Am J Transl Res 2024;16(5):1568-1586
www.ajtr.org /ISSN:1943-8141/AJTR0156174

Original Article

Identification of an 11-miRNA-regulated and surface-protein genes signature predicts the prognosis of lung adenocarcinoma based on multi-omics study

Kunyu Guo[1], Zhenbo Qu[1], Yibo Yu[1], Chendan Zou[2]

[1]*The First Affiliated Hospital of Harbin Medical University, Harbin 150000, Heilongjiang, China;* [2]*Department of Biochemistry and Molecular Biology, Harbin Medical University, Harbin 150000, Heilongjiang, China*

Received February 27, 2024; Accepted April 23, 2024; Epub May 15, 2024; Published May 30, 2024

Abstract: Lung adenocarcinoma (LUAD) is one of the most prevalent and lethal cancers worldwide, signifying a critical need for improved prognostic tools. A growing number of studies have highlighted the role of microRNAs (miRNAs) and their regulatory functions in tumorigenesis and cancer progression. In this context, we performed an extensive analysis of bulk RNA- and miRNA-sequencing to identify LUAD-associated prognostic genes. A risk score system based on 11 miRNA-regulated and surface-protein genes was developed, which was later validated by internally and externally using the Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO), respectively. Further single-cell RNA sequencing analysis revealed significant interactions between various cellular subpopulations within the tumor microenvironment, with the most pronounced differences observed between endothelial and epithelial cells. The mutational analysis highlighted TP53 as a key signaling pathway associated with the risk score. The study underscores that immune suppression, indicated by a positive association with regulatory T cells (Tregs) and an inverse correlation with M1-type macrophages, is prevalent in high-risk LUAD patients. These findings provide a promising prognostic tool for clinical outcomes of LUAD patients, facilitating future development of therapeutic strategies and enhancing our understanding of the regulatory function of miRNAs in LUAD.

Keywords: Bulk-RNA sequencing, lung adenocarcinoma, miRNA sequencing, risk score, single-cell RNA sequencing, TCGA and GEO databases

Introduction

Lung adenocarcinoma (LUAD), a prevalent histologic subtype of non-small-cell lung cancer, is found in 38.5% of all lung cancer cases [1]. LUAD exhibits unique pathological and genetic features that impact patients' prognosis and treatment. The prognosis of LUAD patients remains unsatisfactory despite advances in LUAD diagnosis and management, with a best five-year survival of 70%, which drops to 30% in locally advanced patients [2] and to 5% in those with metastases [3]. The identification of clinically relevant prognostic genome-wide survival models for cancer using gene expression has gained considerable attention [4]. Currently, numerous studies have constructed gene expression-based survival scores in LUAD to predict patients' survival [5-9]. Survival scores are predictive tools that utilize gene expression profiling to stratify patients according to their risk of adverse clinical outcomes. For instance, a recent study identified a 4-mRNA signature-based prognostic model, with its robust prognostic power demonstrated, which is a better predictor of clinical outcomes than traditional methods for LUAD [10]. Moreover, through the identification of crucial molecular pathways involved in the pathogenesis of LUAD, these gene-based prognostic models could enable more personalized treatment options, potentially improving the outcomes of patients with LUAD. Although the early results are promising, there is still a need for innovative and refined gene-based prognostic models, as well as further research exploring the potential of gene-based survival scores in the precision treatment of LUAD.

MicroRNAs (miRNAs) play a fundamental role in controlling gene expression through their bind-

https://doi.org/10.62347/CWMT4815

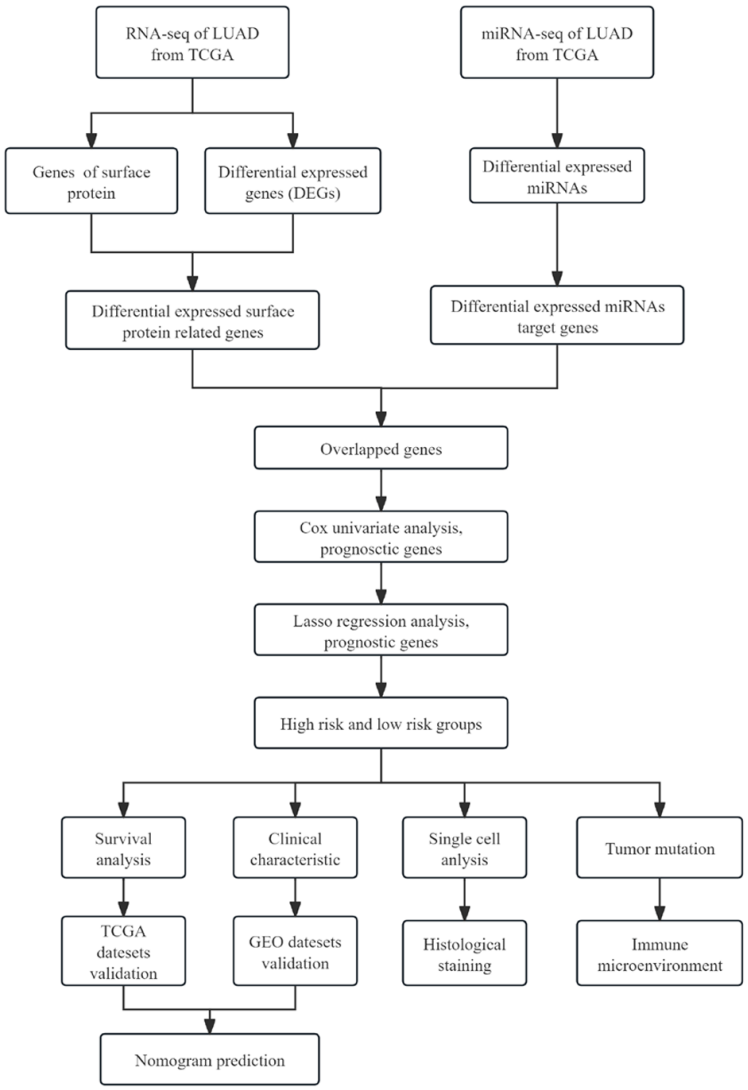


Figure 1. Workflow of data processing, analysis, and validation in the current study. LUAD, lung adenocarcinoma; TCGA, cancer Genome Atlas; GEO, Gene Expression Omnibus.

ing to their target 3' untranslated region (UTR) of mRNAs, resulting in post-translational repression or degradation of target mRNAs [11]. MiRNAs have been shown to considerably contribute to the pathogenesis and progression of various cancers [12, 13], including LUAD [2, 14]. Numerous miRNAs have been identified as prognostic markers, suggesting their potential as critical tools for predicting patient prognosis. For example, miR-21 has repeatedly been linked to poor prognosis in LUAD patients since its upregulation promotes tumor growth, invasion, and metastasis [15]. Additionally, miR-let7 family members, particularly miR-let7a, have suppressive effects on LUAD, and their downregulation is correlated with advanced disease stages and shorter overall survival (OS) [16]. These studies highlight the prognostic significance of specific miRNAs in LUAD, underscoring their potential as essential indicators for guiding treatment decisions and predicting patient prognosis [16]. The miRNA-mRNA regulatory network in LUAD is complex and involves multiple interactions among different molecules. Several studies have explored the correlation between specific miRNAs and target mRNAs in LUAD, highlighting the potential of these regulatory networks as prognostic biomarkers. For instance, miR-21 and miR-221/222 have been identified as oncogenic miRNAs that promote tumorigenesis in LUAD by targeting several tumor suppressor genes, including PTEN, PDCD4, and TIMP3 [17].

The application of computational methods to the miRNA-mRNA regulatory network in LUAD has facilitated the identification of novel therapeutic targets and prognostic markers [17, 18]. Co-expression network analysis, for instance, has been used to identify differentially expressed miRNA-mRNA signatures associated with LUAD prognosis [19]. Other research has explored the use of machine learning algorithms to integrate multiple datasets and identify regulatory modules in LUAD [19]. A comprehensive understanding of the miRNA-mRNA regulatory network in LUAD is essential for identifying key biomarkers, predicting patient outcomes, and developing new interventions.

In this study, we combined mRNA- (mRNA-seq) and miRNA-sequencing (miRNA-seq) results from the open database, The Cancer Genome Atlas (TCGA), as shown by the workflow (**Figure 1**). By analyzing genes from the differentially expressed gene (DEG) list from mRNA-seq and

predicting regulatory genes from the DEG list from miRNA-seq, we generated a series of overlay genes highly likely to be regulated by miRNAs and influence LUAD progression. Based on this gene list, we selected genes related to prognosis and constructed a risk score system (RSS). We used multiple datasets to validate the prognostic value of this risk system, demonstrating its high sensitivity and specificity in predicting patients' survival outcomes.

Data and methods

Data acquisition

We obtained datasets from the open Gene Expression Omnibus (GEO; URL: https://www.ncbi.nlm.nih.gov/geo/) and TCGA (URL: https://portal.gdc.cancer.gov/). The LUAD original mRNA expression data, consisting of a normal (n=59) and a tumor (n=541) group, as well as the LUAD original miRNA data that included a normal (n=46) and a tumor (n=518) group, were downloaded from TCGA. The DEGs (screening criteria: *P*-value <0.05 & |logFC| >1) between groups was identified using the limma package. The following datasets were downloaded from GEO. The Series Matrix File data of the GSE30219 dataset (epigenetic alterations, bulk RNA-seq) was downloaded, and information on survival and complete gene expression profiles of 85 LUAD patients were extracted, with the annotation platform being GPL570. Series Matrix File data (bulk RNA-seq) was also downloaded from the GSE50081 dataset, as well as information on survival and complete gene expression profiles of 127 LUAD patients, with the annotation platform also being GPL570. The GSE149655 data file and data from four patients with LUAD were extracted for single-cell analysis (SCA) by single-cell RNA sequencing.

Selection of surface protein genes

We searched surface protein-associated gene sets in the GeneCards database (URL: https://www.genecards.org/), extracted those with a relevance score >20, and intersected them with the DEGs.

Prediction of miRNA target genes

miRNA-mRNA interactions were predicted by miRDB, miRTarBase, and TargetScan jointly, and the co-recognized targeted mRNAs were used for subsequent research. Cytoscape (v3.8.2) was used for the visualization of the gene miRNA-mRNA networks.

Construction of a nomogram model

A nomogram was built based on regression analysis, which was plotted on the same plane with scaled line segments according to a certain ratio based on risk scores and clinical manifestations to express the relationship among variables in the prediction model. Through the construction of a multivariate Cox regression model, the value of each influencing factor was scored according to its contribution degree in the model to the outcome variable (that is, the size of the regression coefficient). The scores were then added together to get the total score, based on which the predicted value was calculated.

Gene set enrichment analysis (GSEA)

The patients were assigned to either a high- or a low-risk group according to the high or low risk score, for GSEA of the inter-group differences in signaling pathways. The background gene set utilized in the GSEA was obtained from the Molecular Signatures Database (MSigDB, version 4.1.0; URL: https://www.gsea-msigdb.org/gsea/msigdb/) and was designated as the annotated gene set of the subtype pathway. Through differential pathway expression analyses among different subtypes, the differentially enriched gene sets were identified based on the consistency score, whereby those with adjusted *p*-values <0.05 were selected.

Analysis of immune cell infiltration

We used the CIBERSORT algorithm to analyze patient data and infer the relative proportions of 22 kinds of tumor-infiltrating immunocytes. Then, gene expression profiles and immune cell contents were analyzed by Spearman correlation analysis.

Tumor mutation burden (TMB) and neoantigen data analysis

TMB was defined by calculating variant frequencies and variant counts/exon lengths for each tumor sample, dividing the non-synony-

mous mutated loci by the total length of the protein-coding region [20]. The neoantigen of each patient was assessed using NetMHCpanv3.0 [21].

