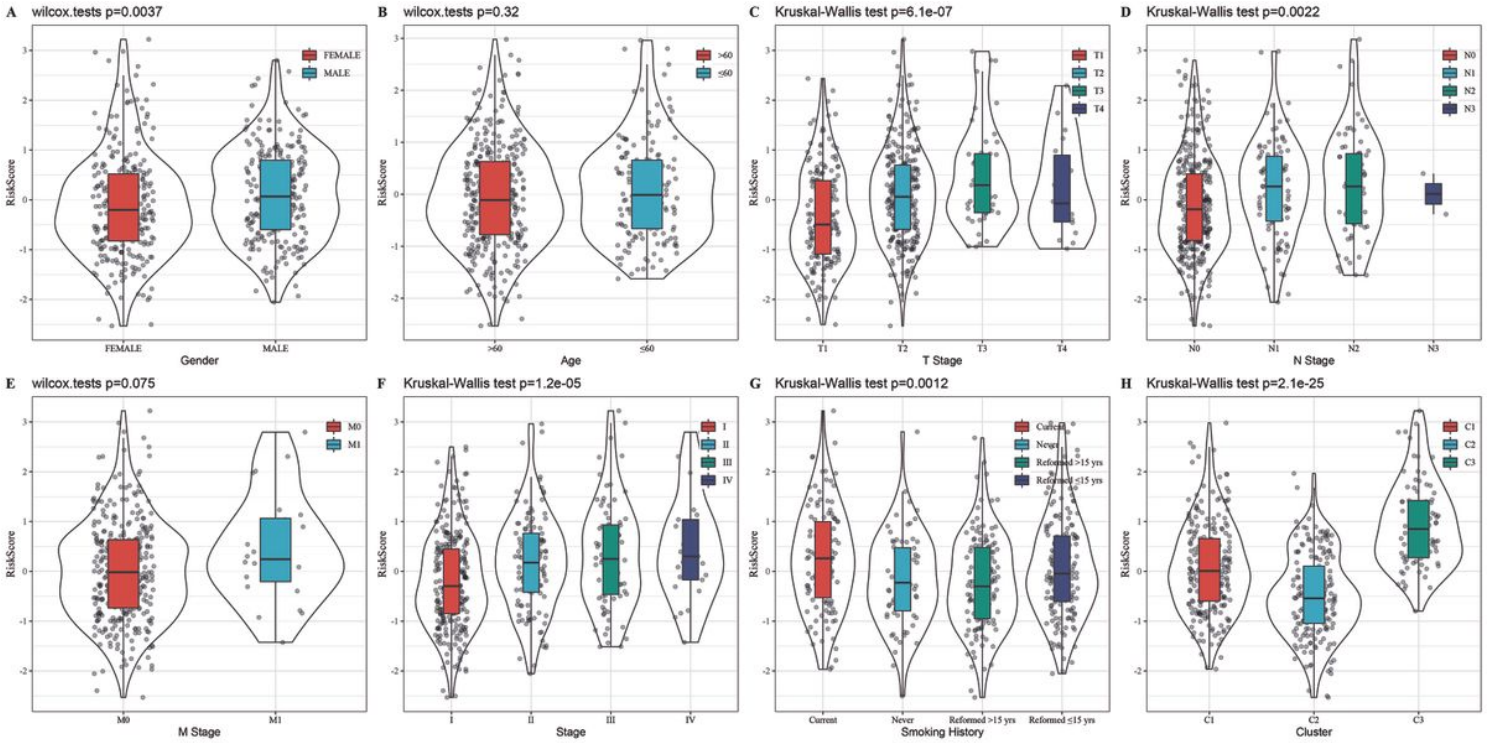

Figure 5

A: Risk score comparison based on sex. B: Risk score comparison based on age. C: Risk score comparison based on T stage. D: Risk score comparison based on N stage. E: Risk score comparison based on M stage. F: Risk score comparison based on clinical stage. G: Risk score comparison based on smoking history. H: Risk score comparison based on molecular subtype.

