Mathematical Biosciences and Engineering

MBE, 20(6): 10626–10658.
DOI: 10.3934/mbe.2023470
Received: 31 January 2023
Revised: 15 March 2023
Accepted: 24 March 2023
Published: 13 April 2023

http://www.aimspress.com/journal/MBE

Research article

Identification of coagulation-associated subtypes of lung adenocarcinoma and establishment of prognostic models

Mengyang Han[1], Xiaoli Wang[1], Yaqi Li[1], Jianjun Tan[2], Chunhua Li[2] and Wang Sheng[1,*]

[1] Department of Pharmacology, Beijing International Science and Technology Cooperation Base of Antivirus Drug, Faculty of Environment and Life, Beijing University of Technology, Beijing 100124, China
[2] Department of Biomedical Engineering, Faculty of Environment and Life, Beijing University of Technology, Beijing 100124, China

* **Correspondence:** Email: shengwang@bjut.edu.cn.

Abstract: Lung adenocarcinoma (LUAD), the most common subtype of lung cancer, is a global health challenge with high recurrence and mortality rates. The coagulation cascade plays an essential role in tumor disease progression and leads to death in LUAD. We differentiated two coagulation-related subtypes in LUAD patients in this study based on coagulation pathways collected from the KEGG database. We then demonstrated significant differences between the two coagulation-associated subtypes regarding immune characteristics and prognostic stratification. For risk stratification and prognostic prediction, we developed a coagulation-related risk score prognostic model in the Cancer Genome Atlas (TCGA) cohort. The GEO cohort also validated the predictive value of the coagulation-related risk score in terms of prognosis and immunotherapy. Based on these results, we identified coagulation-related prognostic factors in LUAD, which may serve as a robust prognostic biomarker for therapeutic and immunotherapeutic efficacy. It may contribute to clinical decision-making in patients with LUAD.

Keywords: lung adenocarcinoma; coagulation; unsupervised clustering; risk score; immunotherapy; cascade

1. Introduction

Lung cancer ranks first in both incidence (11.6%) and deaths (18.4%) of cancer [1]. In China, lung cancer is the most frequent and fatal cancer, with 17.1% morbidity and 21.7% mortality in 2015 [2]. Lung adenocarcinoma (LUAD) is the most common histologic subtype of lung cancer and accounts for approximately 40% of lung cancers [3]. Currently, surgery is the most common therapy for patients with early-stage LUAD, with or without perioperative chemotherapy (such as erlotinib, an epidermal growth factor receptor tyrosine kinase inhibitor, EGFR TKI), immunotherapy (such as bevacizumab), or radiotherapy (such as stereotactic body radiation) [4–7]. In addition to platinum-based dual therapy with or without bevacizumab, patients with advanced LUAD also benefit from targeted molecular therapies such as gefitinib, erlotinib, afatinib, osimertinib (EGFR TKIs) and crizotinib, ceritinib and alectinib (anaplastic lymphoma kinase inhibitor, ALK inhibitors), among others [8,9]. Despite significant clinical improvements in the molecular basis, diagnosis and treatment of LUAD, the recurrence rate remains high [10], and the clinical outcome remains poor, with overall survival of only about 15% at five years. LUAD exhibits a high degree of heterogeneity and a tendency toward early metastasis. Cancer-associated thrombosis (CAT) is the second cause of death after natural disease [11]. In previous study, the incidence of pulmonary embolism (PE) and deep vein thrombosis (DVT) was found to be greater in lung cancer patients than in the general population [12]. Which was consistent with that thrombosis often indicates poor prognosis in lung cancer patients [13] and that coagulation abnormalities are an independent risk factor for death in lung cancer patients [14].

Therefore, in this study, we identified coagulation-related subtypes in patients with LUAD based on public databases and bioinformatics technology and developed a validated prognostic model to improve the diagnosis and prognosis of patients with LUAD.

2. Materials and methods

2.1. Data

Lung adenocarcinoma (LUAD) expression data, including count, fpkm, sample clinical information (phenotype), sample survival information (survival), annotation information of all sample-related genes (annotation.gene.probeMap) and TCGA version of survival information (more data than survival) were downloaded from UCSC Xena database (https://xena.ucsc.edu/) [15] as a training set. A total of 585 samples were selected, including 526 tumor samples and 59 normal samples. A total of 210 (203 after de-duplication) genes from 2 pathways, hsa04610 [16] (complement and coagulation cascade) and hsa04611 [16] (platelet activation), were downloaded as coagulation-related genes (CRGs) using the R package "KEGGREST".

Download the LUAD mutation maf file using the R package "TCGAmutations", then download the CNA data using the R package "TCGAbiolinks" and annotate them. Then we used R package "GEOquery" to download lung adenocarcinoma-related datasets as validation sets (GSE3141 [17] and GSE72094 [17]), including normalized gene expression matrix and sample survival information, and 111 and 442 tumor samples with sample survival information were screened.

2.2. Somatic mutation and copy number alteration analysis

DNA alterations include mutations (truncations and missense) and copy number alterations (amplifications and deep deletions). After reducing the false positive rate, only non-silent mutations such as Missense_Mutation, Nonsense_Mutation, Nonstop_Mutation, Frame_Shift_Del, Frame_Shift_Ins, In_Frame_Del and In_Frame_Ins are retained. For studying the genomic features of CRGs in LUAD, we applied the "maftools" [18] package to analyze the mutation annotation format of TCGA (maf). Furthermore, we studied the relationship between overall survival (OS) and disease-free with SCNA and mutation and differential gene expression between patients with and without DNA alteration (SCNA and mutations).

2.3. Consensus clustering analysis of CRGs

Consensus clustering analysis is a standard method for classifying cancer subtypes, which can be used to differentiate samples into subtypes based on different histological data sets to discover new disease subtypes or to perform a comparative analysis of different subtypes. Consensus clustering is implemented via the resampling method to extract a specific sample of the data set, specify the number of clusters as K and calculate the plausibility under different numbers of clusters. In this step, we perform unsupervised consistency clustering analysis on tumor samples in TCGA-LUAD based on the expression of CRGs using "ConsensusClusterPlus" (https://www.bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html) package to typify the disease samples according to the CDF curves. To evaluate the results of the consensus clustering analysis, we performed T-SNE [19]. T-SNE is to downscale and visualizes the high-dimensional data, using Gaussian distribution to convert the distances into probability distribution in high-dimensional space and using long-tailed distribution to convert the distances into probability distribution in low-dimensional space so that the middle and low distances in high-dimensional space can have a larger distance after mapping, which can avoid focusing too much on local features and ignoring global features when downscaling.

2.4. Identification and verification of the critical CRGs

We aimed to explore differential genes that are both variably expressed among different subtypes and significantly associated with tumors. Differential expression analysis of disease samples and normal samples in the TCGA-LUAD dataset (Tumour vs. Normal) was performed using the R package "DESeq2" [20] to filter out differentially expressed gene sets (DEGs1) with screening thresholds $|\log2FC| > 2.5$ and $p.adj < 0.05$. The results are presented as heatmaps and volcano plots to get an overall picture of the distribution of DEGs. The differential expression analysis of cluster I and cluster II was performed to screen for differentially expressed gene sets (DEGs2) with screening thresholds of $|\log2FC| > 2.5$ and $p.adj < 0.05$. The results are presented in heatmap and volcano plots to get an overall picture of the distribution of genes. The two differential gene sets DEGs1 and DEGs2 obtained from the previous steps were intersected and defined as CRGs-DEGs. Then we performed GO (Gene ontology) [21] and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. GO is a database established by the Gene Ontology Consortium. GO consists of biological processes (BP), molecular functions (MF) and cellular components (CC) to describe the functions of gene products. The KEGG

database is a comprehensive database that contains 17 significant databases in four categories: system information, genomic information, chemical information and health information and the KEGG pathway database contains seven pathways: metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development. This paper analyzes and visualizes the Go system and KEGG pathways involving CRGs-DEGs crossover genes using the "clusterProfiler" [22] package.

2.5. Construction of the coagulation-related risk score

To further reveal the potential prognosis and molecular mechanism of coagulation pathways, we screen out the optimal prognostic biomarkers using the Lasso Cox regression model by the glmnet R package. In the modeling, ten genes (65 intersecting genes were subjected to univariate Cox regression analysis to screen for genes significantly associated with survival, $p < 0.05$) were constructed using the R package "glmnet" [23]. The best λ value for screening was 0.09514943 (indicating the λ corresponding to the minor error mean and the other indicating the maximum λ corresponding to the error mean within one minimum standard deviation). The PCA principal component analysis was performed using the fpkm of the model genes to downscale and visualize the high-dimensional data. The risk scores of the tumor samples were calculated according to the formula as Eq (1). Next, the samples were divided into high-risk and low-risk groups according to the median risk scores.

C-indexes and receiver operating characteristic curves (ROC) were commonly used to test the accuracy of the Cox regression model. The ROC analysis was conducted using the "survivalROC" [24] package. The AUC value of the ROC curve was calculated to evaluate the performance of the prognosis prediction model and compared with other single prognostic biomarkers. Furthermore, we conducted the K-M survival analyses of risk scores and drew a 1-, 3- and 5-year ROC curve using the R package timeROC based on training and test sets (GSE3141 and GSE72094).

$$\begin{aligned}\text{Risk score} = {} & \text{EXP}_{\text{LGI3}} * (-0.1974) + \text{EXP}_{\text{UGT3A1}} * (-0.2078) + \text{EXP}_{\text{HMGA2}} * 0.1580 + \\ & \text{EXP}_{\text{FGA}} * 0.1331 + \text{EXP}_{\text{NEUROD1}} * (-0.3336) + \text{EXP}_{\text{INSL4}} * 0.0436 + \text{EXP}_{\text{SFTPC}} * \\ & (-0.0099) + \text{EXP}_{\text{UGT2B15}} * (-0.1097) \end{aligned} \quad (1)$$

2.6. The clinical correlation analysis

In order to further understand the correlation between risk score and clinical characteristics, the R package "ComplexHeatmap" [25] was employed to map the expression of biomarkers in high-risk and low-risk groups, and in different clinical characteristics. We collected clinical indicators, including coagulation subtype, age, gender, tissue origin, tumor stage, smoking status and pathological stage of LUAD from TCGA, and compared the differences in survival indicators in different risk groups. A box plot was used to visualize the comparative analysis, and the Wilcoxon test was employed to calculate the significance p-value. $p < 0.05$ was used as the cutoff to screen out significant clinical features.

2.7. Construction and evaluation of nomogram

First, we performed the prognostic model construction. Univariate and multivariate analyses of clinical variables and risk scores were performed using Cox regression in training cohorts. Then we constructed a nomogram using the "rms" R package to estimate the 1-, 3- and 5-year survival rates of

LUAD patients. The calibration curves evaluated the discrimination and accuracy of the nomogram. Furthermore, we used Decision curve analysis (DCA) to identify the clinical application of the model.

2.8. Functional enrichment analysis

We first utilized the differential expression analysis of genes in high-risk and low-risk groups was performed using the R package "DESeq2", with the differential gene screening condition of $|\log2FC| > 1$, $p < 0.05$, Mapping the volcano plot with the R package "ggplot2" and differential expression heatmap using the "pheatmap" (https://cran.r-project.org/web/packages/pheatmap/) package. GO and KEGG enrichment analysis of differentially expressed genes between high-risk and low-risk groups were again performed with the R package "Clusterprofiler". Then all genes in the high-risk and low-risk groups were analyzed for GSEA enrichment using the R package "clusterProfiler", with a screening threshold of $|NES| > 1$, NOM $p < 0.05$ and $q < 0.25$. The top 10 were selected for GO enrichment and KEGG pathway mapping.

2.9. The immune microenvironment correlation analysis

Tumor immune cell infiltration refers to the movement of immune cells from the bloodstream to tumor tissue to begin to exert their effects and can be isolated from tumor tissue as infiltrating immune cells. Immune cell infiltration in tumors is closely related to clinical outcomes, and immune cells infiltrating in tumors are most likely to be used as drug targets to improve patient survival. The percentage abundance of tumor-infiltrating immune cells in each sample in TCGA-LUAD was calculated by CIBERSORT [26] algorithm, and the percentage of tumor-infiltrating immune cells in high- and low-risk groups were plotted according to the results, and immune cells of differentially expressed between high-risk and low-risk groups were obtained.

3. **Results**

3.1. CRGs mutation status in LUAD

The CRGs mutation and CNA data showed 439 tumor samples, including 429 mutations, 54 amplifications, eight profound deletions and 50 mutations and CNA changes simultaneously, as shown in Figure 1A.

The mutation data indicate that the 20 genes with the highest number of CRGs-related mutations are exhibited in Figure 1C, accounting for 69.75% of the total number of mutations. Each column represents one sample, different colors represent different mutation types, each row represents one gene and the bar graph on the right indicates the number of samples with mutations in each gene. The CRGs with the highest mutation frequency are *COL3A2* (13%), *ITGAX* (10%), *F8* (10%), *ADCY2* (10%) and *PLCB1* (9%). Figure 1B shows the mutation frequencies of the first 20 CRGs in different mutation groups, and different colors represent different mutation groups. Among the CRGs, *ADCY2*, *C6*, *C7*, *C9*, *C4BPA*, *C4BPB*, *CD46* and *CD55* possess a higher alterations event frequency.