Single cell analysis (SCA)

This study used the Seurat package to process the data, the tSNE algorithm to determine the positional relationship between clusters, and the Celldex package to annotate the clusters; cells that were vital for tumorigenesis were annotated. By setting FindAllMarkers (the logic threshold parameter) to 1, we finally extracted the marker genes for each cell subtype from the single-cell expression profile.

Intercellular communication analysis

CellChat can quantitatively infer and analyze cell-to-cell communication networks from single-cell data. This analysis quantified the closeness of the interactions in terms of the number of times (count), using standardized single-cell expression profiles and cell subtypes obtained from the SCA as the input data and SCA information, respectively, to observe cell activities and influence on the disease.

Immunohistochemical staining (IHC)

The immunohistochemical data of 9 model genes in LUAD or normal lung tissue specimens were obtained from the Human Protein Atlas (HPA; URL: https://www.proteinatlas.org/). Four paired LUAD tissue and normal lung tissue specimens were collected from LUAD patients undergoing pulmonary resection at the Department of Pathology, Fenyang Hospital, Shanxi Province. Tissue sections (thickness: 5 μm) prepared from representative tumor-containing FFPE blocks were subjected to IHC for TGFβ-2 (1:50, Cat# ab80059, Abcam) and ECT2 (1:100, Cat# ab236502, Abcam).

Statistical methods

Patient survival was visualized by Kaplan-Meier curves and comparatively analyzed by the Log-rank test. Multivariate analyses were carried out with the Cox proportional hazards model. R language was employed for statistical analyses, with the significance level set as $P<0.05$.

Results

Screening of cell surface protein-related genes that were differentially expressed in the LUAD cohort

By utilizing mRNA expression profile from the TCGA database, we initially identified 1969 DEGs, comprising 700 up- and 1269 down-regulated genes (**Figure 2A**, **2B**). Surface proteins located on the outer surface of the cell membrane offer convenient access for detection and analysis in both tumor tissue and blood samples. Furthermore, surface proteins show greater stability compared to plasma proteins or nuclear transcription factors, enabling them to endure multiple freeze-thaw cycles and making them more suitable for long-term storage and analysis. Consequently, we proceeded to select all genes that transcribe surface proteins by overlaying the surface protein-coding genes with the entire DEG dataset (**Figure 2C**).

Identification of differentially expressed miRNAs and target gene prediction in the LUAD cohort

Through miRNA expression data analysis from the TCGA database, 243 differentially expressed miRNAs were identified in LUAD, including 132 up- and 111 down-regulated genes (**Figure 3A**, **3B**). miRNAs generally bind to their target 3'UTR of mRNAs, causing mRNA degradation or decreased trancription, thus providing critical insight into post-transcriptional gene regulation. To predict miRNA-mRNA interactions, we integrated miRDB, miRTarBase, and TargetScan databases, resulting in 2,990 predicted miRNA-target mRNAs, which we narrowed down to commonly identified candidates. Cytoscape was then utilized to visualize these miRNA-mRNA interactions (Figure S1). We performed an intersect analysis of the 2,990 identified DEGs with 509 cell surface protein genes and obtained a set of 100 intersecting genes to prioritize future studies (**Figure 3C**), representing potential targets for future investigation in LUAD.

Prognosis-related gene identification and prognosis model building

Clinical data was collected from LUAD patients to identify prognostic genes among the 100 intersection genes using univariate Cox regres-

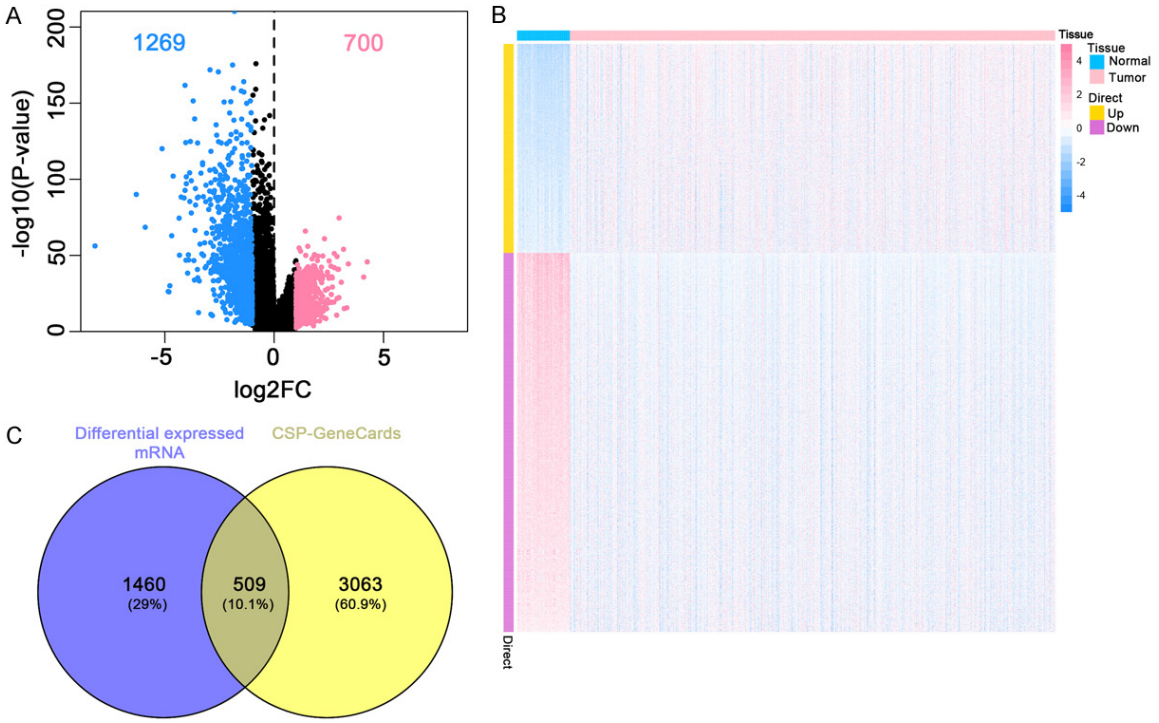


Figure 2. Identification of differentially expressed cell surface protein-related genes. A. Volcano plots of differentially expressed genes in LUAD patients versus normal controls. B. Heatmap for differentially expressed genes in LUAD cases versus normal controls. C. Venn diagrams showing the intersection between differentially expressed genes and surface protein-coding genes. LUAD, lung adenocarcinoma.

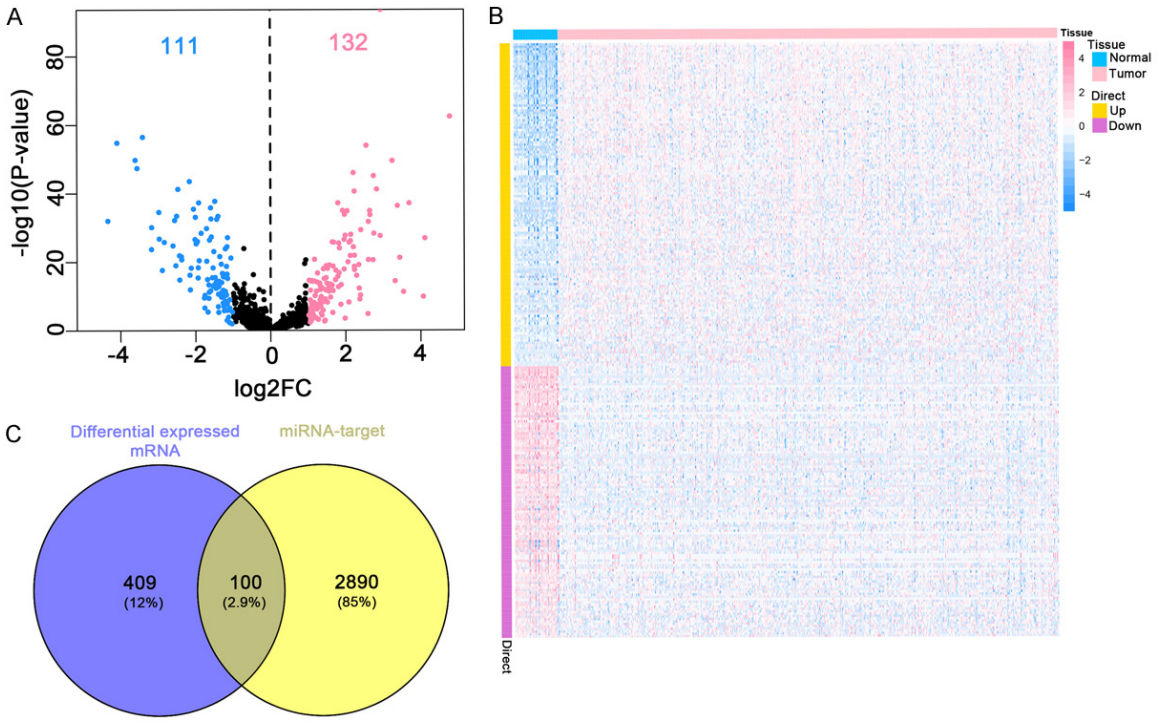


Figure 3. Identification of differentially expressed miRNA-regulated genes. A. Volcano plots of differentially expressed miRNAs in LUAD patients versus normal controls. B. Heatmap for differentially expressed miRNAs in LUAD cases versus normal controls. C. Venn diagrams showing the intersection between differentially expressed surface protein-coding genes and differentially expressed miRNAs. LUAD, lung adenocarcinoma.

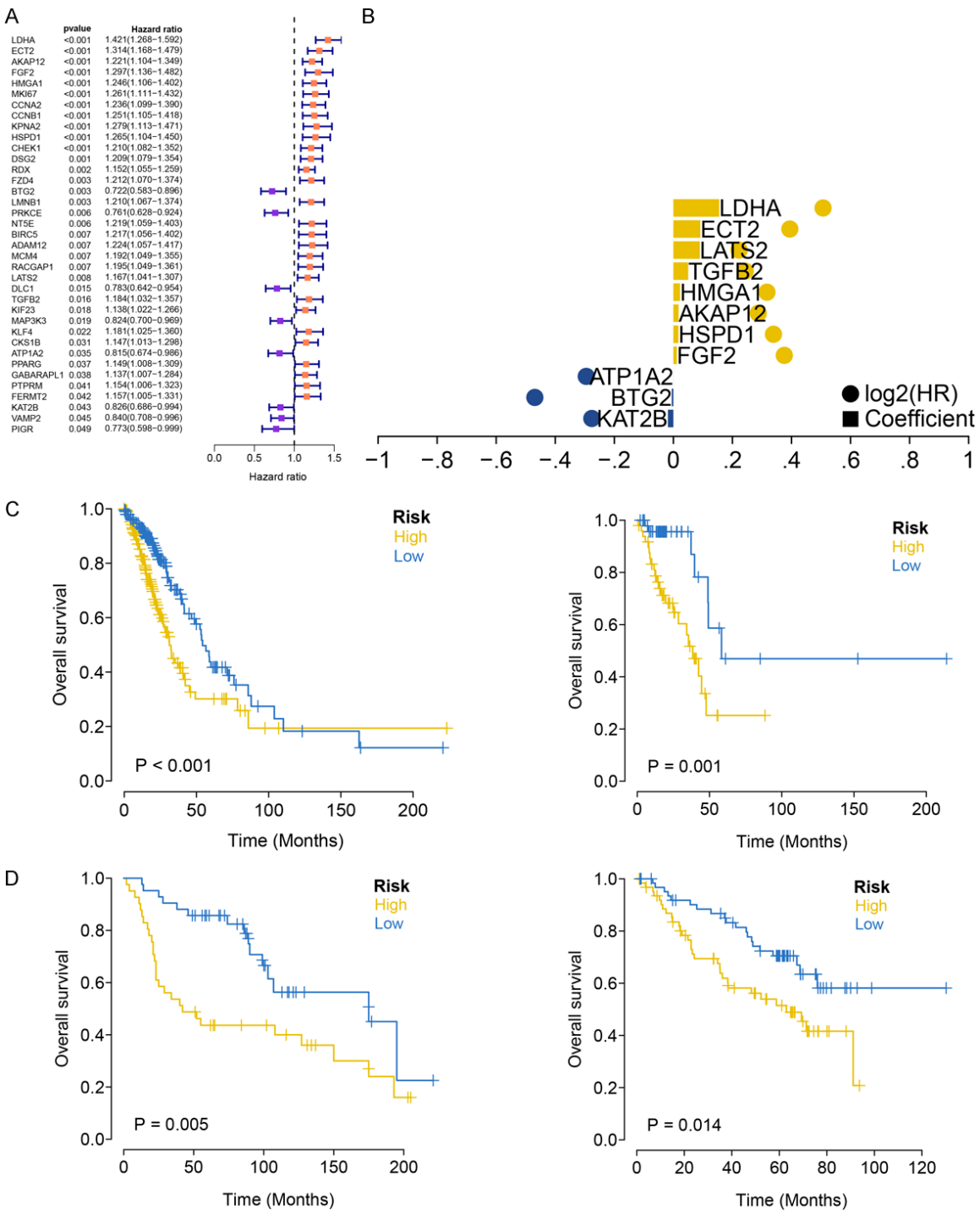


Figure 4. Identification of prognosis-related gene and construction of a prognosis model. A. 37 genes identified by cox-regression that show a significant association with prognosis of LUAD patients. B. 11 key genes selected from 37 prognostic genes, used to construct the risk score formula. C. Association of the risk score with LUAD patients' survival from internal TCGA dataset. D. Association of the risk score with LUAD patients' survival from external GEO dataset. LUAD, lung adenocarcinoma; TCGA, Cancer Genome Atlas; GEO, Gene Expression Omnibus.