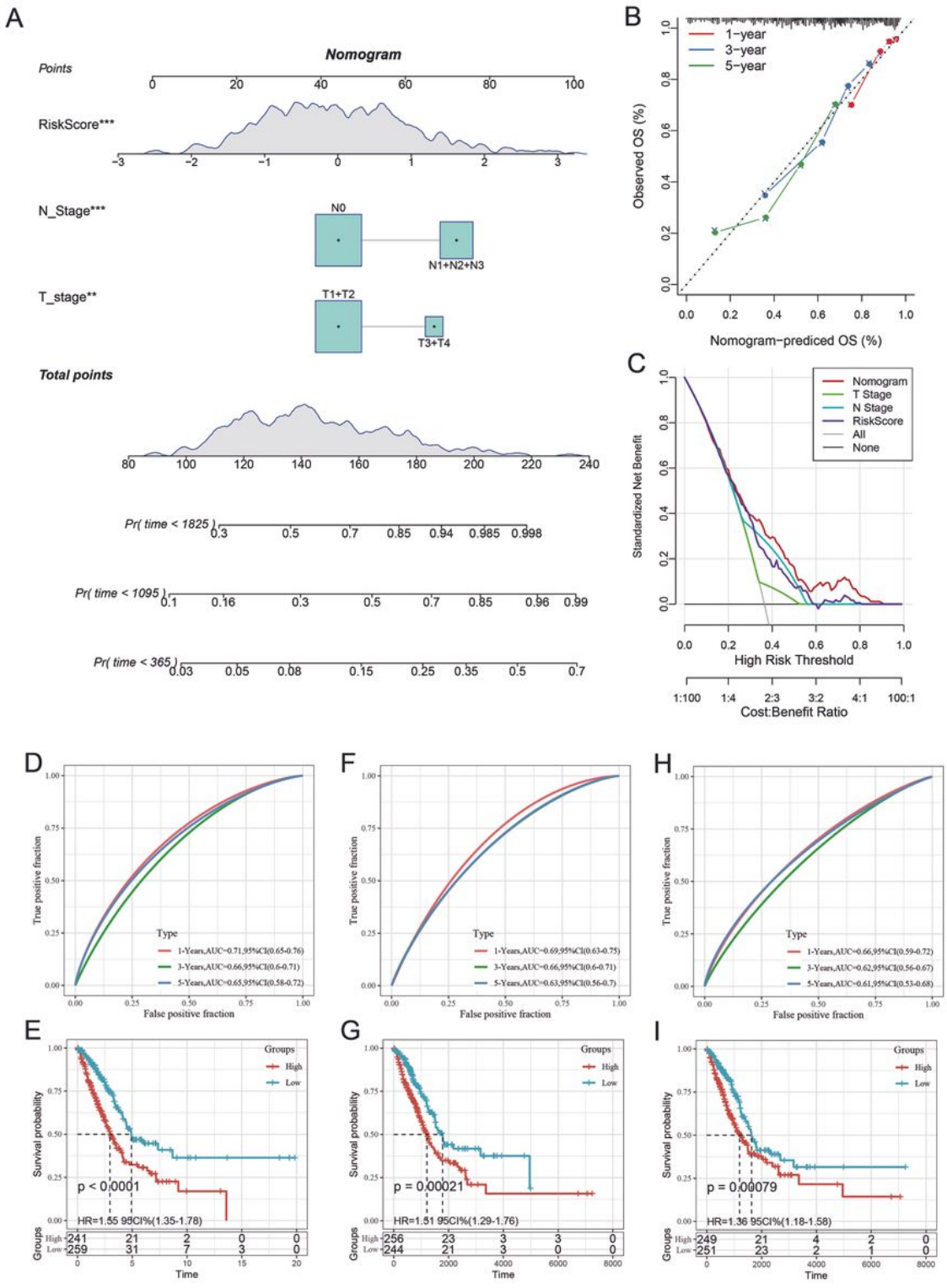

Figure 6

A: Construction of the nomogram model. B: Calibration curves at 1, 3, and 5 years using the nomogram. C. Decision curve analysis of age, M stage, clinical stage, risk score, and nomogram results. D-E: ROC curve of the 8-gene signature risk model (Li) and Kaplan-Meier curves of the high- and low-risk LUAD samples. F-G: ROC curve of the 3-gene signature risk model (Yue) and Kaplan-Meier curves of the high-

and low-risk LUAD samples. H-I: ROC curve of the 3-gene signature risk model (Liu) and Kaplan-Meier curves of the high- and low-risk LUAD samples.