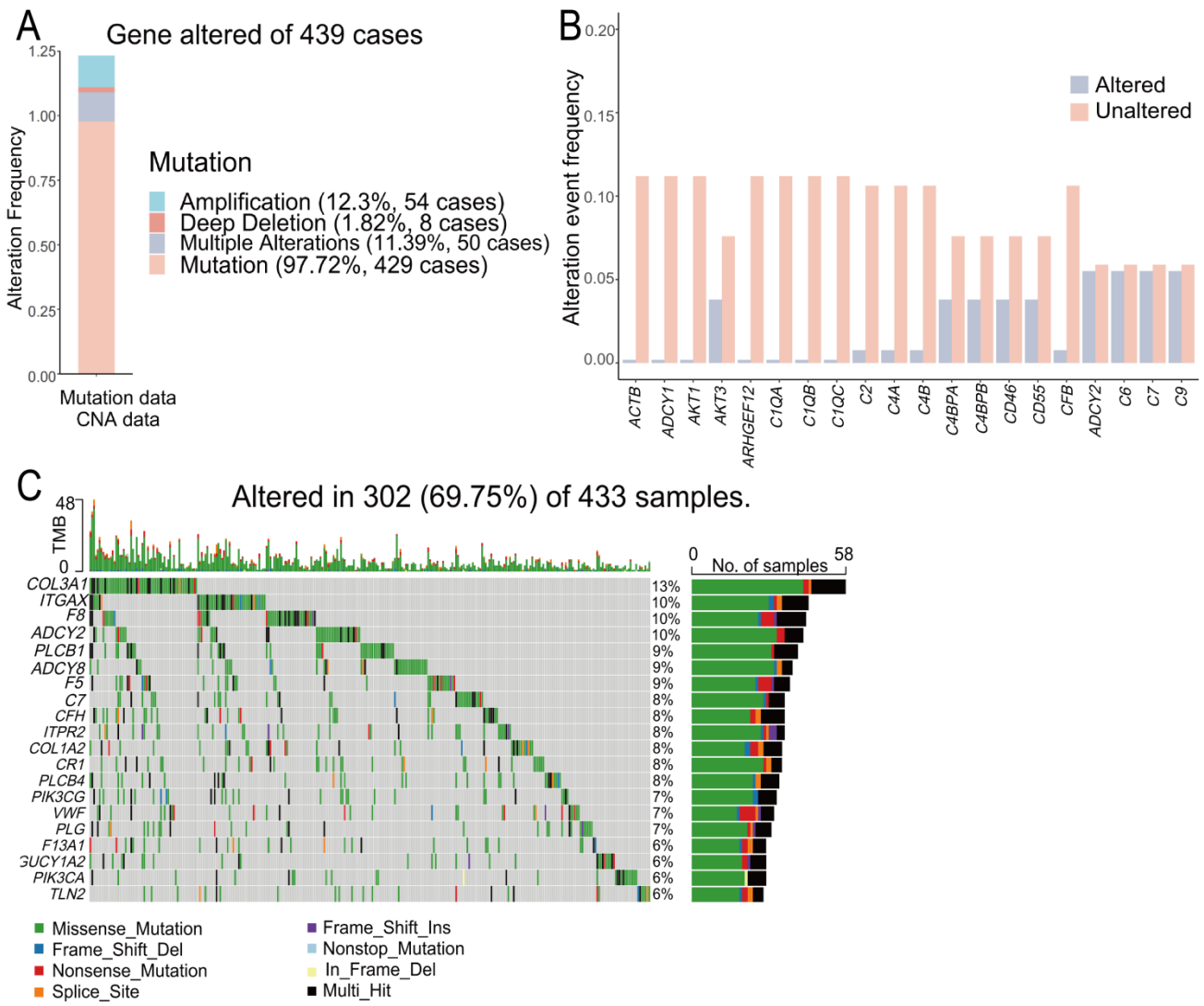

Figure 1. (A) Mutations and CNA alterations; (B) Landscape of genomic alterations of the CRGs in HCC; (C) Frequency of CRGs mutations in different mutation groups (only 20 are manifested).

The Kaplan-Meier survival analysis results for disease-free survival and overall survival for the different mutation groups are presented in Figure A1, which shows that the differences between the two survival analyses for the different subgroups were insignificant ($p = 0.91$, $p = 0.55$). The difference in sample size between survival analyses for the same subgroups was due to the different survival information recorded.

3.2. CNA

Figure 2B reveals the copy number profile of the CRGs. The horizontal coordinates are the genes, the vertical coordinates are the copy number changes of the genes, blue indicates amplification, red indicates deep deletion and the numeric labels indicate the frequency of each mutation. The CRGs are *ADCY2*, *C6*, *C7*, *C9*, *AKT3*, *C4BPA*, *C4BPB*, *CD46*, *CD55*, *CFH*, *CFHR1*, *CFHR2*, *CFHR3* and other

genes possess a higher alterations frequency (3.8–5.7%) coming from amplification but almost no deep deletion. Thus, most CRGs with high CNA frequency tended to be co-amplification rather than co-deletion.

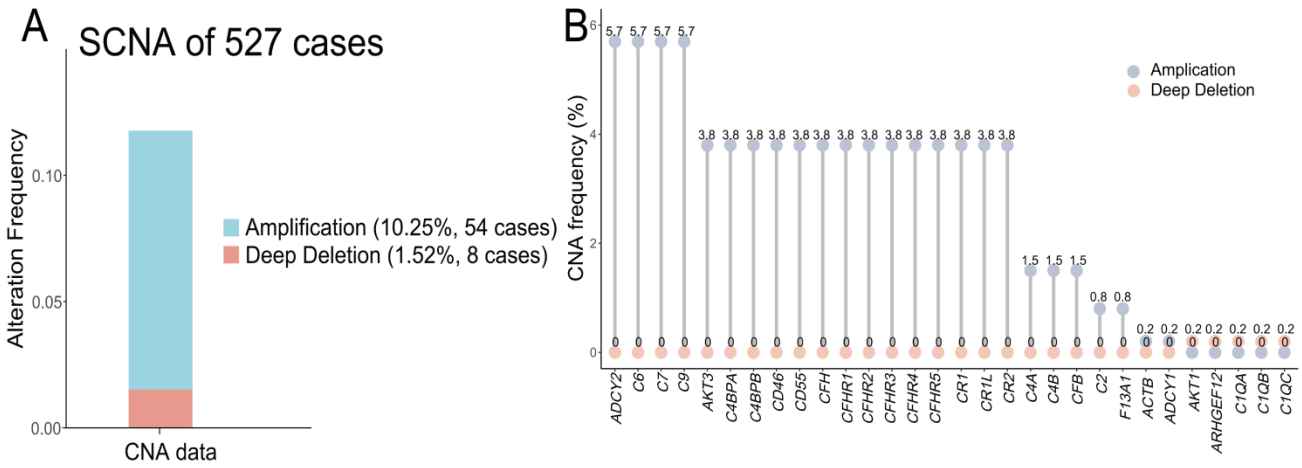

Figure 2. (A) CNA mutation status; (B) CRGs copy number alterations (top 30).

3.3. Identification and evaluation of CRGs-related subtypes in LUAD samples

3.3.1. Consensus clustering analysis

The number of clusters in this project is chosen to be K = 2, and the frequency of clustering is 1000 times to ensure the stability of clustering. The clustering results are manifested in Figure 3A, which shows that the CRGs-related samples are clearly classified into two parts. We performed a t-SNE analysis of the coagulation subtypes in the TCGA cohort to observe whether the results clearly distinguish the two subtypes. Figure 3C manifest the results. Orange indicates the first cluster of coagulation subtypes, and blue indicates the other cluster. The overlap between the two clusters represents some correlation between the groups. The fewer the intersecting areas, the better, as this illustration shows that the two clusters (i.e., the two subtypes) can be distinguished.

3.3.2. Survival analysis in different subtypes

K-M curves were plotted for different subtypes to compare the survival differences between patients of different subtypes, and the results are presented in Figure 3B. The horizontal coordinate is the survival time, and the vertical coordinate is the survival rate. The orange curve is the survival status of coagulation subtype cluster I, and the blue is the survival status of cluster II, $p = 0.0072$, indicating a significant difference in survival status between patients of the two clusters and further indicating that the previous subtypes are well identified.

The program shows the relationship between patients with coagulation subtypes and survival status, tumor stage and gender, as shown in Figure 4A. The curve's width indicates the number of patients, and the first column indicates the cluster, the second column indicates the survival status, the third column indicates the tumor stage and the fourth column indicates the gender. The orange color

indicates patients with subtype cluster I. The second column of the program shows a branch in survival status. The branch into Alive is wider than Dead, indicating that more patients in cluster I survived than died, and the exact correspondence follows. The results of the clustering heatmap using CRGs expression data in lung adenocarcinoma patients are exhibited in Figure 4B. The horizontal coordinates represent the samples, one for each column. The vertical coordinates are the coagulation genes, i.e., one gene per row. Further, the different colors indicate the normalized values of gene expression (only the ten largest and ten most minor gene expression totals are demonstrated). Patients of cluster II had a higher proportion of *C3*, *C7* and *C4BPOA* than cluster I.

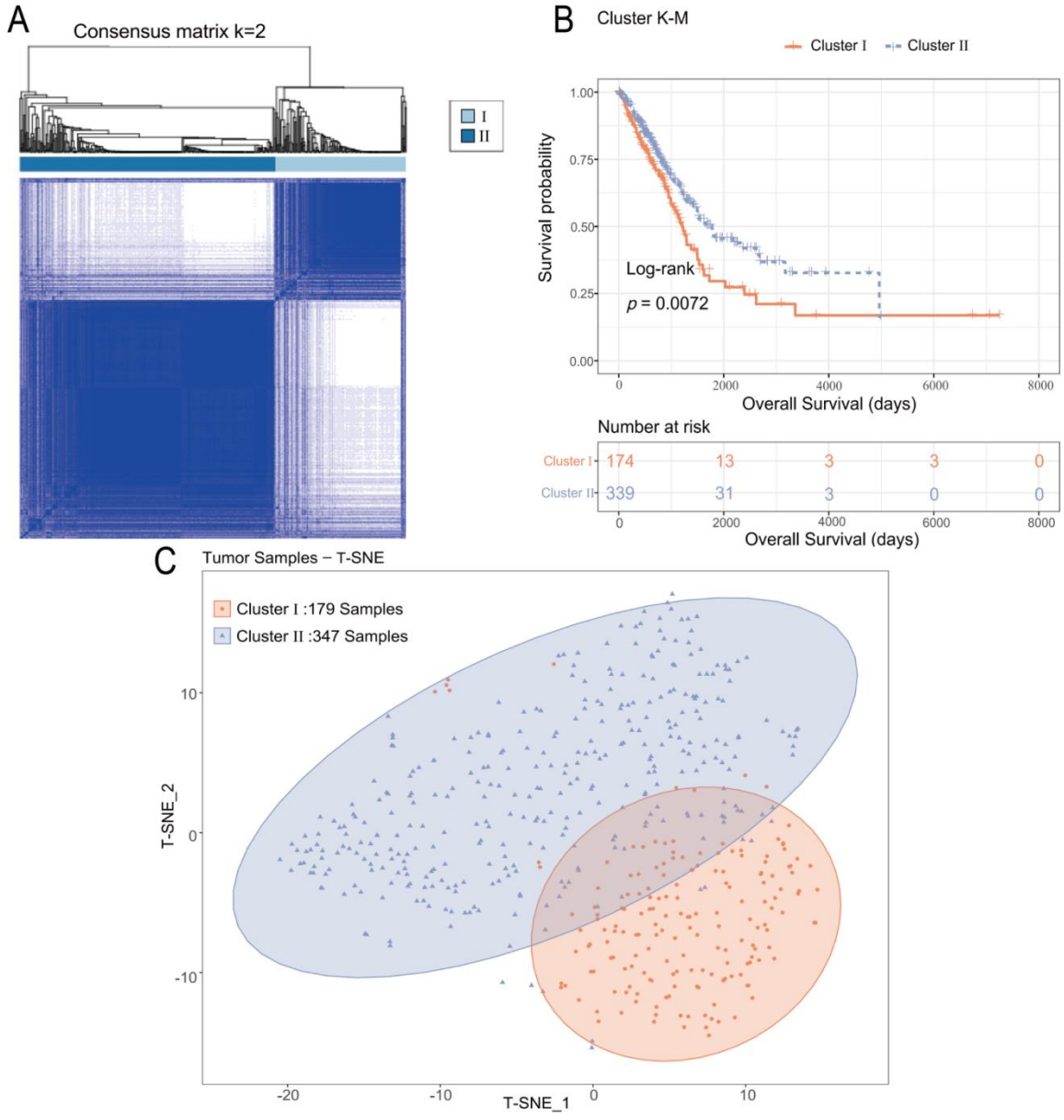

Figure 3. (A) Unsupervised clustering map; (B) K-M curves of different subtypes; (C) Coagulation subtypes TSNE analysis results.