sion, which identified 37 genes with significant association to prognosis (P<0.05) (**Figure 4A**). Further narrowing down the key genes, a Least Absolute Shrinkage and Selection Operator (LASSO) regression feature selection algorithm was performed using clinical information from

LUAD patients (**Figure 4B**). Risk score calculation formula: Risk score = KAT2B × (-0.018519793) + BTG2 × (-0.00544217) + ATP1A2 × (-0.000134602) + FGF2 × 0.01179126 + HSPD1 × 0.017023729 + AKAP12 × 0.017567014 + HMGA1 × 0.022685445 + TGFB2 × 0.050171332 + LATS2 × 0.089590696 + ECT2 × 0.091278183 + LDHA × 0.155482721. The patients were randomized into a training set and a validation set in a 4:1 ratio. Then, the cutoff for each sample was calculated by LASSO regression (Figure S3A, S3B), enabling the classification of patients into either the high- or the low-risk group by the median risk score value. In both the training and test sets, Kaplan-Meier curves demonstrated an obviously lower OS in the high-risk group versus the low-risk group (**Figure 4C**). Additionally, the good performance of the model in both the training and validation sets was confirmed by the receiver operating characteristic (ROC) curve analysis (Figure S3C, S3D). To further explore the prediction stability of our RSS, we utilized an external dataset from GEO to analyze the survival difference between patients at low and high risk of LUAD. Similar to the previous findings, high-risk patients showed statistically lower OS than low-risk patients in the GEO external validation set (**Figure 4D**), demonstrating the robustness of the prognostic model constructed.

Independent prognosis factors of LUAD outcomes by the 11-genes RSS

The patients were assigned to high- and low-risk groups according to the median risk score value to discuss the independent risk and prognostic factors of LUAD. Regression analysis was conducted, and the results were presented in nomograms. As indicated by the logistic regression analysis, the risk score contributed significantly to the scoring process of the nomogram prediction model (**Figure 5A**). Moreover, prediction analysis was performed for both one-year and three-year prognosis of LUAD patients (Figure S2), and the consistent prediction results further confirm the reliability of the findings. The tumor-nodes-metastasis (TNM) stage system is critical for accurately diagnosing and staging LUAD, which, in turn, helps determine appropriate treatment options. To evaluate the correlation of the incidence risk with various clinical stage information, we correlated the risk score with several clinical indicators, including the general stage (**Figure 5B**), T (primary tumor) stage (**Figure 5C**), N (regional lymph nodes) stage (**Figure 5D**), and follow-up state (**Figure 5E**), using the rank sum test. The results showed a significant distribution of risk score values for several clinical indicators among the groups (p-value <0.05), which indicates that the risk score obtained from modeling analysis is applicable to the classification of LUAD samples. Overall, these findings suggest that risk score is a useful predictor of outcomes in LUAD patients and can be effectively used in combination with clinical indicators for this purpose.

Expression profiles of model genes in different cell types in the LUAD tumor microenvironment (TME)

To investigate how our model genes influence LUAD patients' survival, we assessed gene expression profiles across different cell types. We conducted cell clustering using the t-distributed stochastic neighbor embedding (t-SNE) algorithm and identified 16 subtypes (**Figure 6A**). These subtypes were further classified with the R package SingleR into six distinct cell categories, including epithelial cells, endothelial cells, T cells, and tissue stem cells, as well as macrophages and monocytes (**Figure 6B**). The expression patterns of the 11 model genes across these 6 cell types were depicted (**Figure 6C**, **6D**). Interestingly, high-risk patients were found to have a higher proportion of epithelial cells (**Figure 6E**, **6F**), suggesting that our model genes facilitate epithelial cell-related processes that contribute to LUAD progression and poor patient prognosis. To further explore this possibility, we investigated the communication patterns among different cell subtypes and found that epithelial cells and endothelial cells were the centers of cell communication (**Figure 6G**, **6H**). This indicates that these model genes play a key role in the cellular communication that contributes to LUAD progression. To validate our findings, we gathered histological staining data for some of our model genes from the open-access Human Protein Atlas database (**Figure 7**, available at https: //www.proteinatlas.org/) and the results were consistent with our expectations. Furthermore, clinical samples were collected to test the genes associated with our predictive model (**Figure 8**).

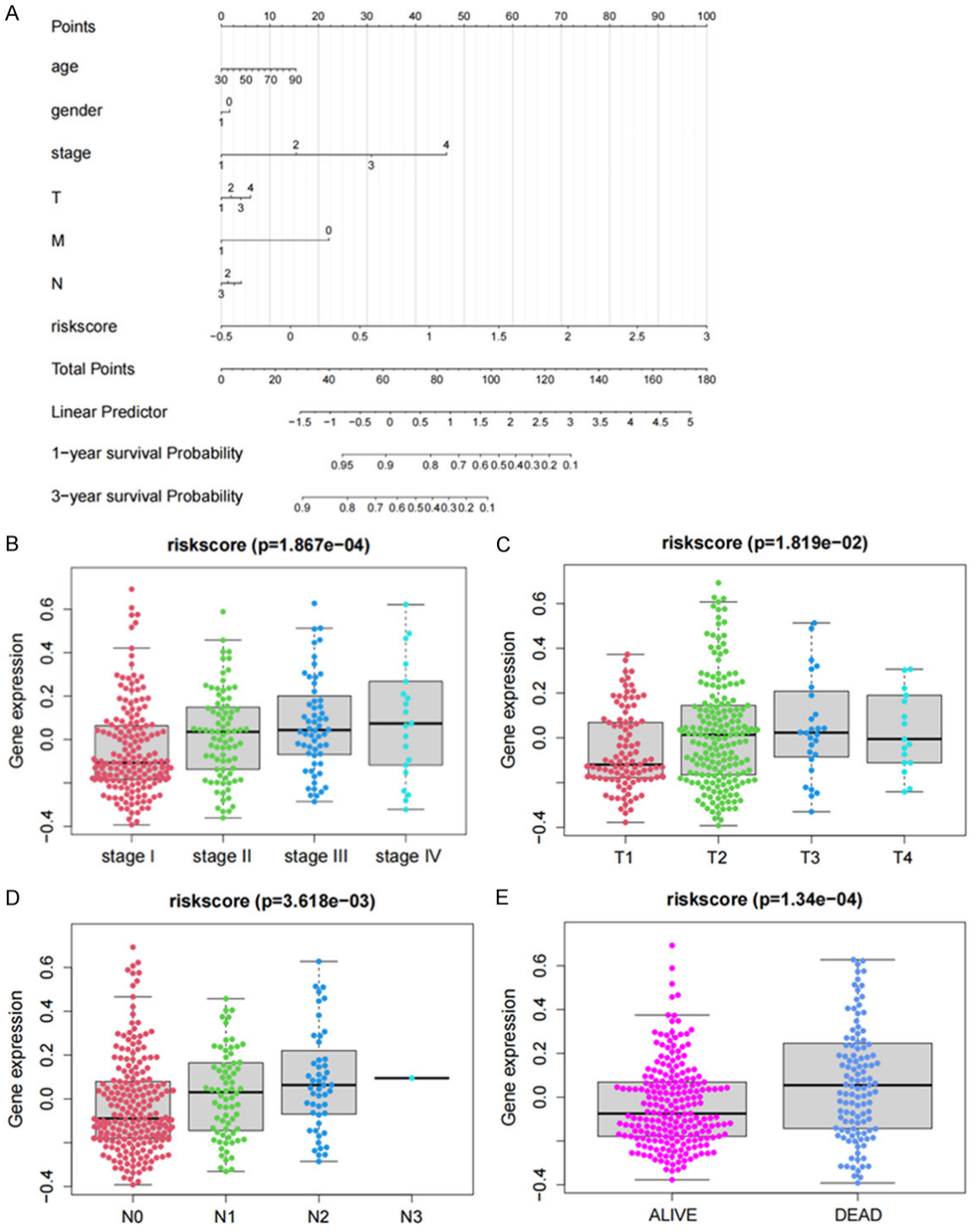
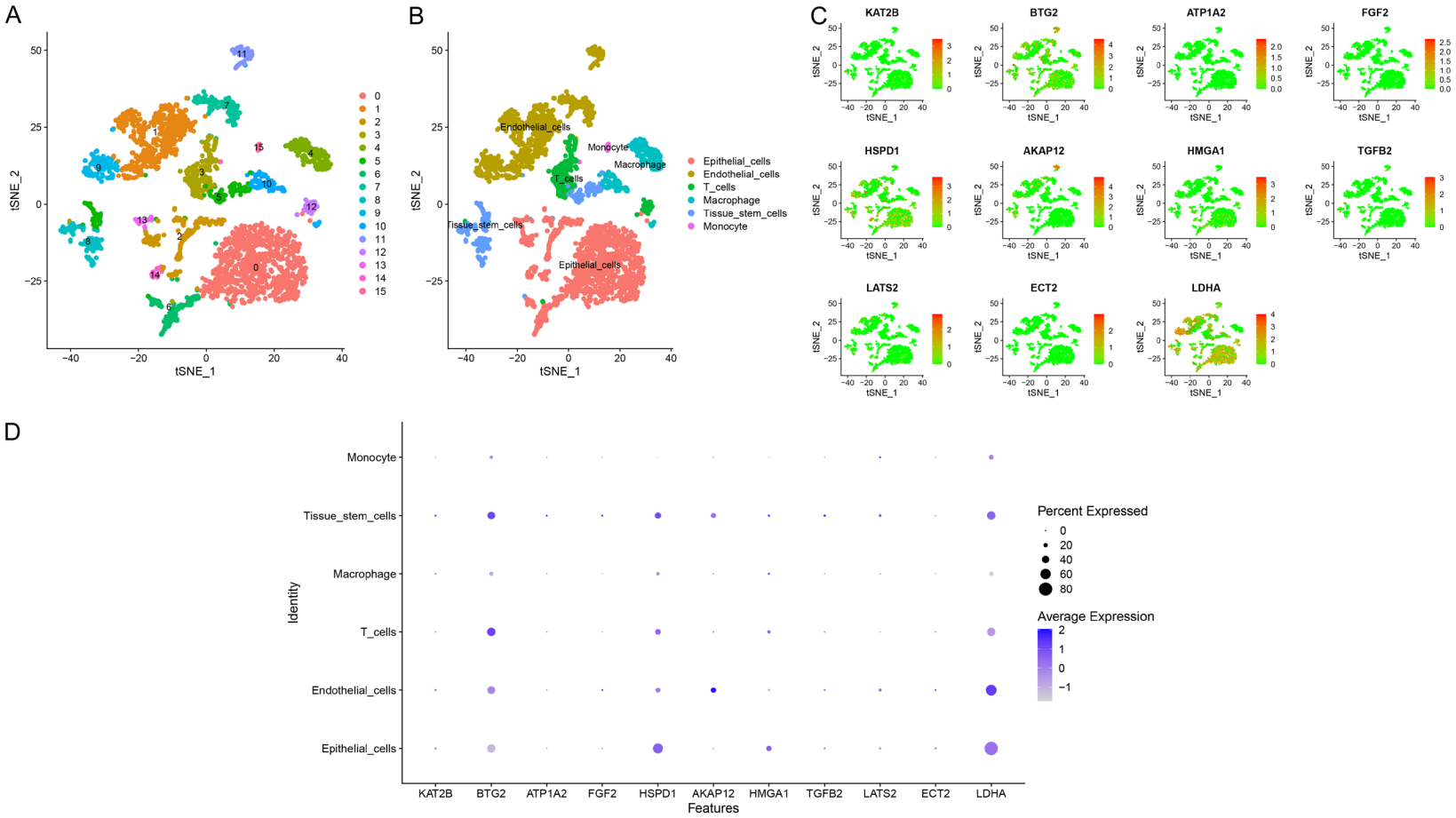