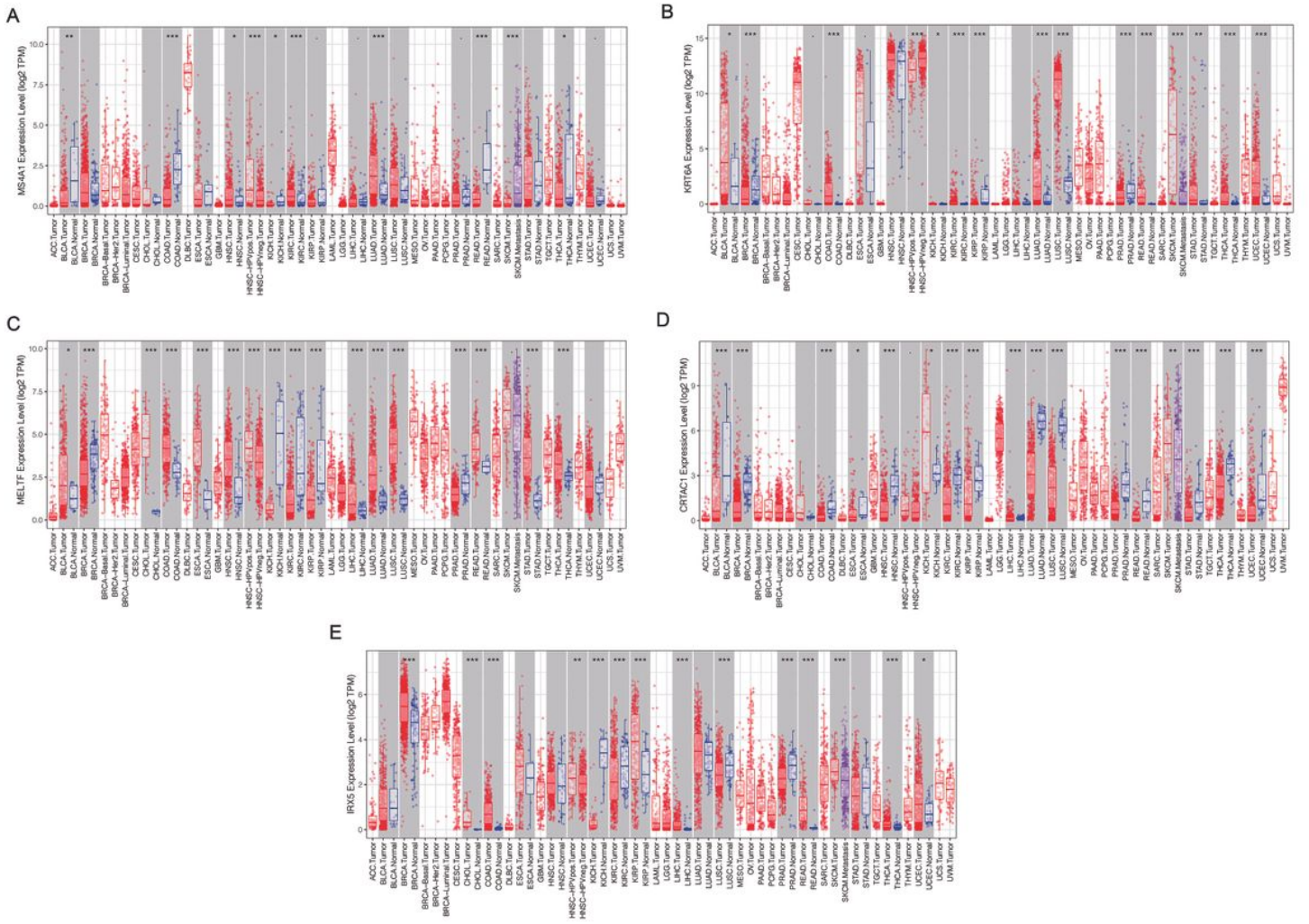

Figure 7

Expression of 5 genes in 33 pan-cancers The box diagram showed that MS4A1 was significantly highly expressed in LUAD, HNSC, and kidney renal clear cell carcinoma, while in bladder carcinoma, colon adenocarcinoma, KICH, and READ tumors, MS4A1 was significantly lowly expressed (Figure 7A). Compared with normal samples, KRT6A and MELTF showed significantly high expression in most cancer types, including LUAD (Figure 7B, C), while CRTAC1 was expressed lowly in most cancer types, including LUAD (Figure 7D, IRX5 was significantly highly expressed in breast cancer, CHOL, colon adenocarcinoma, kidney renal papillary cell carcinoma, liver hepatocellular carcinoma, and READ tumors, while in KICH, kidney renal clear cell carcinoma, lung squamous cell carcinoma, and PRAD, IRX5 was significantly lowly expressed (Figure 7E).