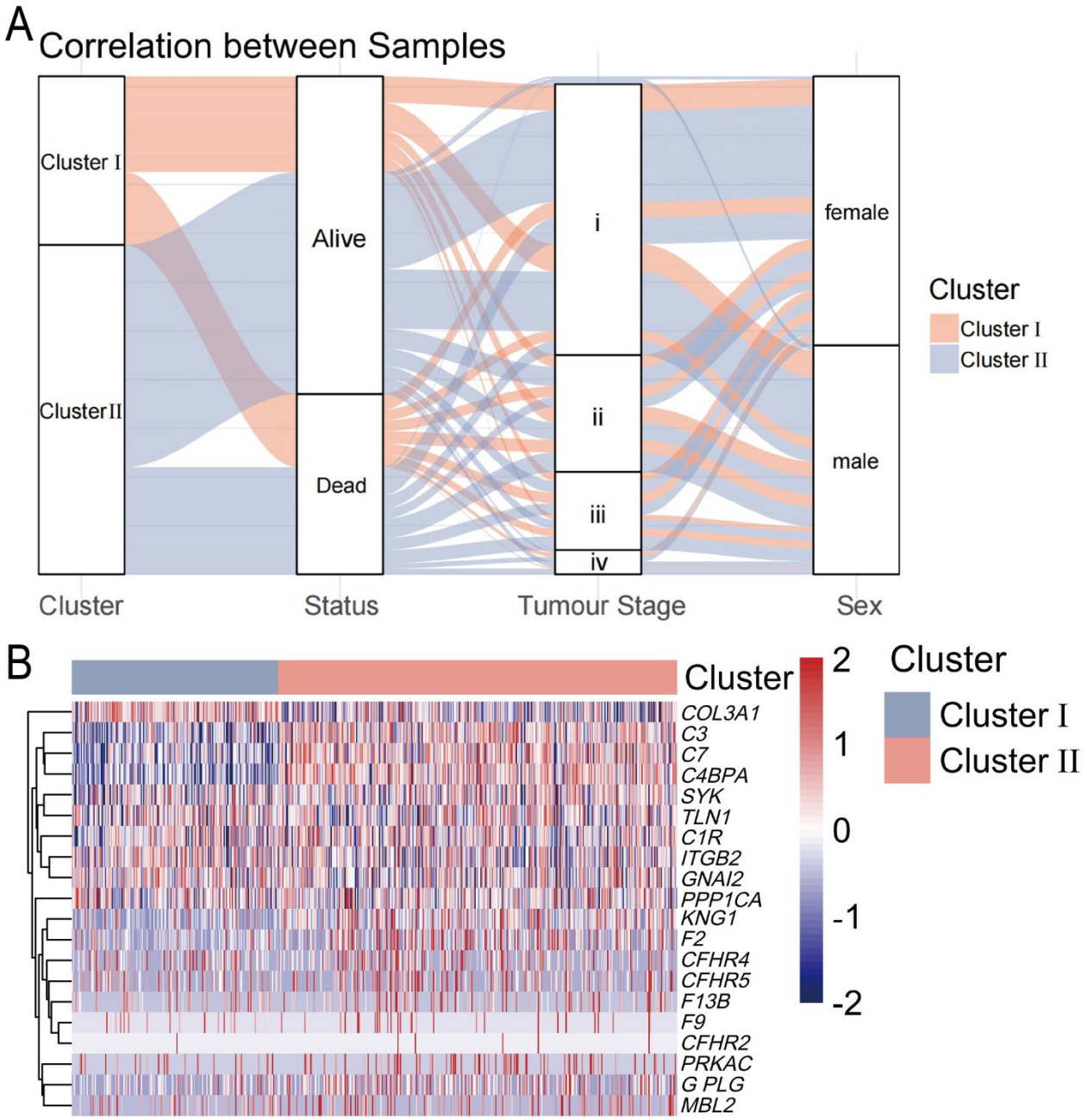

Figure 4. (A) Alluvial diagram of different subtypes; (B) CRGs expression heatmap.

3.4. Differential expression analysis for identification of the CRGs-DEGs

3.4.1. CRGs clustering heatmap

The results of the volcano plot of differential gene expression between tumor and normal samples are exhibited in Figure 5A. The horizontal coordinate is log2FC, where FC is fold change, which represents the ratio of gene expression in the tumor group to that in the normal group, and log2FC is obtained by taking the logarithm of 2. The vertical coordinate is the transformed p.adj. The dotted line is the threshold line, and the dots indicate genes, the gray dots indicate genes without differential

expression outside the threshold range, the blue dots indicate down-regulated genes, the red dots indicate up-regulated genes and the top 10 up-and down-regulated genes (sorted by the Log2FC sorting p.adj) are identified.

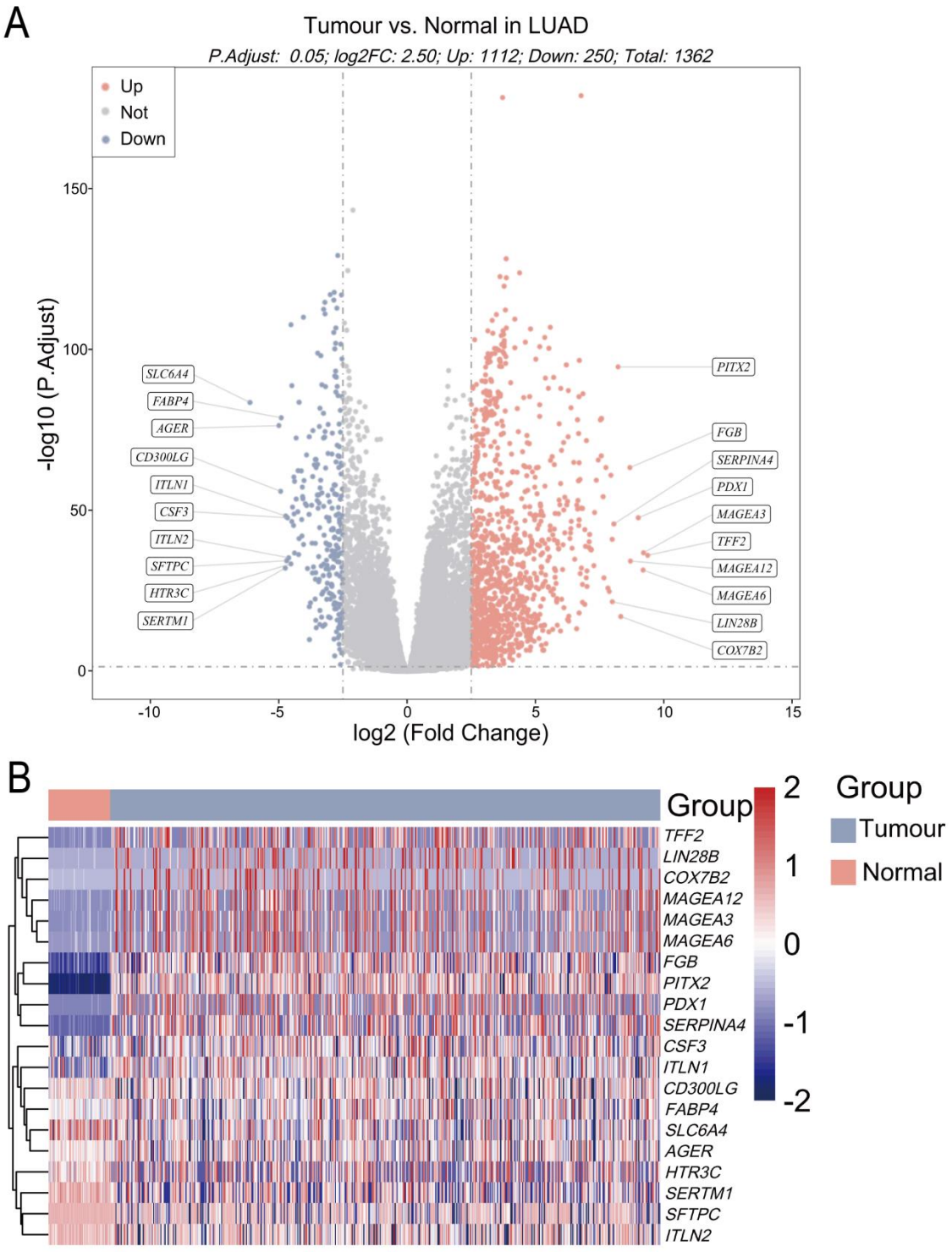

Figure 5. (A) Volcano plot for differential expression analysis of tumor and normal samples; (B) Heatmap of differentially expressed genes in tumor and normal samples.

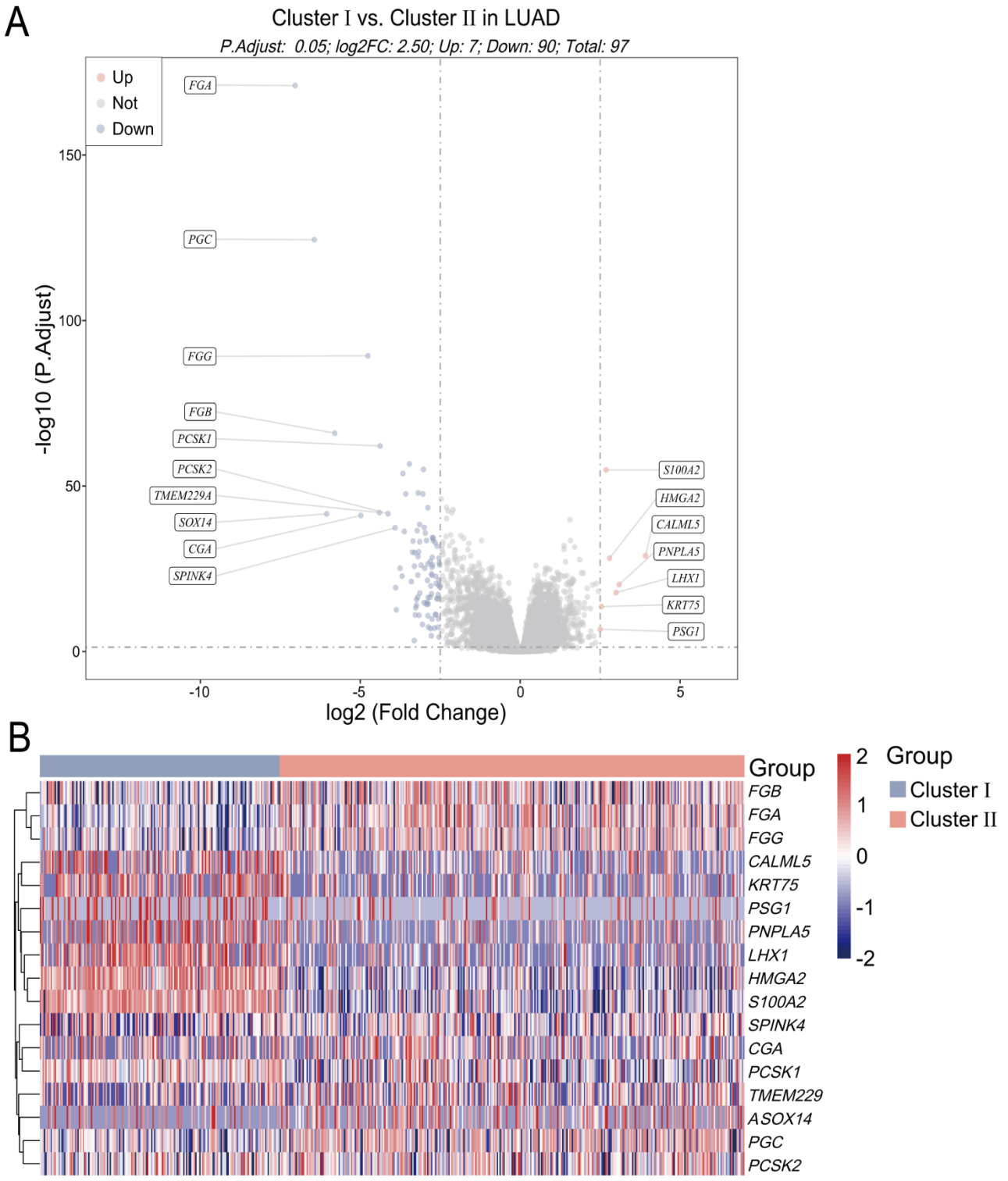

Figure 6. (A) Volcano plot for differential expression analysis of CRGs subtypes; (B) Heatmap of differential expression of CRGs subtypes.

The heatmap of DEGs in tumor and normal samples is shown in Figure 5B. The horizontal coordinates represent the samples, one sample per column and the vertical coordinates are the differentially expressed genes (the top 10 genes are presented for each up- and down-regulation), i.e.,

each row is one gene, and the different colors indicate the normalized values of gene expression. The colors at the top indicate the grouping of normal and tumor samples, and the clustering tree is on the left. The expression in the tumor group is higher than that in the normal group.

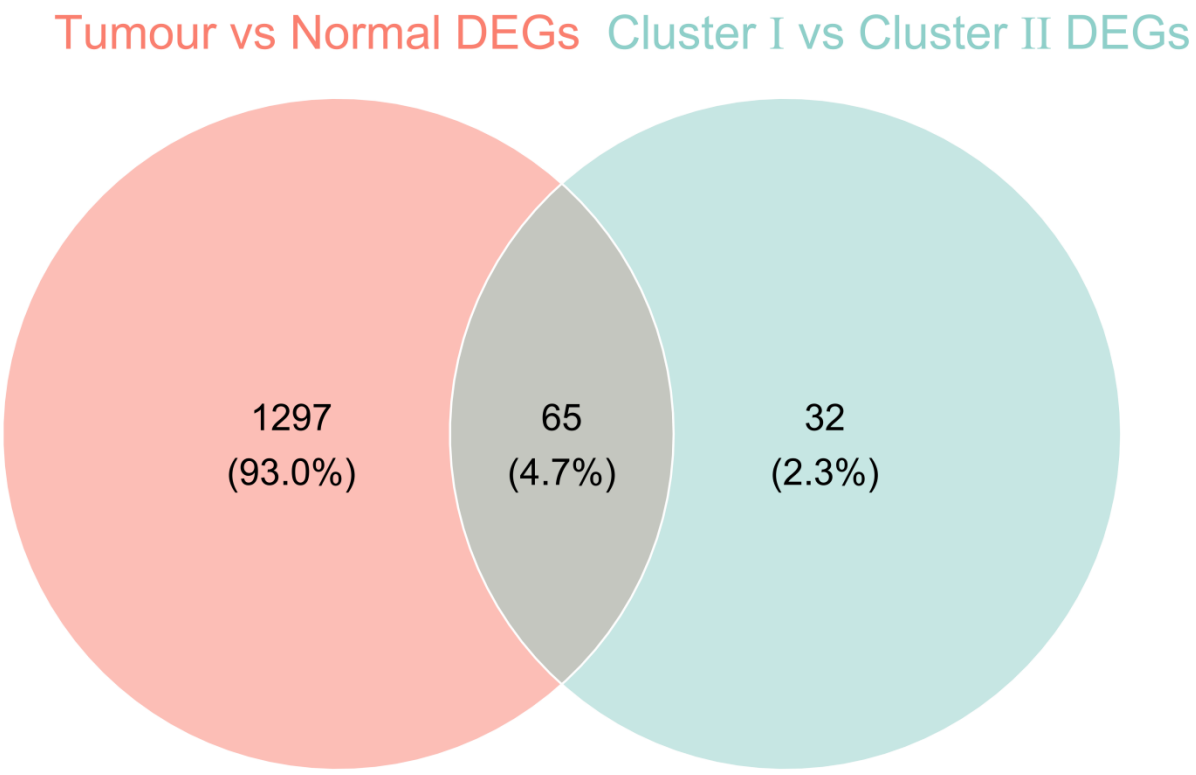

Figure 7. DEGs1 vs. DEGs2 Wayne diagram.