Figure 5. Risk score system and its correlation with various clinicopathological factors. A. Nomogram model for predicting LUAD patients' 1- and 3-year survival. B-E. Correlations between the risk score and general stage, T (primary tumor) stage, N (regional lymph nodes) stage, and follow-up state. LUAD, lung adenocarcinoma.

The histological data corroborated our findings, further confirming the validity and significance of our research. Collectively, these results shed insights into the role of our model genes in LUAD and their impact on cell types associated with disease progression.


A
B
C
D
tSNE_1
tSNE_2
Epithelial_cells
Endothelial_cells
T_cells
Macrophage
Tissue_stem_cells
Monocyte
KAT2B
BTG2
ATP1A2
FGF2
HSPD1
AKAP12
HMGA1
TGFB2
LATS2
ECT2
LDHA
Identity
Features
Percent Expressed
Average Expression

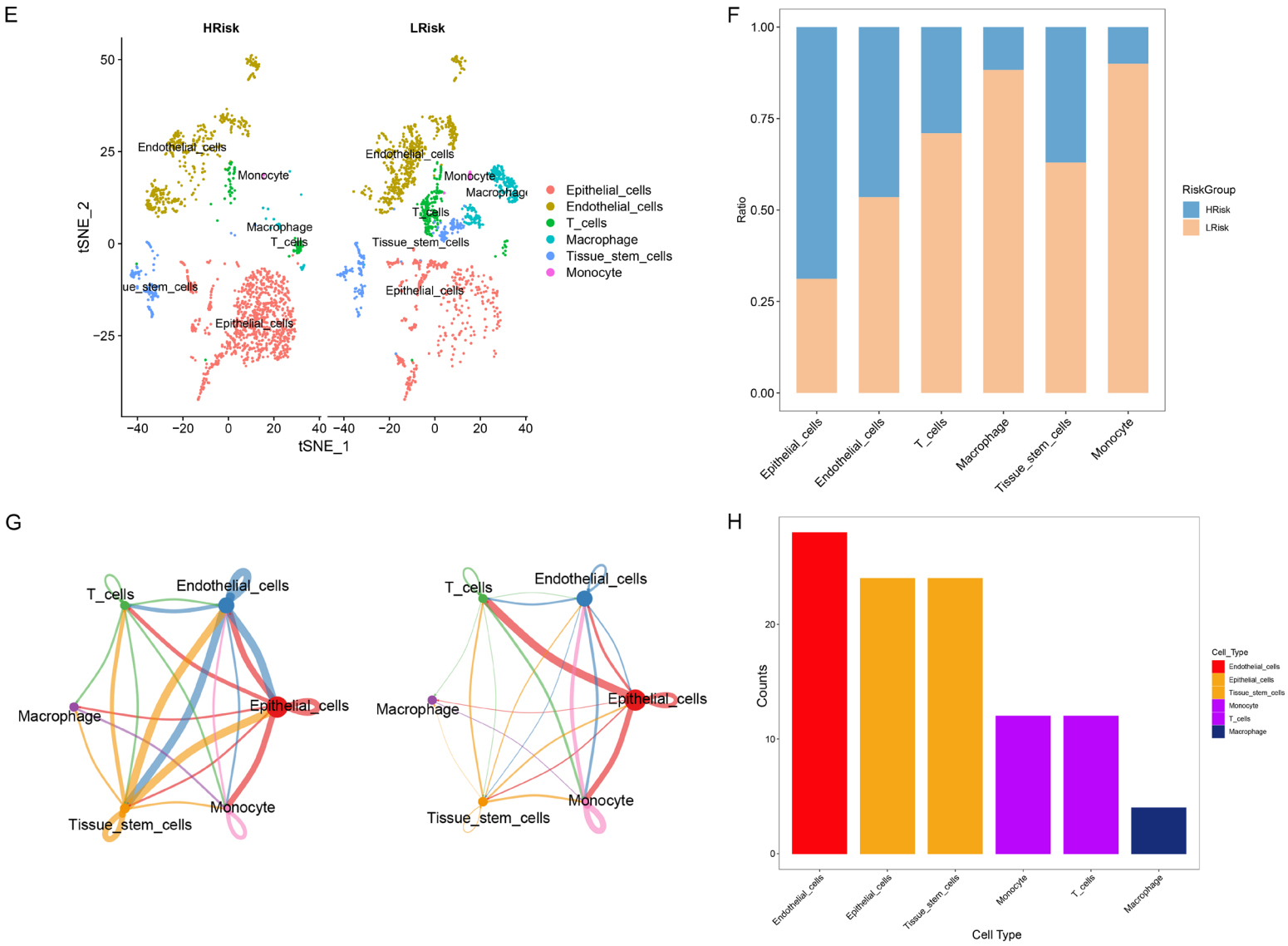


Figure 6. Expression profiles of model genes in the LUAD tumor microenvironment. A. t-SNE algorithm classification of all DEGs into 16 clusters. B. Six cell types were obtained by SingleR package annotation of cell types for all cell subpopulations. C, D. Expression of 11 models in those 6 cell types. E, F. Comparison of cell composition in the tumor microenvironment between patients at high risk and those at low risk. G, H. Communication patterns between different cell subtypes. LUAD, lung adenocarcinoma.

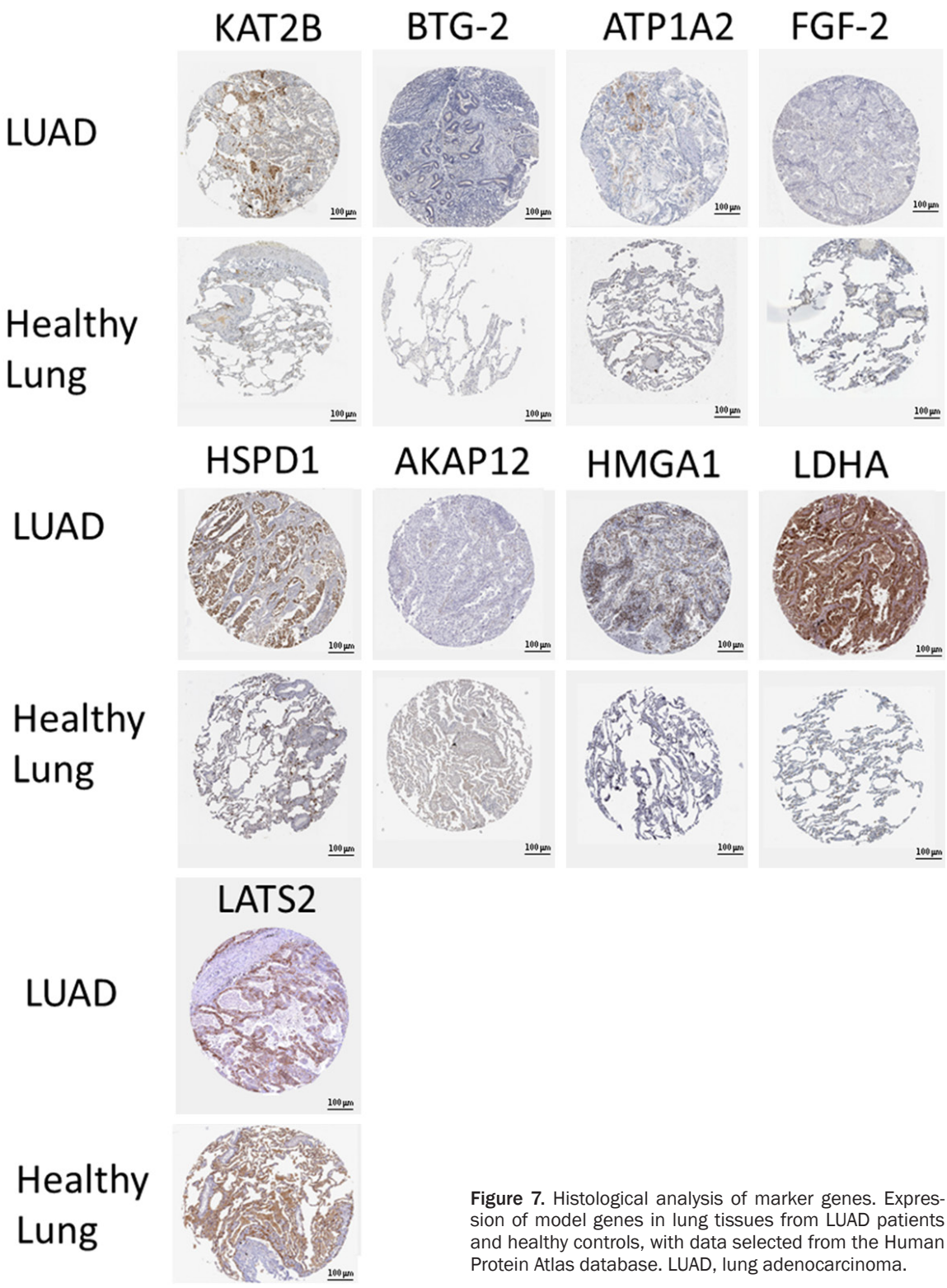


Figure 7. Histological analysis of marker genes. Expression of model genes in lung tissues from LUAD patients and healthy controls, with data selected from the Human Protein Atlas database. LUAD, lung adenocarcinoma.