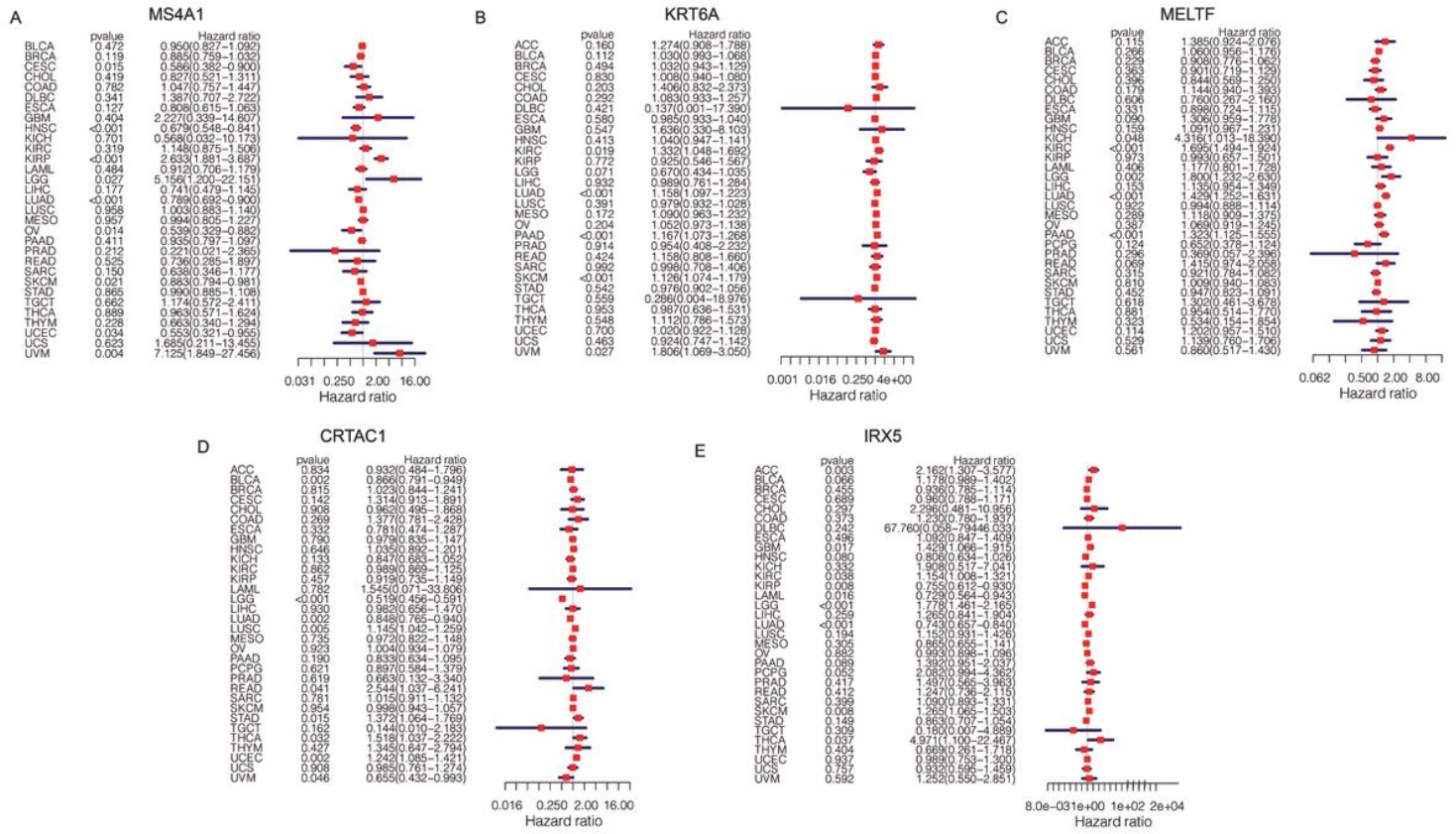

Figure 8

Analysis of prognosis of genes in pan-cancer. A. MS4A1 gene, B. KRT6A gene, C. MELTF gene. D. CRTAC1 gene. E. IRX5 gene. The abscissa represents survival time, and the ordinate represents survival probability.