3.4.2. Differential expression analysis of CRGs subtypes

A volcano plot of the genetic differences associated with CRGs subtypes is manifested in Figure 6A, showing seven up-regulated genes and 90 down-regulated genes.

The results of the CRGs subtype-related gene expression heatmap are exhibited in Figure 6B, where the horizontal coordinates represent the samples, each column is one sample, and the vertical coordinates are the DEGs (top 20), i.e., one gene per row and the different colors indicate the normalized values of gene expression. The gene expression of cluster I is lower than cluster II on the right.

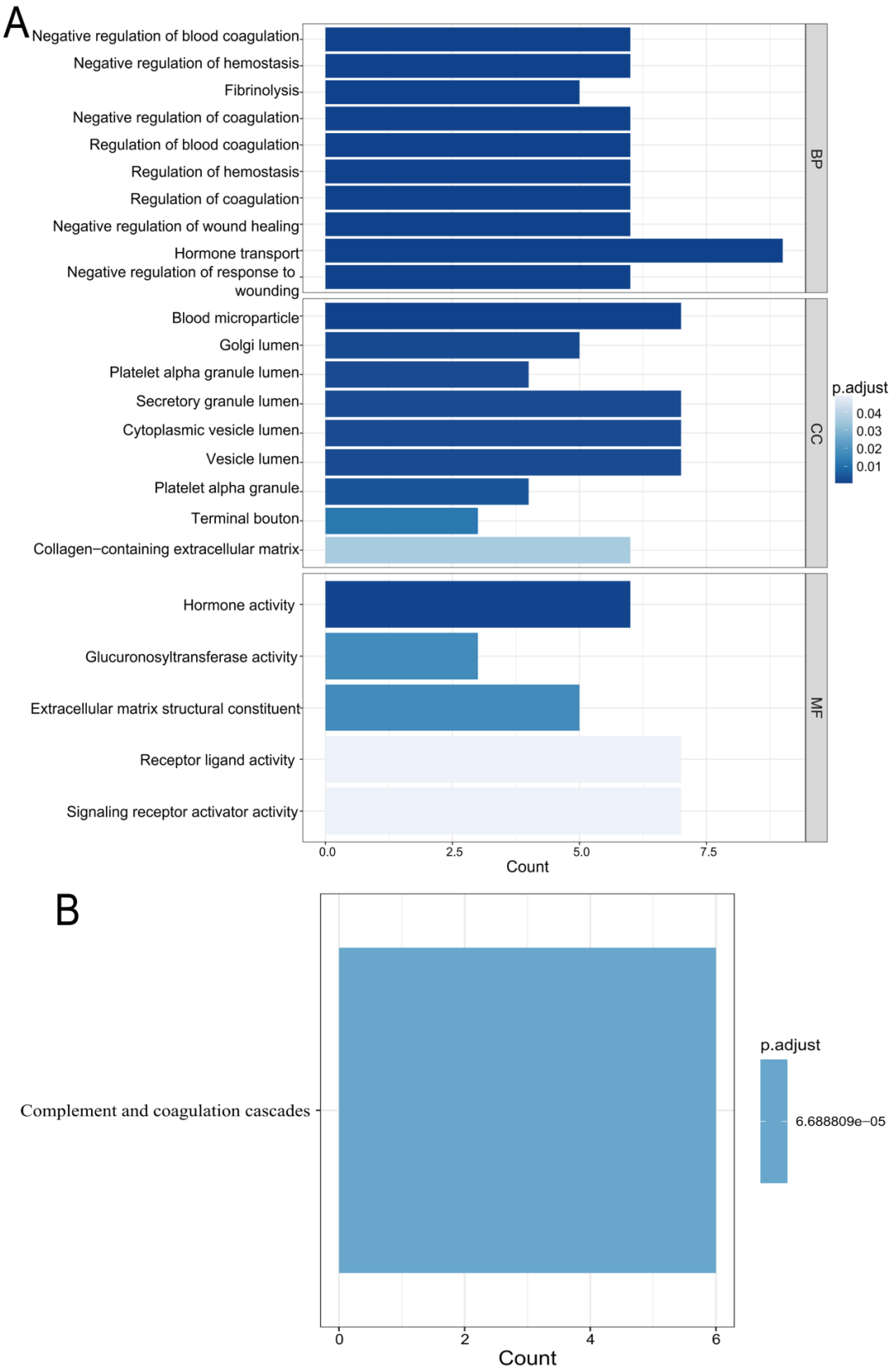

Figure 8. (A) GO enrichment analysis of intersecting genes; (B) KEGG enrichment analysis of intersecting genes.

The two differential gene sets DEGs1 and DEGs2 obtained from the previous steps were intersected and defined as CRGs-DEGs, and the results are shown in Figure 7. The left side shows the differential genes between tumor and normal samples, and the right side shows the differential genes of CRGs subtypes.

3.4.3. Gene ontology, KEGG enrichment analysis

We analyze and visualize the Go system and KEGG pathways involving CRGs-DEGs crossover genes. The results are presented in Figure 8A,B.

The horizontal coordinate is the number of genes enriched into each category. The vertical coordinate is the category of the intersecting genes enriched into the three parts of the GO database. The color shades indicate the p.adj size, only the first ten categories are shown in each part of the graph, and all the categories with less than ten are demonstrated. The top 10 categories of biological processes (BP) are negative regulation of blood coagulation, negative regulation of hemostasis, fibrinolysis, regulation of coagulation, regulation of hemostasis, regulation of coagulation, hormone transport and negative regulation of wound healing. From hormone transport and negative regulation of response to wounding, we can see that the crossover genes are mainly involved in biological processes related to coagulation and hemostasis. There are nine categories of cellular components (CC): blood microparticle, Golgi lumen, platelet alpha granule lumen, secretory granule lumen, cytoplasmic vesicle lumen, vesicle lumen, platelet alpha granule, terminal bouton and collagen-containing extracellular matrix, which can be seen as the intersecting genes are mainly the cellular components of various particles and granular lumens. There are five categories of molecular function (MF): hormone activity, glucuronosyltransferase activity, extracellular matrix structural constituent, receptor-ligand activity and signaling receptor activator activity.

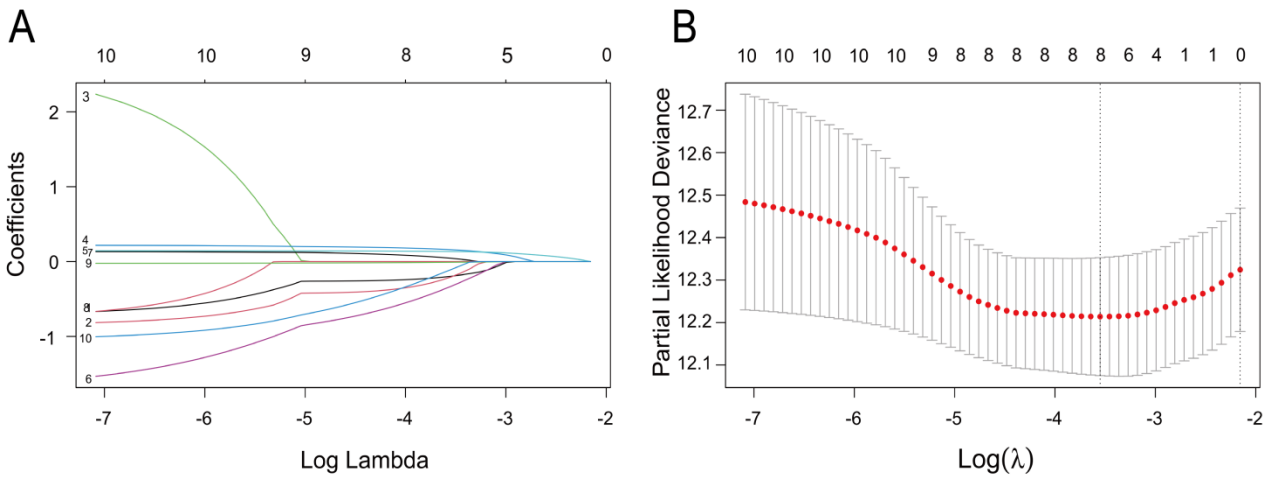

Figure 9. (A) Plot of Lasso Cox regression coefficients corresponding to λ; (B) λ screening diagram.

3.5. Construction and validation of Lasso Cox regression model

Lasso regression is a reduced form of linear regression that allows for variable and parameter estimation by introducing penalty terms.

3.5.1. Lasso Cox regression modeling and survival evaluation

As demonstrated in Figure 9, the best λ value for screening was 0.09514943. The coefficients of *PHOX2B* and *KCNU1* were penalized to 0 and were not included in the model. The coefficients of *LGI3*, *UGT3A1*, *HMGA2*, *FGA*, *NEUROD1*, *INSL4*, *SFTPC* and *UGT2B15* were not penalized to 0 and were included in the model construction.

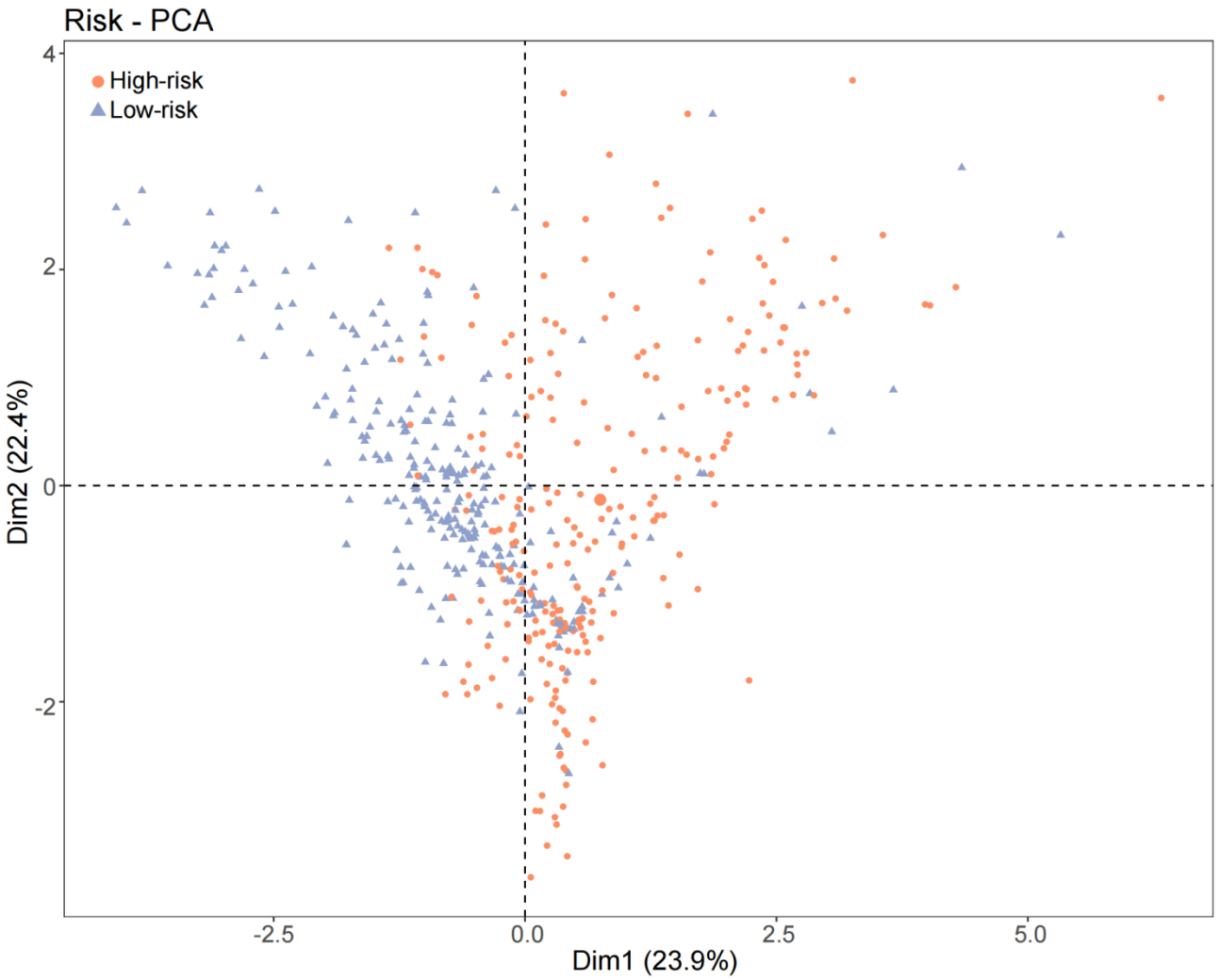

Figure 10. Results of PCA analysis in the high-risk group and low-risk group.