Core signal pathway analysis

To comprehensively investigate the molecular mechanisms underlying the correlation of risk scores with tumor progression, we studied specific signal pathways. Patients were further assigned to high- and low-risk groups based on their risk level for GSEA analysis of the inter-

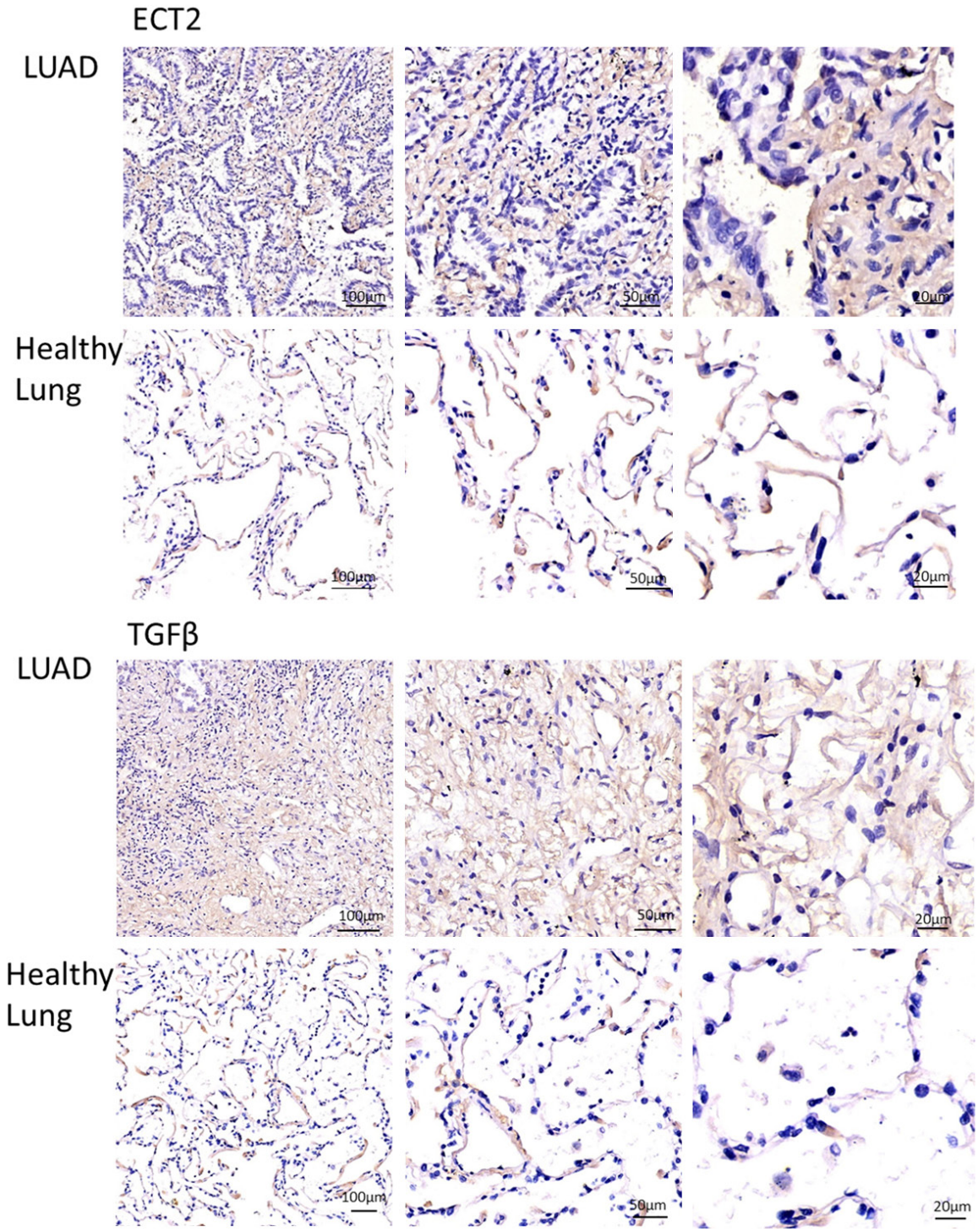


Figure 8. Histological analysis of marker genes. Expression of model genes in lung tissues from LUAD patients and healthy controls. LUAD, lung adenocarcinoma.

group differences in signal pathways. Interestingly, we found that central carbon metabolism in cancer, HIF-1 signaling, and p53 signaling pathway were among the pathways (**Figure 9A**). To gain a deeper understanding of these pathways, we mapped out the molecular interaction network for each pathway (**Figure 9B**). To investigate the genetic landscape of high- and

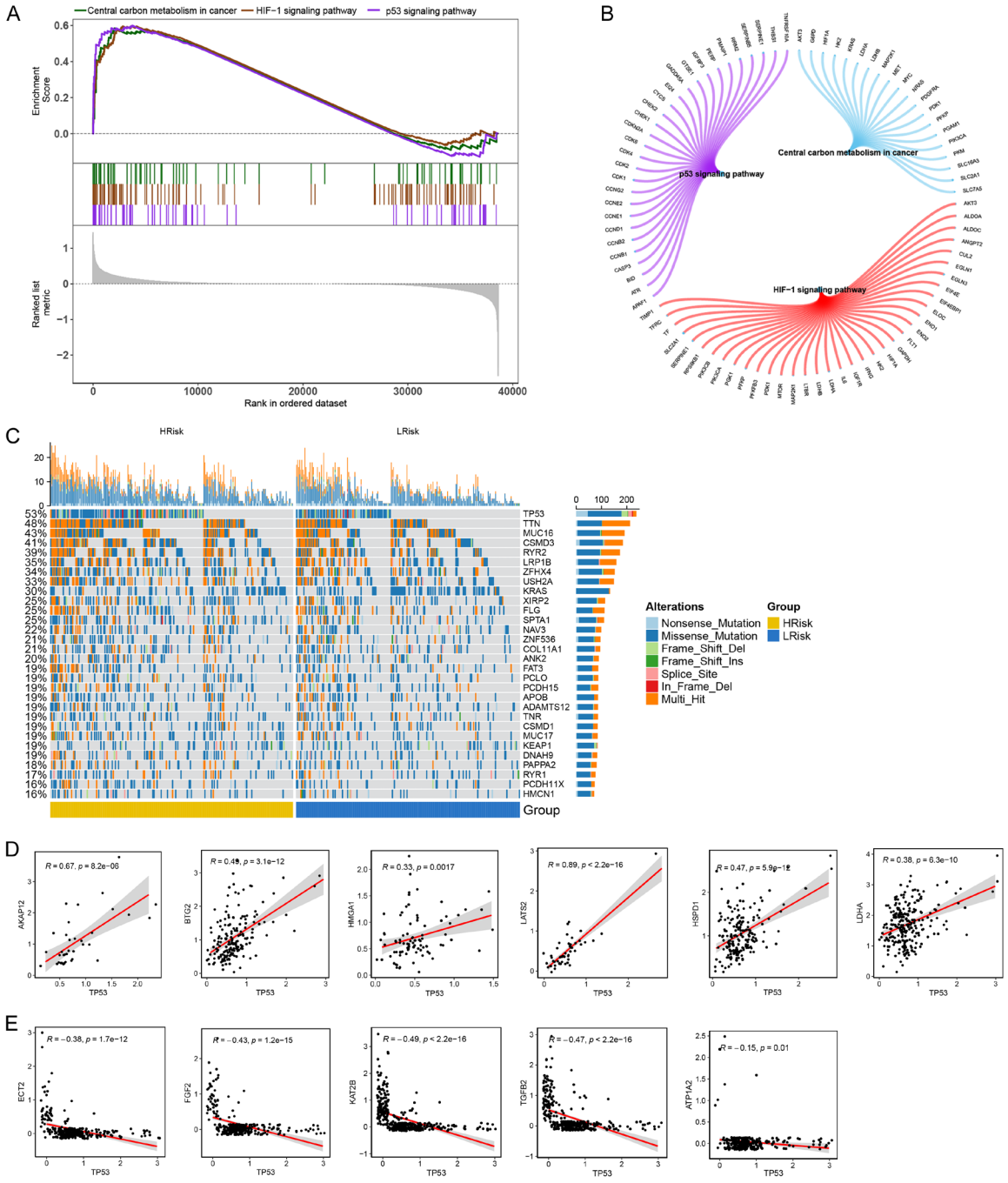


Figure 9. Core signal pathway analysis. A. Gene set enrichment analysis of high- and low-risk patients. B. Molecular interactions between those signal pathways. C. Mutation profiles of high- and low-risk patients. D. Model genes that are positively correlated with TP53. E. Model genes that are negatively correlated with TP53.

low-risk patients, we examined the mutation profiles. Strikingly, the high-risk group showed an evidently higher proportion of TP53 and other gene mutations than the low-risk group (**Figure 9C**), highlighting the potential role of TP53 in predicting high-risk status and tumor progression. Following these findings, we visualized the co-expression of the model genes and the tumor progression gene TP53 across six cell marker genes (**Figure 9D**, **9E**). This analysis provides insight into the potential interaction between the model gene and TP53 in different cell types, shedding light on possible mechanisms underlying their cooperative

effect in driving tumor progression. Thus, our comprehensive analysis of the signaling pathways, immunotherapy-related tumor markers, mutation profiles, and co-expression patterns reinforces the significance of our risk model in understanding LUAD progression. These findings provide a basis for further exploration of the mechanisms and potential therapeutic targets associated with these model genes and their interactions with TP53 in various cell types.

The association of the prognosis model with the immune microenvironment

We further investigated the correlation of the risk score with common immunotherapy-related tumor markers to gain insights into the immunogenic profiles of the high- and low-risk patients. The two groups are statistically different in TMB (**Figure 10A**) and neoantigens (**Figure 10B**). Specifically, the high-risk group showed higher levels of TMB and neoantigens. These observations suggest the potential of the risk score as a predictor of immunotherapy response. To further explore this possibility, we examined the contents of immune cells. The T cells focal helper, regulatory T cells (Tregs), natural killer (NK) cells activated dendritic cells resting, and mast cells resting were found to be obviously reduced in high-risk patients than in low-risk patients. Conversely, high-risk patients had obvious increases in activated CD4 memory T cells and M0 and M1 macrophages than low-risk patients (**Figure 10C**). In addition, the investigation of the association of the risk score with immune cell contents revealed a positive connection between the risk score and activated CD4 memory T cells and resting NK cells but an inverse relationship with resting mast cells and Tregs (**Figure 10D**). Next, we conducted a detailed analysis of immune regulatory genes by examining differences in the expression profiles of immune-associated chemokines, immunosuppressors, immunostimulating factors, and immunoreceptors between the two groups (Figure S4A-D). These findings expand our understanding of the immune landscape associated with high- and low-risk groups and provide greater insights into the correlation of risk scores with immunotherapy response. Moreover, to determine the sensitivity of high- and low-risk patients to anti-tumor immunotherapy, we predicted their response to immunotherapy, and the findings depicted poorer response in high-risk patients (**Figure 10E**). Based on our observed immune cell contents and immune-related gene expression patterns, the poor response to immunotherapy in high-risk patients may be attributed to an increase in M1 macrophages, which promotes tumor development. All in all, the above findings provide a rounded analysis of the immune landscape associated with risk groups and offer potential avenues for developing personalized immunotherapeutic approaches to treat LUAD patients based on risk score.

Discussion

LUAD, the most prevalent type of lung cancer, is associated with an unsatisfactory prognosis despite advances in clinical management. A recent domain of investigation in LUAD prognosis is the role of miRNAs in regulating gene expression. miRNAs, as a class of endogenous non-coding RNAs, play dynamic roles in the modulation of genetic expression at the post-transcriptional level. A recent review summarized a close link between differentially expressed miRNAs to LUAD cell invasion, metastasis, and malignant behavior [14]. Due to their important regulatory roles in these processes, miRNAs represent important biomarkers for assessing the severity and prognosis of LUAD. Additionally, miRNAs may serve as crucial therapeutic targets in designing novel therapies for LUAD treatment.