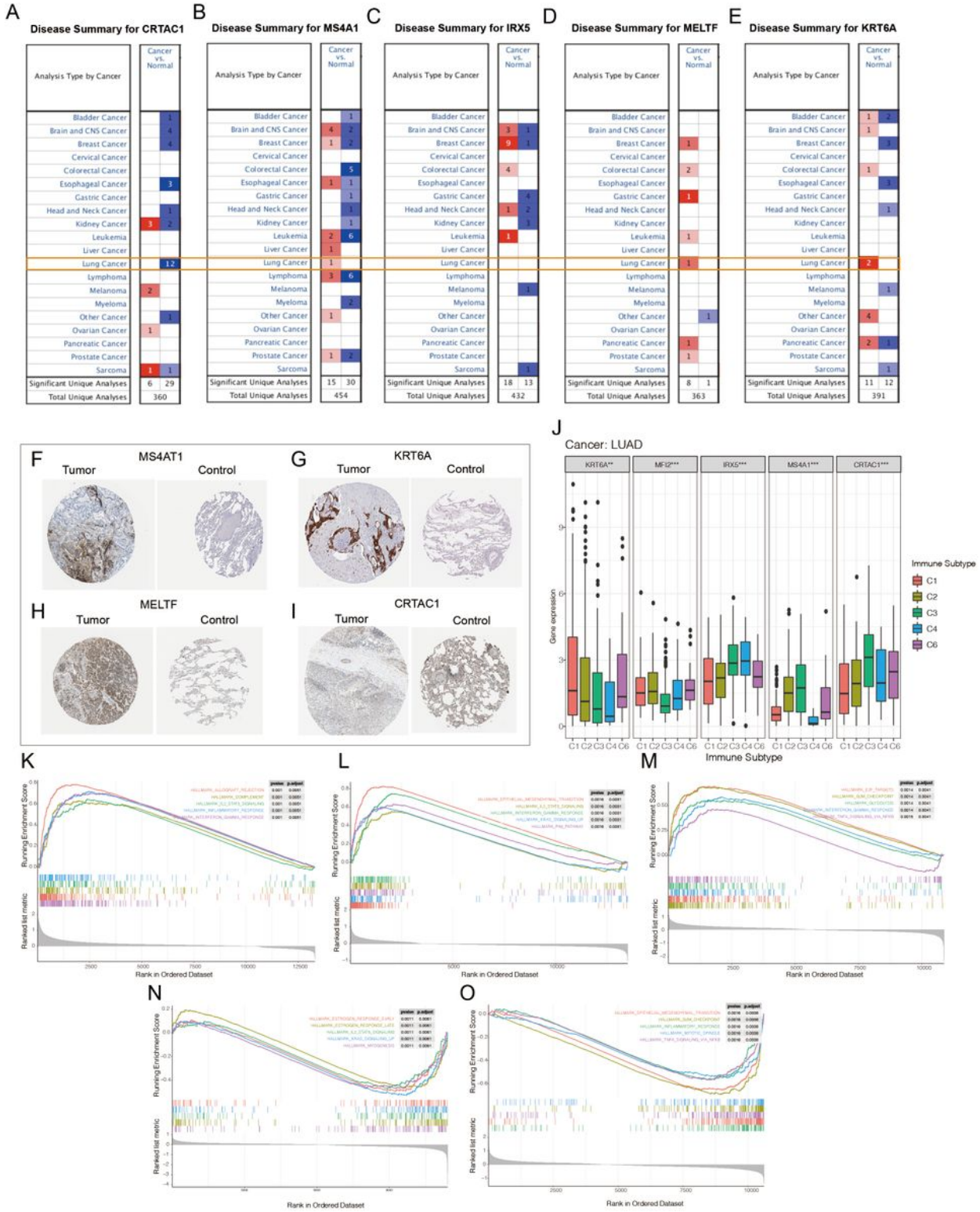

Figure 9

Expression box diagrams of gene expression in pan-cancer. Expression of MS4A1 (A), KRT6A (B), MELTF (C), CRTAC1 (D), and IRX5 (E) in different tumors. F: MS4A1 protein expression in cancer and normal control samples. G: KRT6A protein expression in cancer and normal control samples. H: MELTF protein expression in cancer and normal control samples. I: CRTAC1 protein expression in cancer and normal control samples. J. Differential expression of genes in pan-cancer immune subtypes. K-O. Gene

functional enrichment analysis: K. MS4A1 gene, L. KRT6A gene, M. MELTF gene. N. CRTAC1 gene. O. IRX5 gene.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigure1.jpg
- SupplementaryFigure2.jpg
- SupplementaryFigure3.jpg
- SupplementaryFigure4.jpg
- SupplementaryFigure5.jpg
- SupplementaryTable1.docx
- SupplementaryTable2.docx
- SupplementaryTable3.docx
- SupplementaryTable4.docx