The PCA results are presented in Figure 10. The samples were divided into high-risk and low-risk groups according to the median risk scores, which reveals in different colors in the figure. The points in the two colors do not have overlapping areas, indicating that the results of PCA analysis based on gene expression are consistent with the results of grouping by median risk score.

Figure 11 displays the results of the Lasso model. The risk score distribution for the sample is exhibited on the top left, and the survival status distribution is exhibited on the bottom left. The vertical coordinate of the survival status distribution is the survival time, which decreases as the patient's risk score increases, where the red dot represents the patient's death.

Survival analysis was performed for the low-risk and high-risk groups. The results are presented in the K–M plot on the top right of the figure, with $P < 0.0001$ for the log-rank test, indicating a significant difference in survival between the two groups. False positives and true positives were calculated, and the results were used to plot ROC curves, shown in the figure's bottom right. The curve divides the whole graph into two parts. The area under the curve is called AUC (Area Under Curve), which is used to indicate the prediction accuracy. The higher the AUC value, the higher the prediction accuracy. The closer the curve is to the upper left corner (the smaller the X and the larger the Y), the higher the prediction accuracy. Here, the AUC of 1, 2 and 3 years is > 0.65, which has high accuracy.

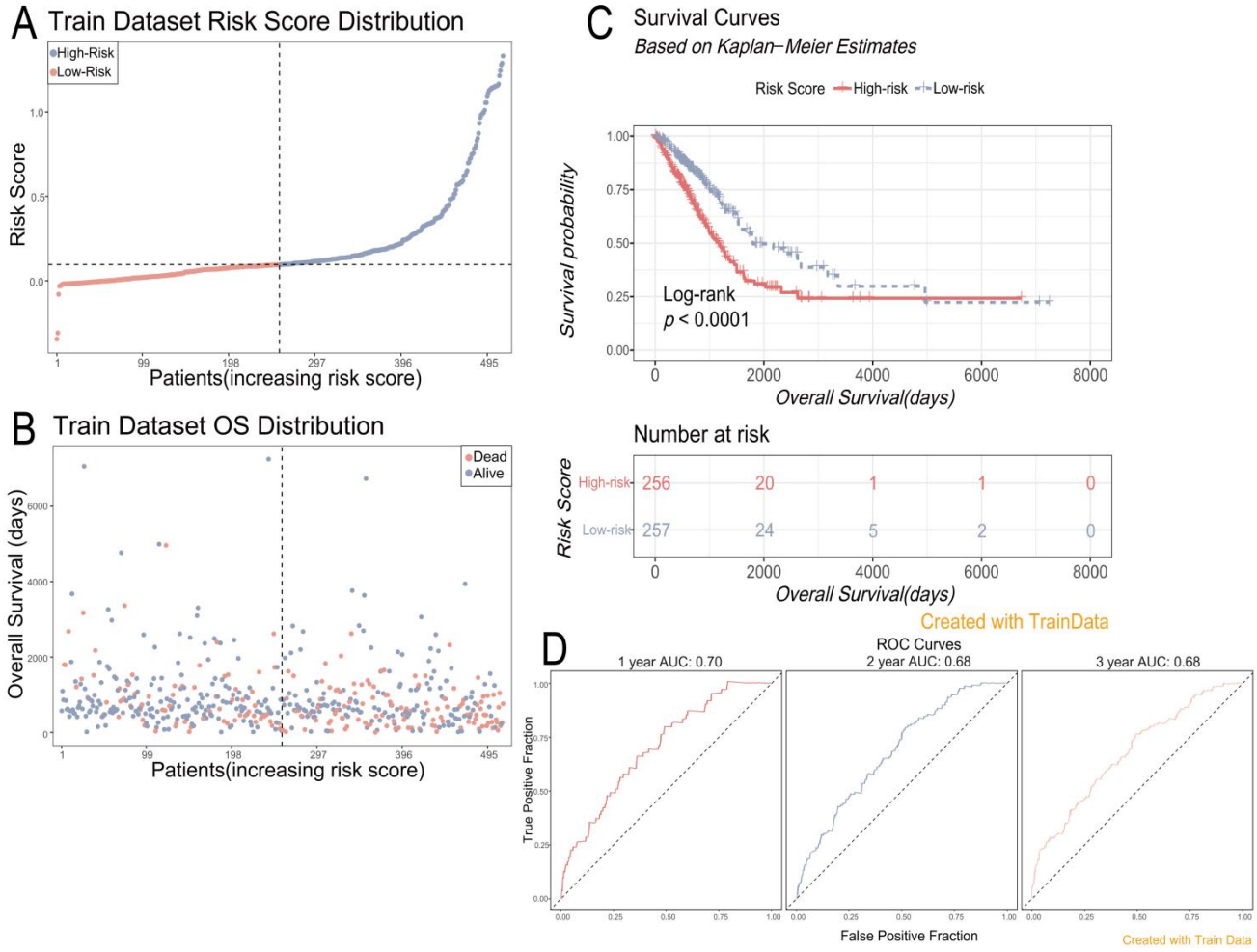

Figure 11. (A) Risk score distribution of the sample; (B) Survival state distribution; (C) Survival analysis of the high-risk group and low-risk group; (D) ROC Curve.

3.5.2. Validation of the Lasso Cox regression model

The model validation was performed using external datasets GSE3141 and GSE72094. The results in Figures 12 and 13 shows that the K-M curves log-rank test for both datasets have $p < 0.05$ and AUC > 0.6 at 1, 2 and 3 years, indicating that the constructed model can better predict the overall survival rate of lung cancer patients.

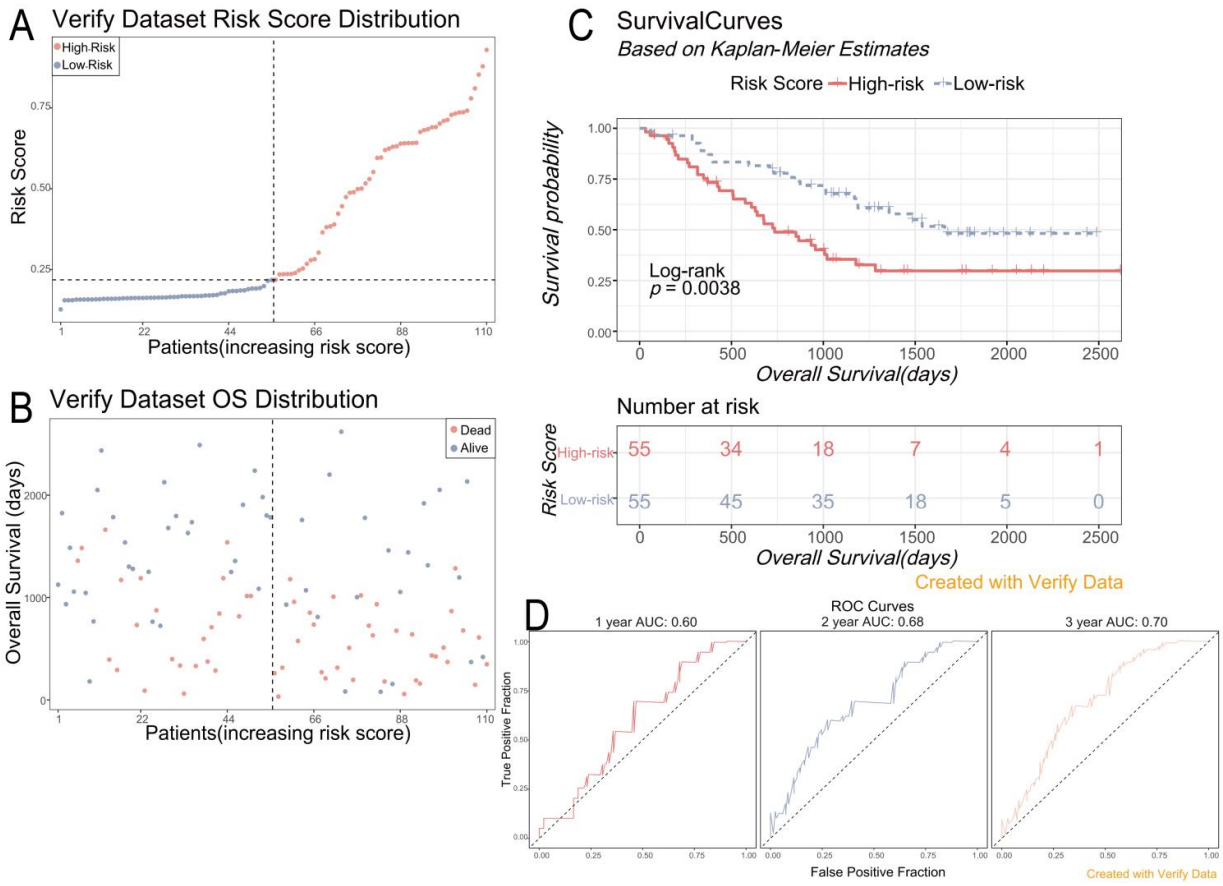

Figure 12. Lasso Cox model validation-GSE3141.

3.6. Correlation analysis between risk scores and clinical characteristic

3.6.1. Heatmap of different clinical features expression

As presented in Figure 14, we mapped out the expression of biomarkers in high-risk and low-risk groups and different clinical characteristics (coagulation subtype, age, gender, tissue origin, tumor stage, smoking status and pathological stage). The horizontal coordinates are the samples, the vertical coordinates are the model genes and the expression of the genes is exhibited from red to blue, with higher expression nearer to red and lower expression nearer to blue.

3.6.2. The correlation of risk score and clinical features

In order to verify the reliability of this model, we analyzed the differences in risk score between the cohorts with different clinical characteristic subgroups, including coagulation subtype (clusters I and II), age (> 60 and <= 60), gender (male and female), tissue origin (Lobe lower/middle/upper and NOS/Overlap/Bronchus), tumor stage (Stage i/ii and Stage iii/iv), smoking status (1–3 years, 4–5 years), pathological stage M (M0 and M1), pathological stage N (N0/N1 and N2/N3) and pathological stage T (T1, T2 and T3/T4). The results are exhibited in Figure 15 (and Figure A2). The horizontal coordinate is the subgroup category for each clinical feature, and the vertical coordinate is the risk

score, with the Wilcoxon test for two subgroup categories and the Kruskal-Wallis test for more than two. The results showed that $p > 0.05$ for tissue origin, pathological stage M, coagulation subtype and age were not significantly correlated; $p < 0.05$ for tumor stage, gender, smoking status, pathological stage N and pathological stage T were significantly correlated.

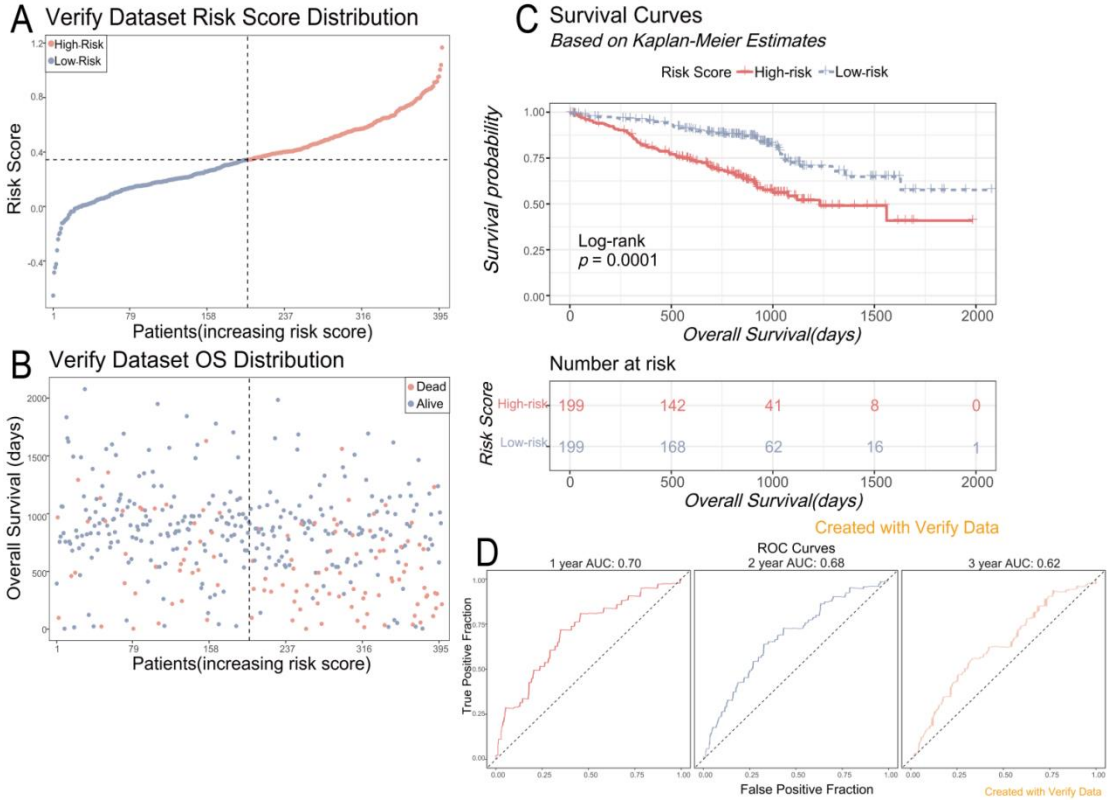

Figure 13. Lasso Cox model validation-GSE72094.