Novel miRNA-associated target genes have been identified for prognostic prediction and therapeutic decision-making in LUAD. For example, Wei et al. [22] identified eight miRNA-associated target genes and constructed a prognostic model, revealing that high-risk patients had poorer prognoses, lower immune scores, and a lower response to immunotherapy. However, surface-protein gene signatures were not reported. To harness this potential, we conducted an extensive analysis of bulk RNA-seq and miRNA-seq data. We identified a range of overlapping, miRNA-regulated genes, which were subsequently subjected to Cox regression and LASSO regression processes, to pinpoint gene that is closely associated with patient prognosis. A highly selective 11 gene-based risk score formula was then established: KAT2B × (-0.018519793) + BTG2 ×

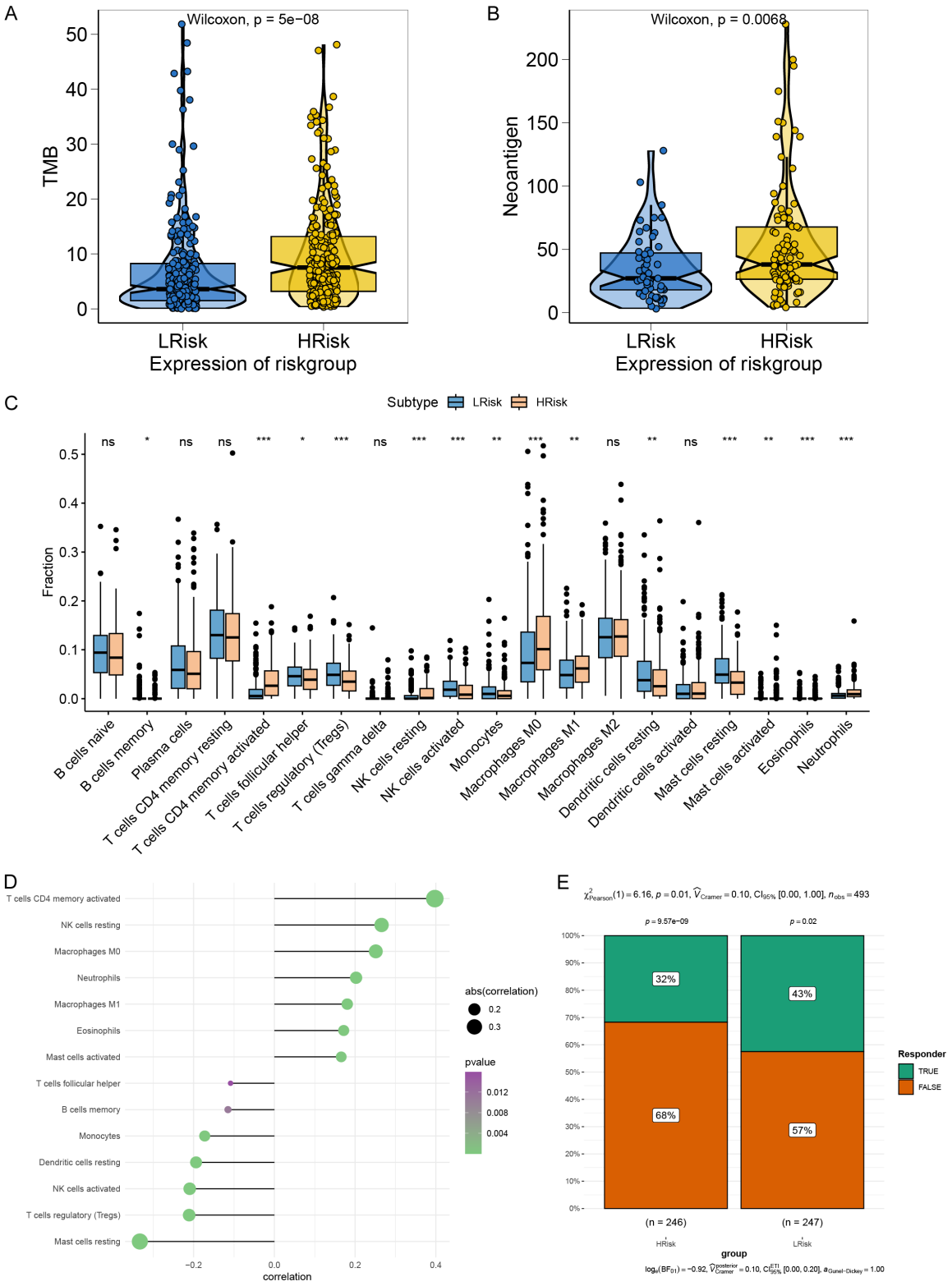


Figure 10. The correlation of the prognosis model with the immune microenvironment. A, B. Comparison of tumor mutation burden (TMB) and neoantigens between two groups. C. Comparison of immune cell components between high- and low-risk patients. D. Correlation of the risk score with immune cell content. E. Predicted response to immunotherapy between high- and low-risk patients.

(-0.00544217) + ATP1A2 × (-0.000134602) + FGF2 × 0.01179126 + HSPD1 × 0.017023729 + AKAP12 × 0.017567014 + HMGA1 × 0.022685445 + TGFB2 × 0.050171332 + LATS2 × 0.089590696 + ECT2 × 0.091278183 + LDHA × 0.155482721. These genes are miRNA-regulated and surface-protein encoded, enhancing their potential for future clinical applications as easily detectable surface markers. This marks a significant advancement in understanding the intricate role of miRNAs in LUAD and their regulatory influence on mRNAs.

The 11 gene-based risk score formula developed in this study consists of miRNA-regulated genes, each of which plays a specific role in LUAD pathogenesis. For instance, the first two genes, KAT2B and ATP1A2, are essential in maintaining genome integrity and regulating cell proliferation, growth, and differentiation [23, 24]. Meanwhile, KAT2B has been associated with higher immune filtration and more satisfactory efficacy of immunotherapy in NSCLC [25]. Additionally, BTG2 is an anti-oncogene that is usually downregulated in several cancer types, including LUAD [26]. Its silencing in various cancer cell lines has been shown to enhance cellular migration, resulting in the promotion of tumor growth [27, 28]. In addition to these three protective genes, the remaining eight risk-adding genes also contribute significantly to cancer biology. HMGA1, another gene included in the 11-gene panel, promotes cell proliferation and presents elevated levels in cancers characterize by increased tumor malignancy and enhanced invasion [29-31]. TGFB2 plays a crucial role in regulating the response to cellular stress, including inflammation and immune regulation [32]. It has also been suggested to have a dual function in cancer development, possessing both tumor suppressive and oncogenic effects [33]. FGF2 is essential for the growth and differentiation of multiple cell types in the body, including cells in the lung parenchyma and vasculature [34]. Recently, this gene has been identified as a marker for drug resistance in LUAD [35].

Altered glucose metabolism in cancers has long been recognized, which is reflected in the increased activity of lactate dehydrogenase A (LDHA) in tumors [36]. LDHA catalyzes the final step of glycolysis, producing lactate and ultimately promoting tumor growth. Therefore, its upregulation is indicative of a higher risk in LUAD patients [37]. On the other hand, LATS2, ECT2, and AKAP12 are involved in cell cycle regulation and checkpoint control. LATS2, known as an anti-oncogene, regulates the Hippo signaling pathway to control cellular proliferation and differentiation [38]. Previous research suggests that AKAP12 influences cell migration and invasion in multiple cancer types, including ovarian, breast, and prostate cancers, and its downregulation is associated with aggressive tumor behavior in LUAD [39]. ECT2 is a crucial regulator of cytokinesis and cell division and has been linked to tumor progression, migration, and invasion [40]. Meanwhile, high expression of ECT2 is considered as an independent prognostic factor for poor OS and recurrence in LUAD [41].

Internal datasets from TCGA and external datasets from the GEO database were utilized to validate the prognostic significance of this RSS. The highly specific and sensitive results derived from both datasets bolster the credibility of this RSS. Further analysis using single-cell RNA sequencing facilitated a deeper exploration of the differences between high- and low-risk patients in cell sub-population inside the TME, leading to a more in-depth understanding of how our model genes express themselves differently across various cell subtypes. It was concluded that the largest inter-group differences lay in the endothelial and epithelial cells. We also conducted an inter-group analysis of gene mutations and identified the TP53 signaling pathway as a key factor. A great deal of effort and attention has been devoted to exploring the complex mechanisms of LUAD, but the current understanding, especially in terms of the TME, therapeutic targets, and prognostic factors, remains far from being satisfactory. In this study, we first constructed the LUAD TME using gene expression profiling and further investigated its immune infiltration landscape. As expected, some immune cells showed significant differences between the hyper- and hypo-immune groups. The TME is often defined as the environment surrounding the tumor, including the extracellular matrix, vasculature, and cellular players such as immune cells and neurons, all of which are strongly associated with tumor progression and treatment outcomes [42]. A growing number of studies have experimentally elucidated the facilitating role of

TME infiltration in immunotherapeutic responses and drug resistance in different types of tumors and explored its impact on patient prognosis [43-45]. In another important observation, it was deduced that the risk score was positively associated with Treg cells and inversely related to anti-tumor M1 macrophages, indicating an immunosuppressive environment in high-risk patients. Subsequently, the immunotherapy prediction analyses also revealed a lower response in high-risk patients.

Although this study provides a comprehensive view of LUAD and establishes a powerful model for prognostic prediction, there are still two major drawbacks that require further investigation. First, we were limited to data solely from TCGA and GEO portals and could not cover other data sources. This hindered us from testing the robustness of the model when used for other data. Second, the risk signatures were established using retrospective data from public databases. Therefore, prospective and multi-center LUAD cohorts are required to eliminate bias. Third, we only validated the protein levels of genes in clinical samples. Thus, we will subsequently collect clinical data for further testing of the prognostic model. Collectively, the 11 genes identified in this study play a vital role in modulating cellular activities (e.g., cell proliferation, differentiation, migration, and survival). Understanding the specific role of these genes and how they interact with miRNAs have significant implications for the development of targeted therapies for LUAD. Moreover, their surface expression makes them ideal candidates for potential therapeutic targets and diagnoses. Therefore, the risk score formula developed in this study sheds valuable light on the potential mechanisms of LUAD and can be considered a significant contribution to the field. Our novel RSS shows tremendous potential in predicting clinical outcomes in LUAD patients while laying a strong foundation for further investigation into the regulatory function of miRNAs in LUAD.

Disclosure of conflict of interest

None.

Address correspondence to: Chendan Zou, Department of Biochemistry and Molecular Biology, Harbin Medical University, Harbin 150000, Heilongjiang, China. Tel: +86-18604506752; E-mail: zouchendan1114@126.com

References

[1] O'Brien TD, Jia P, Aldrich MC and Zhao Z. Lung cancer: one disease or many. Hum Hered 2018; 83: 65-70.
[2] Song Y, Kelava L and Kiss I. MiRNAs in lung adenocarcinoma: role, diagnosis, prognosis, and therapy. Int J Mol Sci 2023; 24: 13302.
[3] Huang CY, Chen BH, Chou WC, Yang CT and Chang JW. Factors associated with the prognosis and long-term survival of patients with metastatic lung adenocarcinoma: a retrospective analysis. J Thorac Dis 2018; 10: 2070-2078.
[4] Smith JC and Sheltzer JM. Genome-wide identification and analysis of prognostic features in human cancers. Cell Rep 2022; 38: 110569.
[5] Wang Z, Zhang J, Shi S, Ma H, Wang D, Zuo C, Zhang Q and Lian C. Predicting lung adenocarcinoma prognosis, immune escape, and pharmacomic profile from arginine and proline-related genes. Sci Rep 2023; 13: 15198.
[6] Ren Q, Li Q, Shao C, Zhang P, Hu Z, Li J, Wang W and Yu Y. Establishing a prognostic model based on immune-related genes and identification of BIRC5 as a potential biomarker for lung adenocarcinoma patients. BMC Cancer 2023; 23: 897.
[7] He J, Li W, Li Y and Liu G. Construction of a prognostic model for lung adenocarcinoma based on bioinformatics analysis of metabolic genes. Transl Cancer Res 2020; 9: 3518-3538.
[8] Chen Q, Chen S, Wang J, Zhao Y, Ye X, Fu Y and Liu Y. Construction and validation of a hypoxia-related risk signature identified EXO1 as a prognostic biomarker based on 12 genes in lung adenocarcinoma. Aging (Albany NY) 2023; 15: 2293-2307.
[9] Ma B, Geng Y, Meng F, Yan G and Song F. Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method. J Cancer 2020; 11: 1288-1298.
[10] Liu L, He H, Peng Y, Yang Z and Gao S. A four-gene prognostic signature for predicting the overall survival of patients with lung adenocarcinoma. PeerJ 2021; 9: e11911.
[11] O'Brien J, Hayder H, Zayed Y and Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. Front Endocrinol (Lausanne) 2018; 9: 402.
[12] Peng Y and Croce CM. The role of microRNAs in human cancer. Signal Transduct Target Ther 2016; 1: 15004.
[13] Ali Syeda Z, Langden SSS, Munkhzul C, Lee M and Song SJ. Regulatory mechanism of microRNA expression in cancer. Int J Mol Sci 2020; 21: 1723.
[14] Liu J, Zhang F, Wang J and Wang Y. MicroRNA-mediated regulation in lung adenocarcinoma:

signaling pathways and potential therapeutic implications (Review). Oncol Rep 2023; 50: 211.
[15] Du J, Qian J, Zheng B, Xu G, Chen H and Chen C. miR-21-5p is a biomarker for predicting prognosis of lung adenocarcinoma by regulating PIK3R1 expression. Int J Gen Med 2021; 14: 8873-8880.
[16] Pop-Bica C, Pintea S, Magdo L, Cojocneanu R, Gulei D, Ferracin M and Berindan-Neagoe I. The clinical utility of miR-21 and let-7 in non-small cell lung cancer (NSCLC). A systematic review and meta-analysis. Front Oncol 2020; 10: 516850.
[17] Kunz M, Göttlich C, Walles T, Nietzer S, Dandekar G and Dandekar T. MicroRNA-21 versus microRNA-34: lung cancer promoting and inhibitory microRNAs analysed in silico and in vitro and their clinical impact. Tumour Biol 2017; 39: 1010428317706430.
[18] Malik S, Zafar Paracha R, Khalid M, Nisar M, Siddiqa A, Hussain Z, Nawaz R, Ali A and Ahmad J. MicroRNAs and their target mRNAs as potential biomarkers among smokers and non-smokers with lung adenocarcinoma. IET Syst Biol 2019; 13: 69-76.
[19] Wang XJ, Gao J, Wang Z and Yu Q. Identification of a potentially functional microRNA-mRNA regulatory network in lung adenocarcinoma using a bioinformatics analysis. Front Cell Dev Biol 2021; 9: 641840.
[20] Zhou B and Gao S. Construction and validation of a novel immune and tumor mutation burden-based prognostic model in lung adenocarcinoma. Cancer Immunol Immunother 2022; 71: 1183-1197.
[21] Castillo-Peña A and Molina-Pinelo S. Landscape of tumor and immune system cells-derived exosomes in lung cancer: mediators of antitumor immunity regulation. Front Immunol 2023; 14: 1279495.
[22] Wei Y, Zhong W, Bi Y, Liu X, Zhou Q, Liu J, Wang M, Zhang H and Chen M. Molecular subtypes and prognostic models for predicting prognosis of lung adenocarcinoma based on miRNA-related genes. Curr Med Chem 2023; [Epub ahead of print].
[23] Walters BW, Tan TJ, Tan CT, Dube CT, Lee KT, Koh J, Ong YHB, Tan VXH, Jahan FRS, Lim XN, Wan Y and Lim CY. Divergent functions of histone acetyltransferases KAT2A and KAT2B in keratinocyte self-renewal and differentiation. J Cell Sci 2023; 136: jcs260723.
[24] Zhang B, Zhu Z, Zhang X, Li F and Ding A. Inhibition of the proliferation, invasion, migration, and epithelial-mesenchymal transition of prostate cancer cells through the action of ATP1A2 on the TGF-β/Smad pathway. Transl Androl Urol 2022; 11: 53-66.
[25] Zhou X, Wang N, Zhang Y, Yu H and Wu Q. KAT2B is an immune infiltration-associated biomarker predicting prognosis and response to immunotherapy in non-small cell lung cancer. Invest New Drugs 2022; 40: 43-57.
[26] Zhang XZ, Chen MJ, Fan PM, Jiang W and Liang SX. BTG2 serves as a potential prognostic marker and correlates with immune infiltration in lung adenocarcinoma. Int J Gen Med 2022; 15: 2727-2745.
[27] Li YJ, Dong BK, Fan M and Jiang WX. BTG2 inhibits the proliferation and metastasis of osteosarcoma cells by suppressing the PI3K/AKT pathway. Int J Clin Exp Pathol 2015; 8: 12410-12418.
[28] Mao B, Zhang Z and Wang G. BTG2: a rising star of tumor suppressors (Review). Int J Oncol 2015; 46: 459-464.
[29] Wang Y, Hu L, Zheng Y and Guo L. HMGA1 in cancer: cancer classification by location. J Cell Mol Med 2019; 23: 2293-2302.
[30] Zhong J, Liu C, Zhang QH, Chen L, Shen YY, Chen YJ, Zeng X, Zu XY and Cao RX. TGF-β1 induces HMGA1 expression: the role of HMGA1 in thyroid cancer proliferation and invasion. Int J Oncol 2017; 50: 1567-1578.
[31] Pang B, Fan H, Zhang IY, Liu B, Feng B, Meng L, Zhang R, Sadeghi S, Guo H and Pang Q. HMGA1 expression in human gliomas and its correlation with tumor proliferation, invasion and angiogenesis. J Neurooncol 2012; 106: 543-549.
[32] Massagué J and Sheppard D. TGF-β signaling in health and disease. Cell 2023; 186: 4007-4037.
[33] Principe DR, Doll JA, Bauer J, Jung B, Munshi HG, Bartholin L, Pasche B, Lee C and Grippo PJ. TGF-β: duality of function between tumor prevention and carcinogenesis. J Natl Cancer Inst 2014; 106: djt369.
[34] Danopoulos S, Shiosaki J and Al Alam D. FGF signaling in lung development and disease: human versus mouse. Front Genet 2019; 10: 170.
[35] Jiang H, Li C, Gong Q and Qie H. Identification and validation of basic fibroblast growth factor as a prognostic biomarker for the response of lung adenocarcinoma patients to bevacizumab treatment. Immunobiology 2023; 228: 152764.
[36] Feng Y, Xiong Y, Qiao T, Li X, Jia L and Han Y. Lactate dehydrogenase A: a key player in carcinogenesis and potential target in cancer therapy. Cancer Med 2018; 7: 6124-6136.
[37] Hou XM, Yuan SQ, Zhao D, Liu XJ and Wu XA. LDH-A promotes malignant behavior via activation of epithelial-to-mesenchymal transition in lung adenocarcinoma. Biosci Rep 2019; 39: BSR20181476.

[38] Luo SY, Kwok HH, Yang PC, Ip MS, Minna JD and Lam DC. Expression of large tumour suppressor (LATS) kinases modulates chemotherapy response in advanced non-small cell lung cancer. Transl Lung Cancer Res 2020; 9: 294-305.
[39] Chang J, Liu S, Li B, Huo Z, Wang X and Zhang H. MiR-338-3p improved lung adenocarcinoma by AKAP12 suppression. Arch Med Sci 2019; 17: 462-473.
[40] Fields AP and Justilien V. The guanine nucleotide exchange factor (GEF) Ect2 is an oncogene in human cancer. Adv Enzyme Regul 2010; 50: 190-200.
[41] Zhou S, Wang P, Su X, Chen J, Chen H, Yang H, Fang A, Xie L, Yao Y and Yang J. Correction: high ECT2 expression is an independent prognostic factor for poor overall survival and recurrence-free survival in non-small cell lung adenocarcinoma. PLoS One 2018; 13: e0196354.
[42] Zhang Y, Yang M, Ng DM, Haleem M, Yi T, Hu S, Zhu H, Zhao G and Liao Q. Multi-omics data analyses construct TME and identify the immune-related prognosis signatures in human LUAD. Mol Ther Nucleic Acids 2020; 21: 860-873.
[43] de Visser KE and Joyce JA. The evolving tumor microenvironment: from cancer initiation to metastatic outgrowth. Cancer Cell 2023; 41: 374-403.
[44] Kang W, Qiu X, Luo Y, Luo J, Liu Y, Xi J, Li X and Yang Z. Application of radiomics-based multiomics combinations in the tumor microenvironment and cancer prognosis. J Transl Med 2023; 21: 598.
[45] Naser R, Fakhoury I, El-Fouani A, Abi-Habib R and El-Sibai M. Role of the tumor microenvironment in cancer hallmarks and targeted therapy (Review). Int J Oncol 2023; 62: 23.

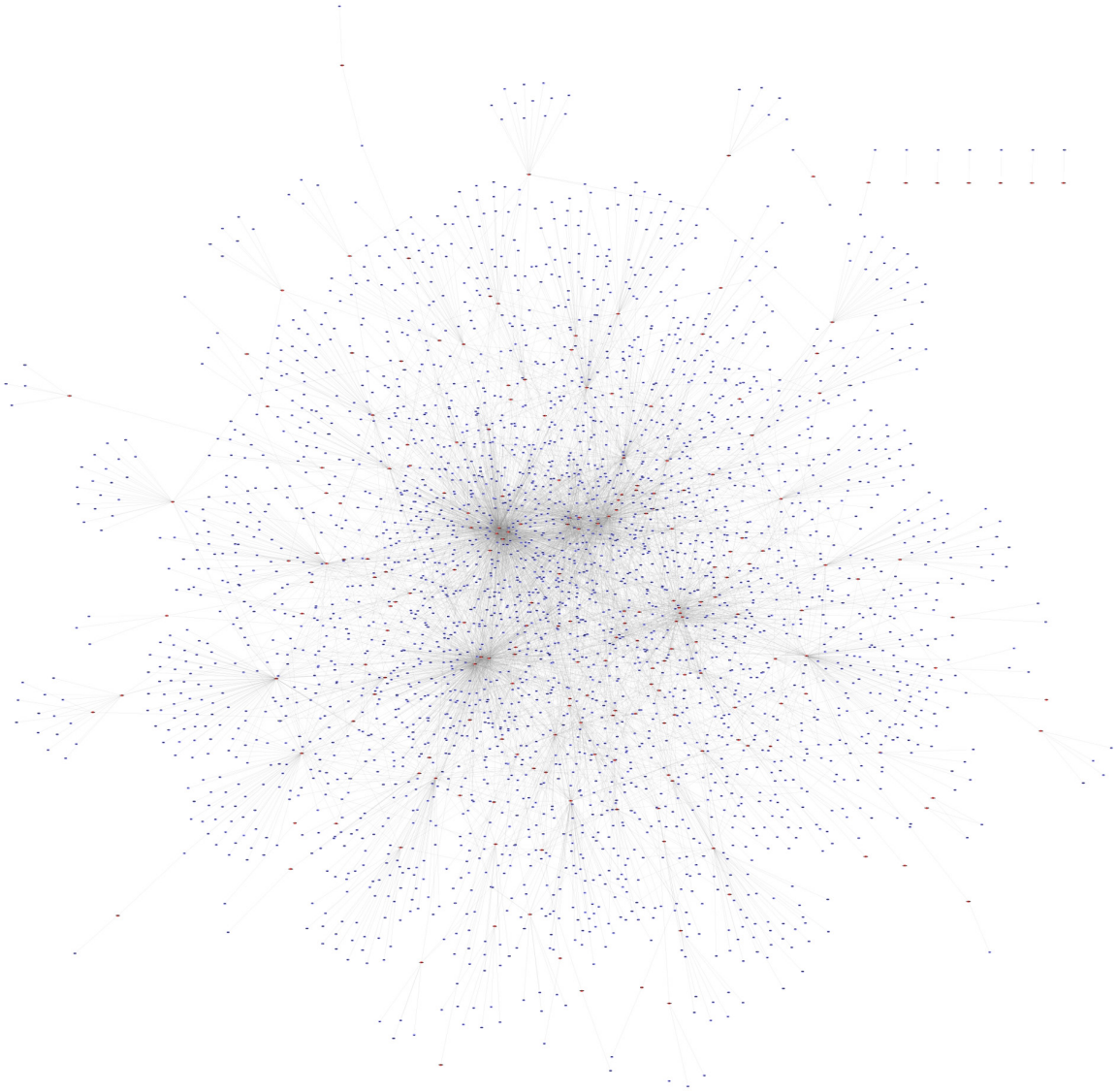


Figure S1. 2990 prediceted mRNAs visualized using cytoscape.