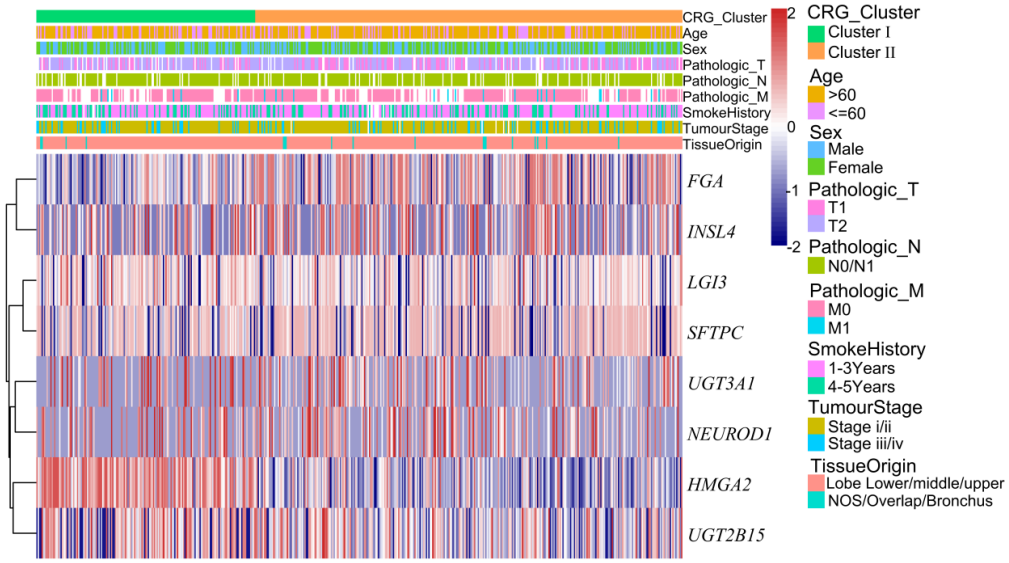

Figure 14. Heatmap of different clinical features expression.

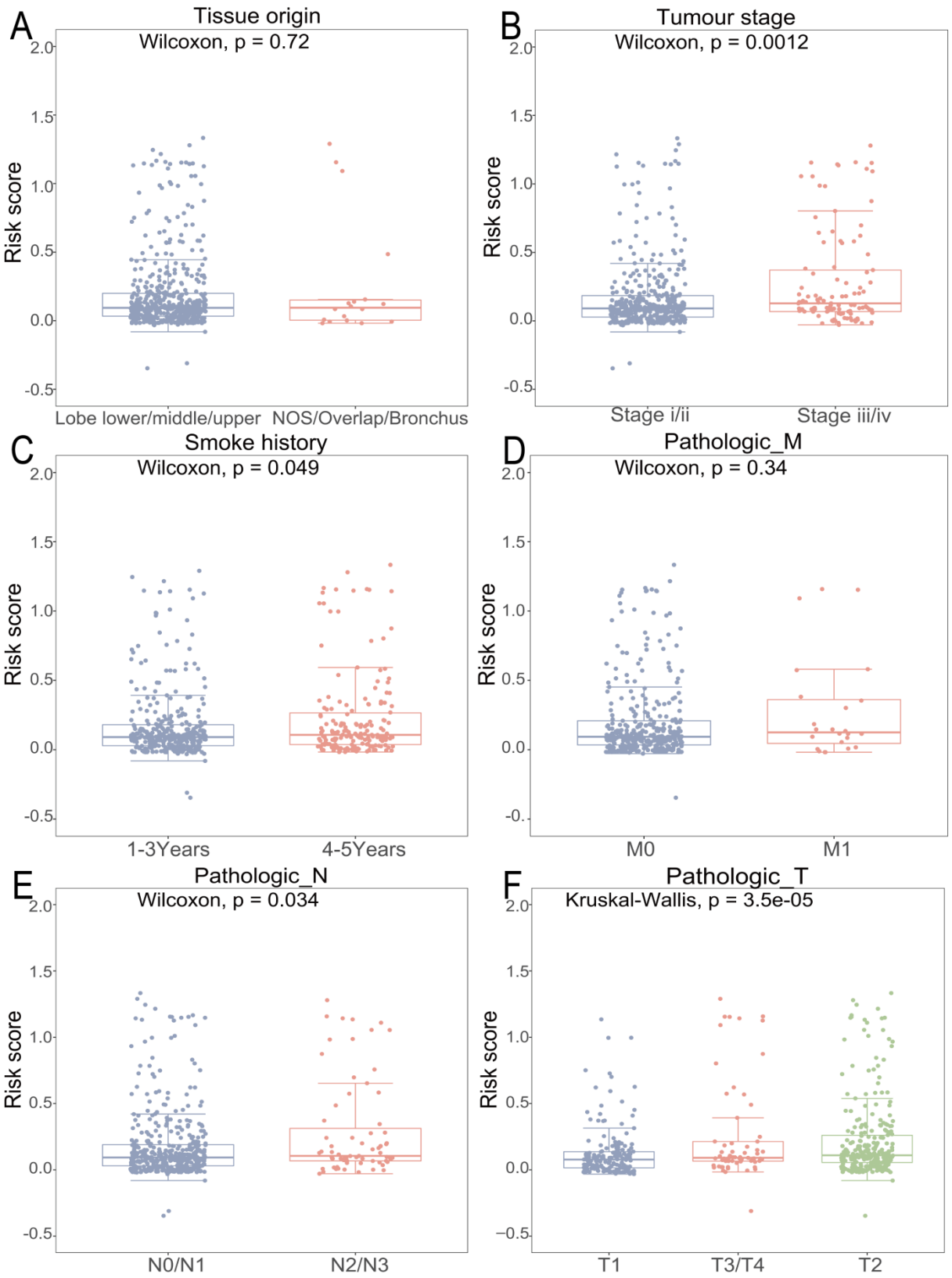

Figure 15. Correlation analysis of different clinical characteristics.

3.6.3. Survival analysis by subgroup for clinical features

The nine clinical characteristics were then further refined to analyze the risk scores of each subgroup and plot the K-M survival curves, and the results are presented in Figure 16 (and Figure A3). The results show that the risk model can be applied to different clinicopathological characteristics, except for cluster I, age <= 60, pathological stage M1 and tissue origin NOS/Overlap/Bronchus, where the log-rank test $p > 0.05$, and the rest of the stratification tests $p < 0.05$.

3.7. Independent prognostic analysis and evaluation

3.7.1. Construction of prognostic nomogram

Combining risk scores and clinical characteristics (coagulation subtype, age, gender, tissue source, tumor staging, smoking status, pathological staging M, pathological staging n, pathological staging T), the prognosis nomogram model for patients' survival is constructed. First, the risk score, coagulation subtype, age, gender, tumor staging, smoking stage and tumor MNT staging. These nine variables perform univariate Cox regression analysis. The analysis results show that risk scores, coagulation subtypes, tumor staging and pathological MNT staging were related to the patient's survival prognosis, and the results are shown in Figure 17A.

Therefore, a Multivariate cox regression model was constructed using risk score, coagulation subtype, tumor stage and pathological MNT stage. After the PH hypothesis test, the tumor stage and pathological stage M that did not pass the test were removed. Only the risk score, coagulation subtype, pathological stage N and pathological stage T were used to construct the Multivariate cox regression model (prognostic model). The forest diagram of the model is manifested in Figure 17B, and the model results are manifested in Figure 17C.

3.7.2. Validation of prognostic nomogram

The prognostic model was evaluated using calibration curves and DCA curves; the results are exhibited in Figure 18. The results show that the survival prediction is close to the theoretical straight line, and the decision curves for clinical characteristics are under the ProModel, indicating the good predictive performance of the model.

3.7.3. Differential gene expression analysis of high-risk and low-risk groups

After we utilized differential expression analysis of genes in high-risk and low-risk groups, the volcano plot in figure 19 shows there were 3776 up-regulated genes and 442 down-regulated genes, and the top 10 up-regulated and down-regulated genes were displayed. Heatmap shows the top 20 differentially expressed genes.

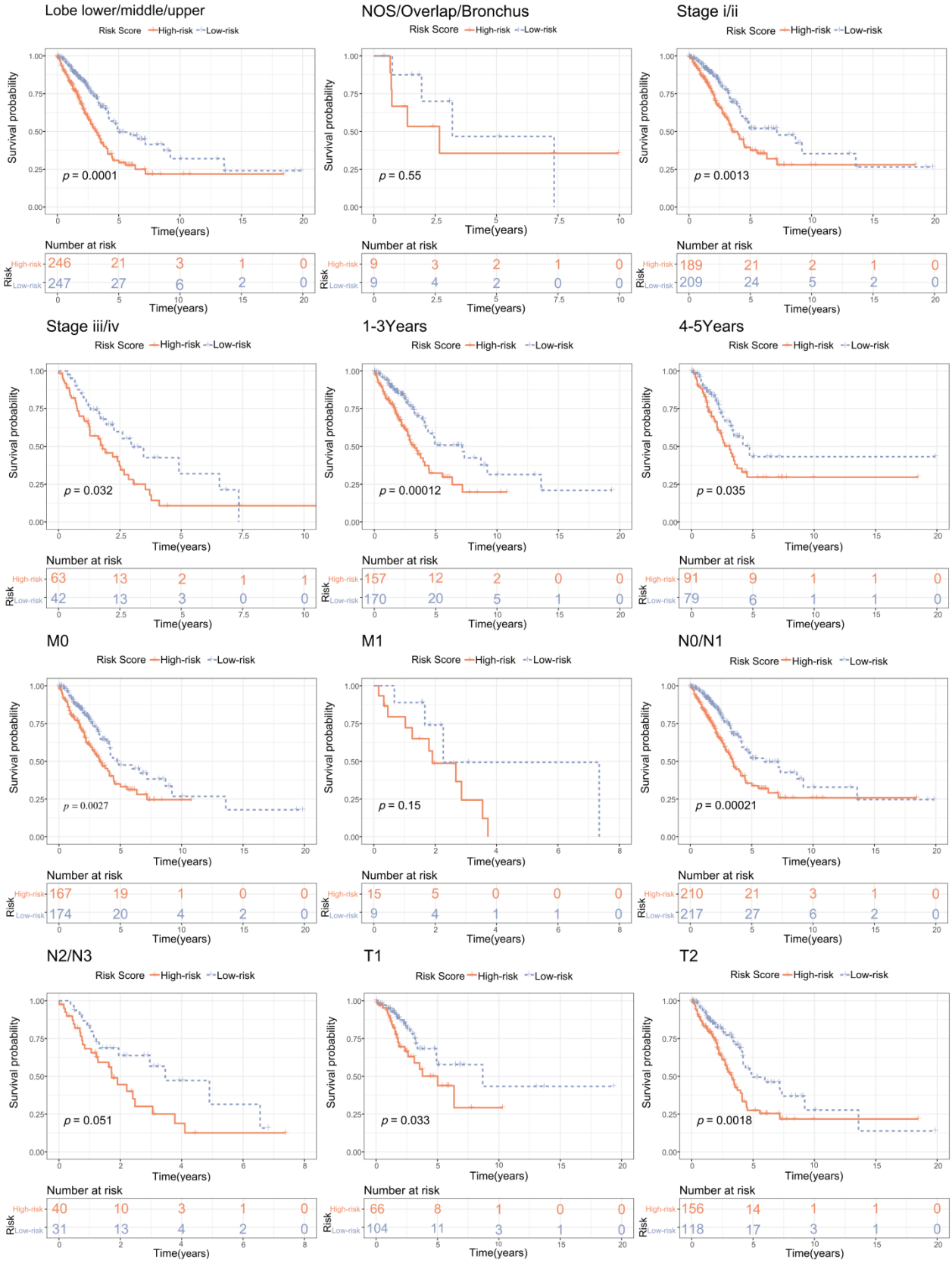

Figure 16. Stratified survival analysis with different clinical characteristics K-M curve.

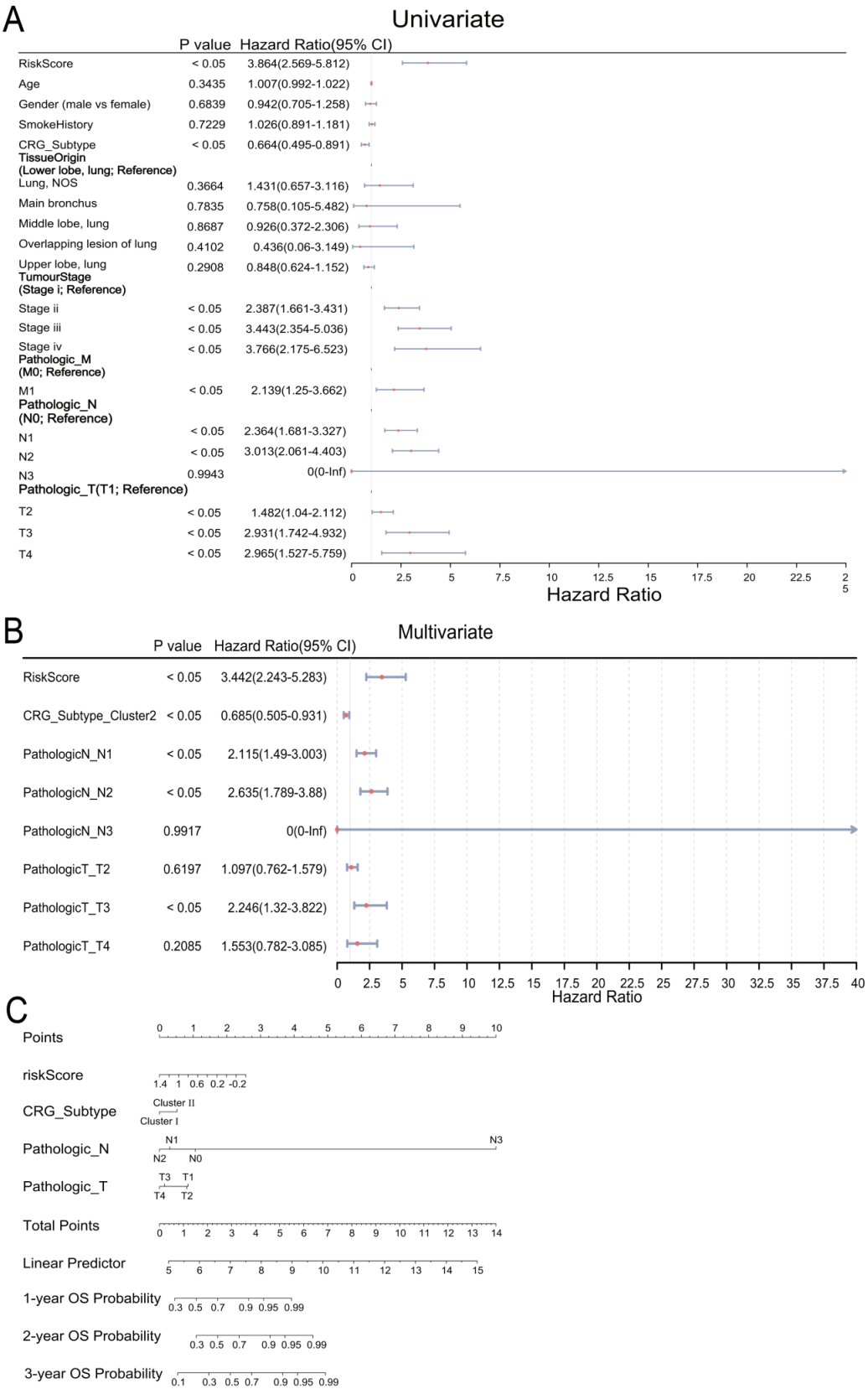

Figure 17. (A) Prognosis-related clinical indicators; (B) Multivariate Cox regression analysis; (C) Nomogram.

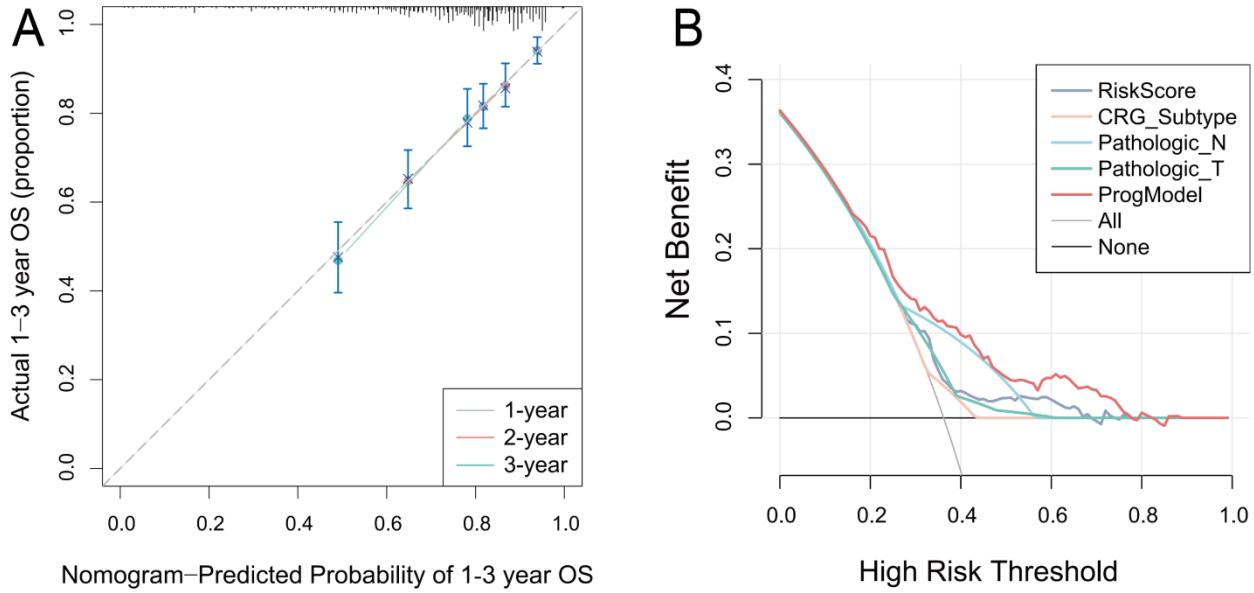

Figure 18. (A) Calibration curve; (B) DCA curves.

3.7.4. GO, KEGG enrichment analysis

For GO enrichment, the top ten categories of BP reveal the relevant biological processes mainly involved in glucuronidation. The first ten categories of CC are mainly cellular components of synaptic membrane and blood microparticles. The first ten categories of MF are hormone activity, glucuronosyltransferase cellular activity, monocarboxylate binding, receptor-ligand activity, gated channel activity, signal receptor activator activity, ligand-gated ion channel activity, gated ion channel activity, ligand-gated channel activity, channel activity and passive transmembrane transporter activity. GO enrichment results are displayed in Figure 20A.

The top ten pathways enriched to the KEGG pathway are neuroactive ligand-receptor interaction, bile secretion, steroid hormone biosynthesis, pentose and glucuronate interconversions, metabolism of xenobiotics by cytochrome P450, ascorbate and aldarate metabolism, retinol metabolism, Complement and coagulation cascades, drug metabolism-cytochrome P450, porphyrin metabolism. The results are exhibited in Figure 20B.

3.7.5. GSEA enrichment analysis

In GO enrichment, the top 10 categories of BP can be seen as the main biological processes involved in the various protein complexes. The first ten categories of CC are dynein complex, axoneme, an integral component of lumenal side of endoplasmic reticulum membrane, lumenal side of endoplasmic reticulum membrane, MHC protein complex, MHC class II protein complex, lumenal side of the membrane, axonemal dynein complex, lamellar body, multivesicular body. The first ten categories of MF are MHC class II protein complex binding, purinergic nucleotide receptor activity, nucleotide receptor activity, MHC protein complex binding, dynein intermediate chain binding, minus-end-directed microtubule motor activity, G protein-coupled purinergic nucleotide receptor activity,

sialic acid binding, chemokine binding, C-C chemokine binding. GO enrichment maps are demonstrated in Figure 21A–C.

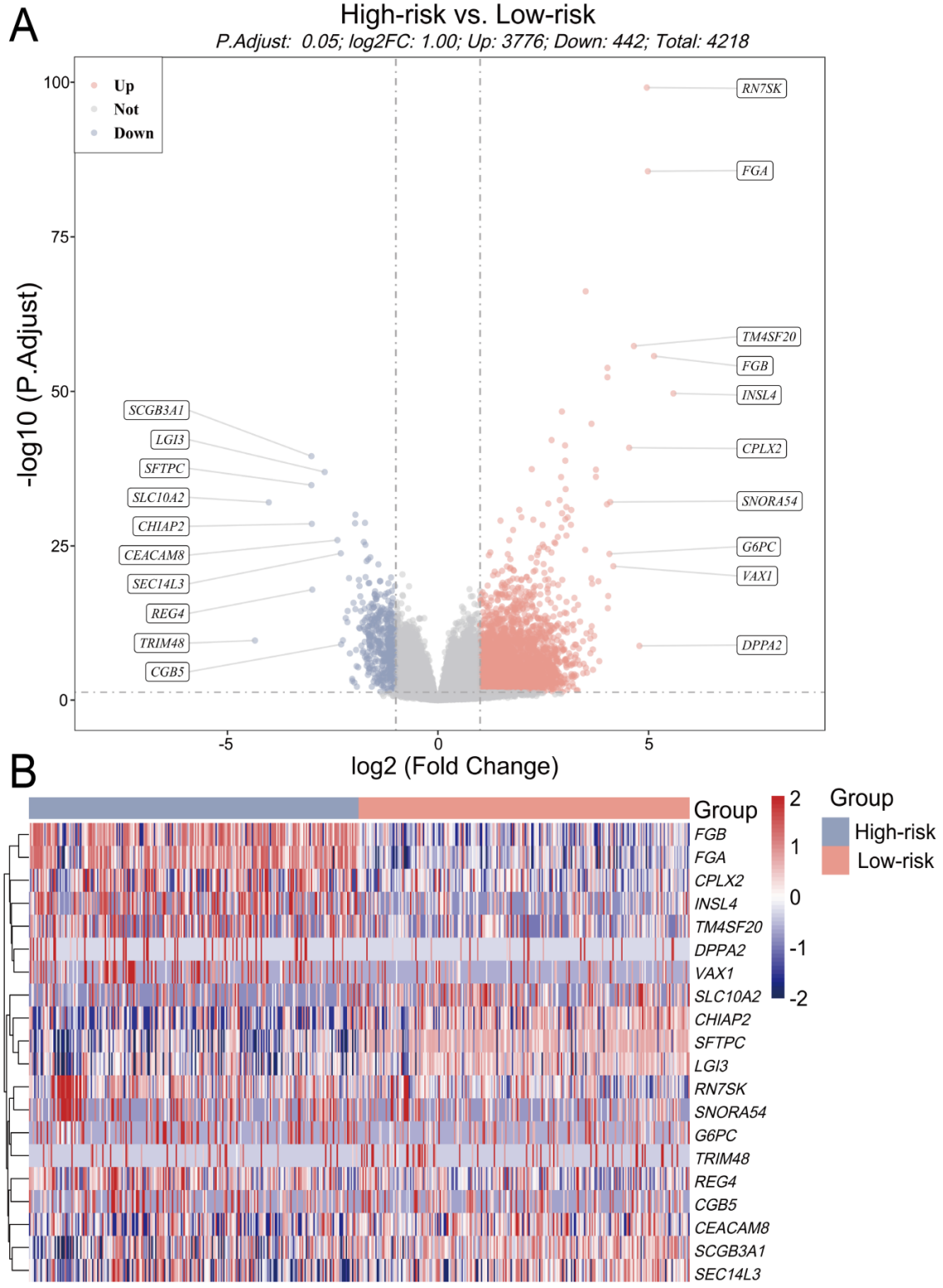

Figure 19. (A) Genetic differences expression volcano map; (B) Heatmap of differentially expressed genes (top 20).

Figure 20. (A) GO enrichment analysis; (B) KEGG enrichment analysis.

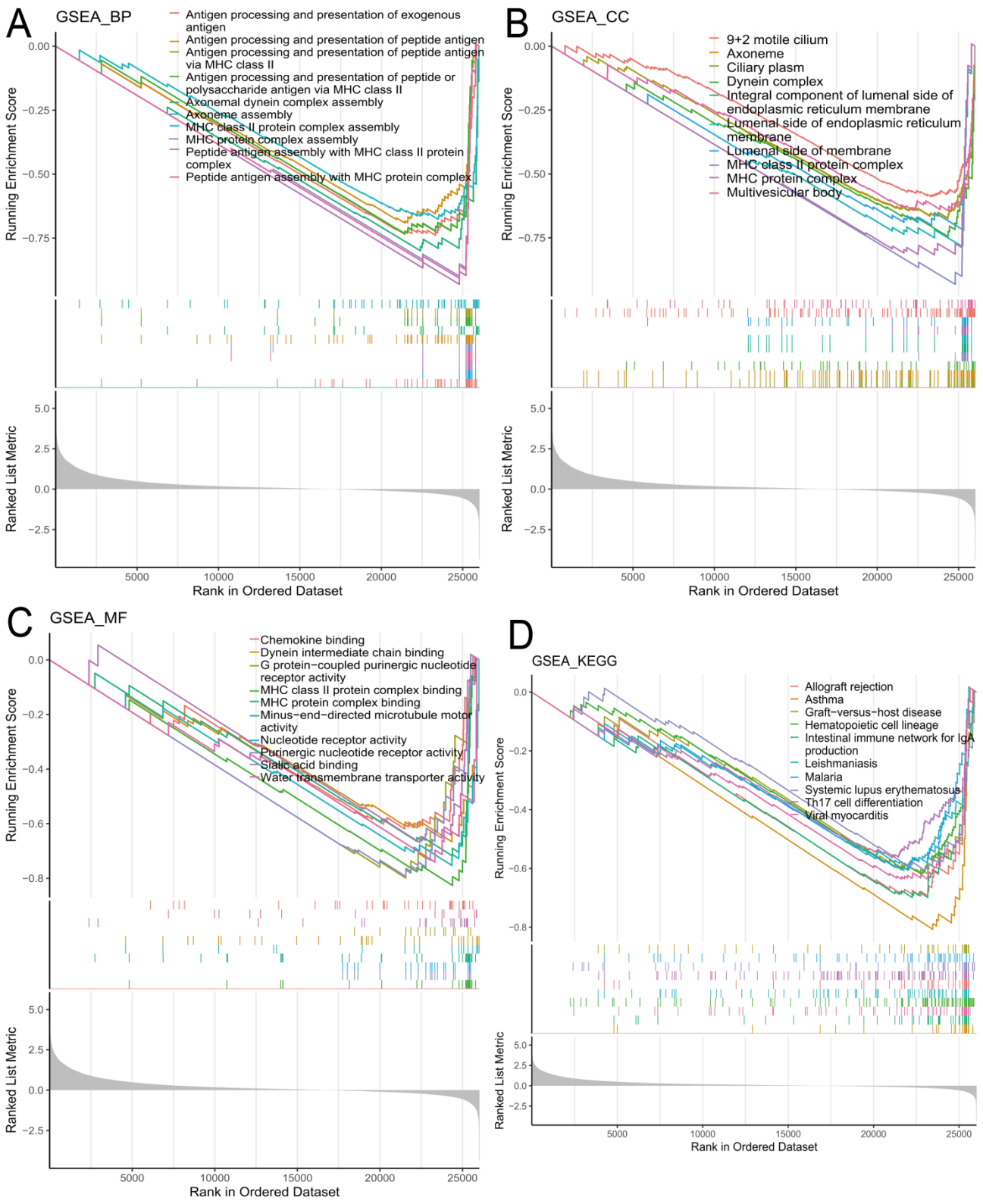

Figure 21. (A) GO Enrichment Analysis (BP); (B) GO Enrichment Analysis (CC); (C) GO enrichment analysis (MF); (D) KEGG enrichment analysis.