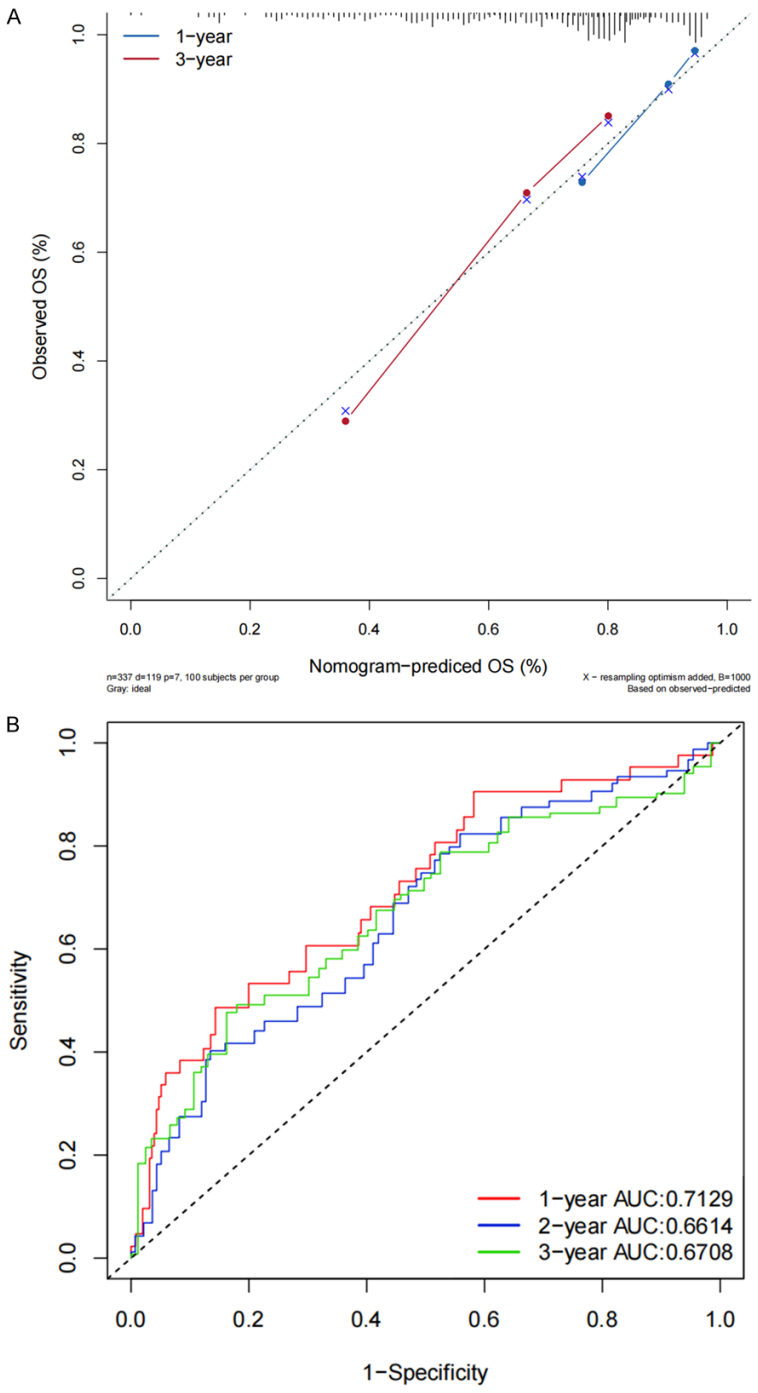


Figure S2. Prediction analysis on the one-year and three-year periods of lung adenocarcinoma. A: Calibration curve; B: ROC curve.

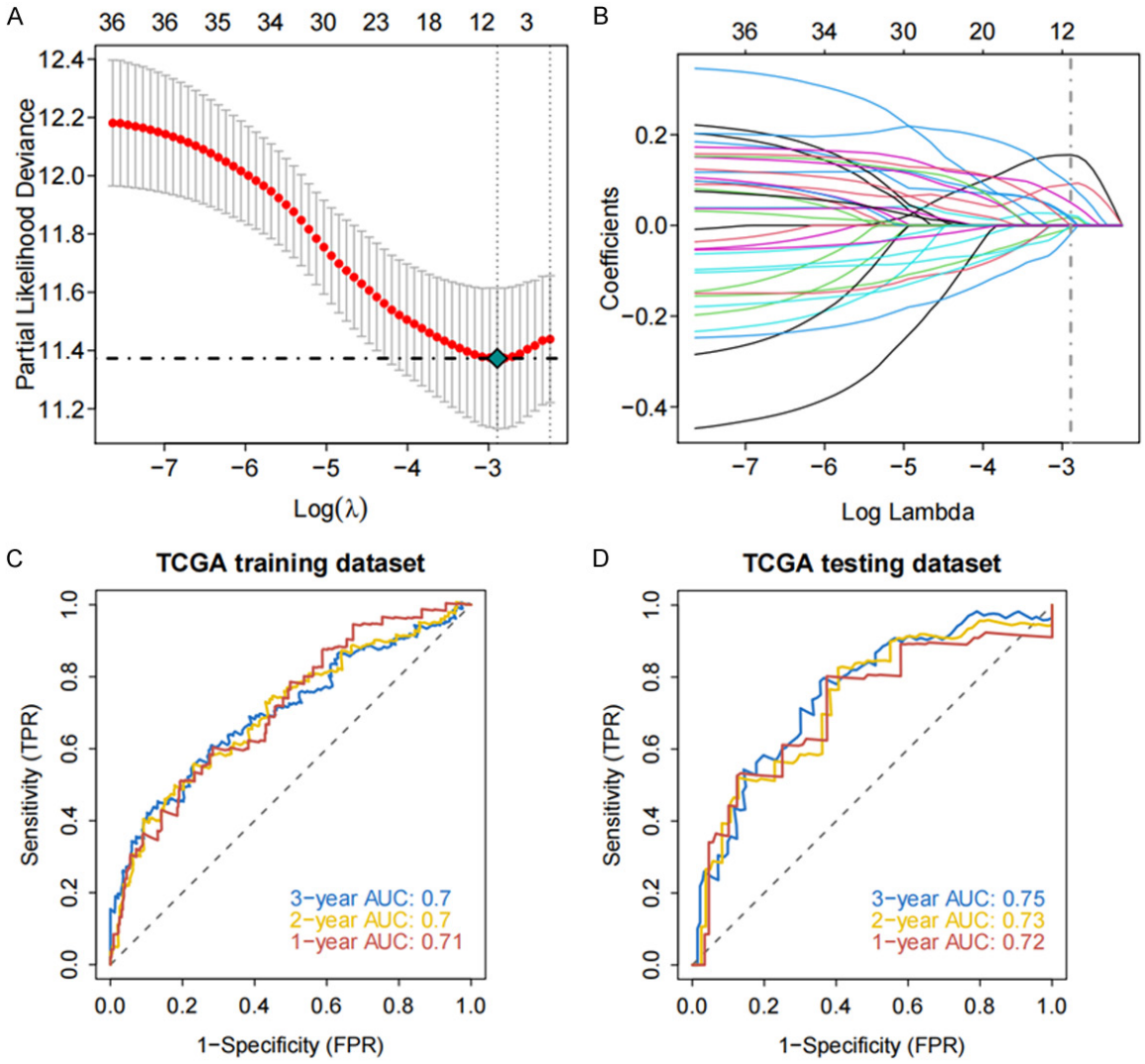


Figure S3. The characteristic genes in lung adenocarcinoma screened by lasso regression feature selection algorith. A: LASSO coefficient profiles of genes; B: Cross-validation for tuning parameter selection in the LASSO model; C: ROC curveof the genes in training dataset; D: ROC curve of the genes in testing dataset.

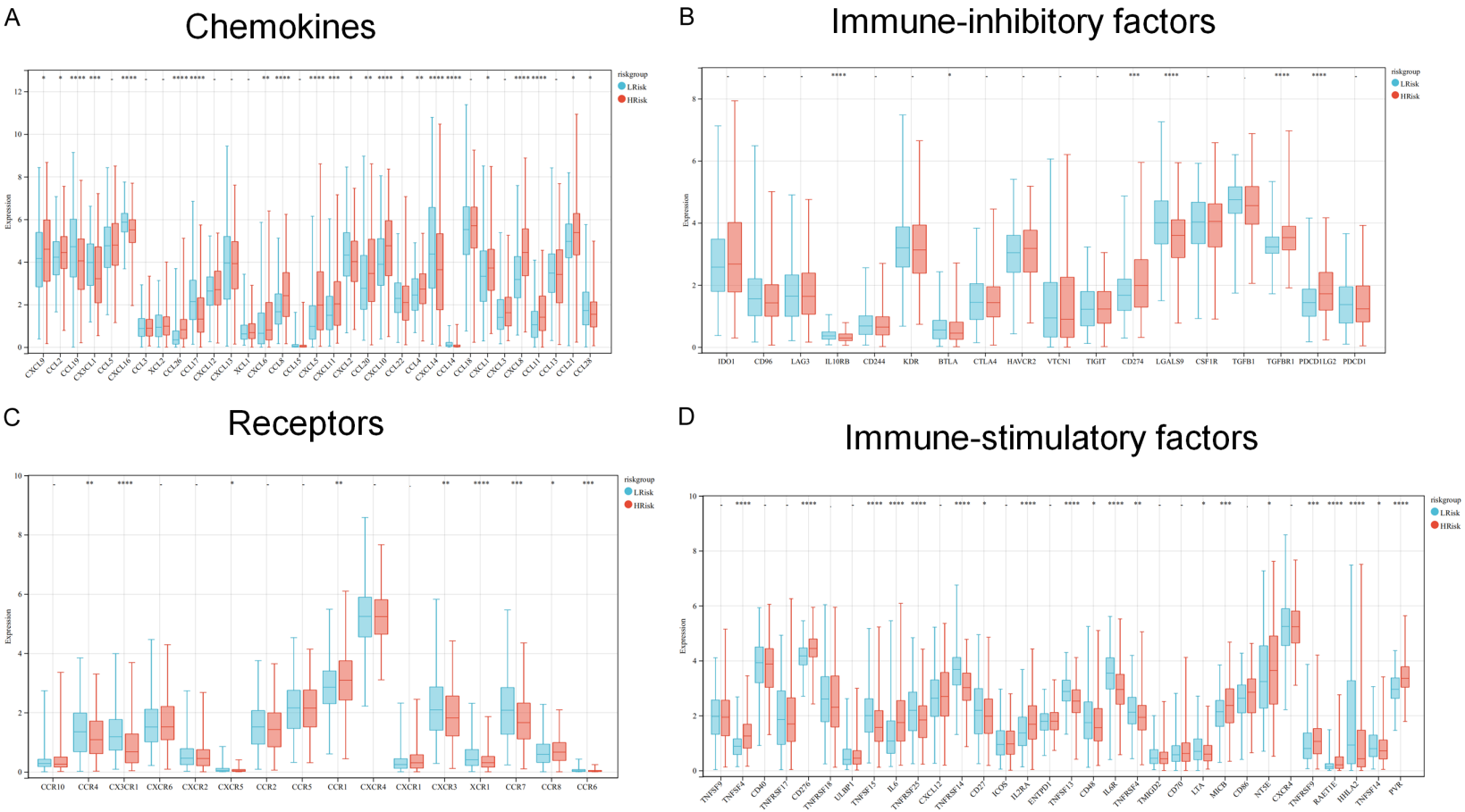


Figure S4. Immune status in high and low risk group. A: The expression differences of immune-related chemokines between the high and low risk groups; B: The expression differences of immune-inhibitory factors between the high and low risk groups; C: The expression differences of immune receptors between the high and low risk groups; D: The expression differences of immune stimulatory factors between the high and low risk groups; *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001.