The top ten pathways enriched to the KEGG pathway are Asthma, Hematopoietic cell lineage,

Intestinal immune network for IgA production, Allograft rejection, Viral myocarditis, cell differentiation, Graft-versus-host disease, Leishmaniasis, Malaria, Systemic lupus erythematosus. The results are shown in Figure 21D.

3.8. Tumor microenvironment analysis and validation

3.8.1. Immune cell screening

We calculated the percentage abundance of tumor-infiltrating immune cells in each sample in TCGA-LUAD, and obtained eight immune cells of differentially expressed between high-risk and low-risk groups were obtained: Dendritic cells resting, Macrophages M0, Macrophages M1, Mast cells resting, Monocytes, NK cells resting, T cells CD4 memory activated and T cells CD4 memory resting, cells resting, T cells CD4 memory activated, T cells CD4 memory resting and the box plot of abundance ratio is manifested in Figure 22.

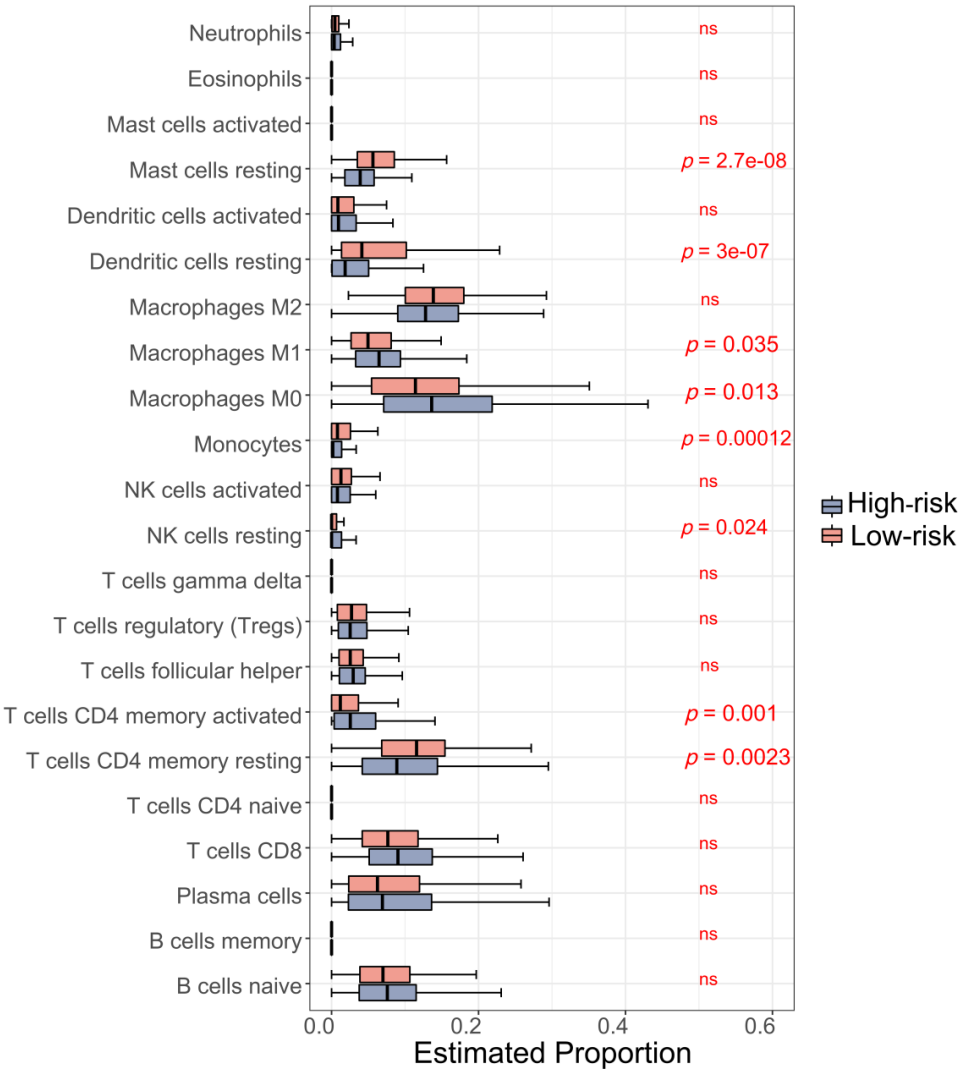

Figure 22. Box plot based on the abundance of immune cells between the high-risk and low-risk groups.

4. Discussion

Cancer-associated thrombosis (CAT) has been recognized since 1865, when Trousseau was first suggested to predispose cancer patients to thrombosis. Nowadays, venous thromboembolism (VTE) is widely recognized as a common complication of malignant tumors and a significant cause of death in cancer patients. The latest epidemiological data show that lung cancer has become the most deadly malignancy among men and women with cancer worldwide and also one of the malignancies with a high incidence of VTE. The incidence of VTE associated with lung cancer ranges from 3% to 13.9% [27], is influenced by several factors, and is most likely to occur within the first few months of diagnosis or when distant metastases occur. Patients with lung cancer with VTE have an approximately 50% increased risk of death compared to patients without VTE, and their 1-year survival rate is significantly shorter than those without VTE. Currently, the most common pathological type of lung cancer is adenocarcinoma of the lung, and the incidence is increasing significantly. Therefore, it is essential to identify and screen high-risk groups with VTE for early prevention and treatment of lung adenocarcinoma, which has a high incidence.

In this study, we identified coagulation-related subtypes in patients with LUAD based on coagulation pathways, developed a coagulation-related risk score prognostic model in the TCGA cohort, and validated the predictive value of the coagulation-related risk score in prognosis and immunotherapy. This study is the first bioinformatics analysis of coagulation-related genes in LUAD, and we found that risk models based on these genes showed excellent prognostic value in LUAD.

First, to investigate the genomic characteristics of CRGs in LUAD, we performed mutation and CNA copy number variation analysis on CRGs-associated tumor samples. The genetic analysis indicated a high frequency of copy number alternations of CRGs in the LUAD cohort. Patients with copy number alternations of CRGs had a poor prognosis. Meanwhile, copy number alternation was one of the reasons for the CRGs expression imbalance. *ACTB* and *ADCY1* were highly altered in LUAD patients and were associated with worse patient prognosis, suggesting that they may be driver genes in cancer development.

We performed unsupervised cluster analysis on the tumor samples and obtained two subtypes of CRGs, cluster I with 179 samples and cluster II with 347 samples. The T-SNE analysis and survival analysis of both subtypes clearly identified subtypes. Significant differences in subgroup survival based on CRG expression yielded preliminary evidence of the prognostic potential of coagulation-related genes in LUAD.

First, to explore the differential genes that were both variably expressed among subtypes and significantly associated with tumors, the two differential gene sets were intersected to obtain 65 CRG-DEGs, from which eight essential genes with prognostic significance were further selected to construct a risk model. A previous study reported an association between *LGI3* expression levels and cancer prognosis in brain cancer (astrocytoma), colorectal cancer and non-small cell lung cancer. In these cancer cohorts, lower *LGI3* expression was significantly associated with poor patient survival [29]. *HMGA* gene is a primary cancer gene, and *HMGA* protein plays a particular role in cell hyperplasia. Because the *HMGA* protein family can change the chromosomal structure, it can regulate a variety of destination gene expressions, which is usually considered a structural transcription factor [30]. Therefore, it may be related to the formation of thrombosis. *FGA* encodes the α subunit of the coagulation factor fibrinogen a component of the blood clot. At the same time, some studies have shown that *FGA* is an indicator of targeted therapy for EGFR-mutated lung adenocarcinoma [31].

The historical research on these genes is consistent with our research and proves the accuracy of our research.

Then we calculated the risk scores, and the samples were divided into high and low-risk groups according to the median risk scores. The risk models were evaluated and validated with log-rank test $P < 0.05$ for K-M curves and AUC > 0.6 for ROC curves, indicating good risk models.

Next, we analyzed the correlation between risk scores and nine clinical characteristics (coagulation subtype, age, sex, tissue origin, tumor stage, smoking status, pathological stage M, pathological stage N and pathological stage T), and the results showed that the correlation was not significant for tissue origin, pathological stage M, coagulation subtype and age, and was not significant for tumor stage, sex, smoking status, pathological stage N and pathological stage T, with $P < 0.05$. Survival analysis of 19 subgroups of clinical characteristics showed that the risk model could be applied to different clinicopathological characteristics except for cluster I, age <= 60, pathological stage M1 and tissue origin NOS/Overlap/Bronchus with log-rank test $p > 0.05$, and $p < 0.05$ for each subgroup. Univariate cox analysis in constructing the independent prognostic model showed that risk score, coagulation subtype, tumor stage and pathological MNT stage were associated with the survival prognosis of patients. The prognostic model was then constructed by multifactorial cox analysis of the clinical characteristics associated with prognosis. The prognostic model was then evaluated using calibration curves and decision curves, and the results showed good predictive performance.

Next, differential gene expression analysis was performed for the high-risk and low-risk groups, and 3776 up-regulated genes and 442 down-regulated genes were obtained. The GO enrichment analysis of the differential genes yielded 197 categories of BP (biological process), 38 categories of CC (cellular component) and 90 categories of MF (molecular function). The KEGG pathway enriched 261 pathways. Two hundred twenty-two categories of BP (biological process), 34 categories of CC (cellular component) and 19 categories of MF (molecular function) were enriched by GO in the GSEA enrichment analysis. Among them, BP is mainly involved in the related biological processes of various protein complexes, and its process may be related to the formation of thrombus. Thirty-eight pathways were enriched by the KEGG pathway. The enriched cell differentiation is related to the generation of blood vessels [32], which may lead to the development of cancer and the formation of thrombosis.

The tumor samples were then analyzed for immune infiltrating cells, and eight immune cells were differentially expressed between the high-risk and low-risk groups.

In conclusion, this study identifies prognostic factors associated with coagulation in LUAD that can help optimize risk stratification and individualized management of LUAD patients and explore the underlying molecular mechanisms of LUAD. There were some limitations in this study. In the future, we prepare to combine the results of the analysis of immune cells with single-cell technology and combine it with spatial transcriptomics to deepen the association of immune cells with coagulation-related genes.

Acknowledgments

This work is financially supported by the National Natural Science Foundation of China (NSFC, 31770999; 21173014) and the Beijing Natural Science Foundation (No. 2202002).

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J. Clin.*, **68** (2018), 394–424. https://doi.org/10.3322/caac.21492
2. W. Chen, R. Zheng, P. D. Baade, S. Zhang, H. Zeng, F. Bray, et al., Cancer statistics in China, 2015, *CA Cancer J. Clin*., **66** (2016), 115–132. https://doi.org/10.3322/caac.21338
3. D. Yang, Y. Liu, C. Bai, X. Wang, C. A. Powell, Epidemiology of lung cancer and lung cancer screening programs in China and the United States, *Cancer Lett*., **468** (2020), 82–87. https://doi.org/10.1016/j.canlet.2019.10.009
4. R. Ruiz-Cordero, W. P. Devine, Targeted therapy and checkpoint immunotherapy in lung cancer, *Surg. Pathol. Clin*., **13** (2020), 17–33. https://doi.org/10.1016/j.path.2019.11.002
5. J. Vansteenkiste, L. Crinò, C. Dooms, J. Y. Douillard, C. Faivre-Finn, E. Lim, et al., 2nd ESMO consensus conference on lung cancer: early-stage non-small-cell lung cancer consensus on diagnosis, treatment and follow-up, *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, **25** (2014), 1462–1474. https://doi.org/10.1093/annonc/mdu089
6. F. R. Hirsch, P. A. Bunn Jr, Adjuvant TKIs in NSCLC: what can we learn from RADIANT, *Nat. Rev. Clin. Oncol.*, **12** (2015), 689–690. https://doi.org/10.1038/nrclinonc.2015.202
7. S. Sampath, Treatment: Radiation therapy, in *Lung Cancer*, Springer, **170** (2016), 105–118. https://doi.org/10.1007/978-3-319-40389-2_5
8. M. Ahn, J. M. Sun, S. H. Lee, J. S. Ahn, K. Park, EGFR TKI combination with immunotherapy in non-small cell lung cancer, *Expert Opin. Drug Saf.*, **16** (2017), 465–469. https://doi.org/10.1080/14740338.2017.1300656
9. J. F. Gainor, A. M. Varghese, S. H. Ignatius Ou, S. Kabraji, M. M. Awad, R. Katayama, et al., ALK rearrangements are mutually exclusive with mutations in EGFR or KRAS: an analysis of 1683 patients with non-small cell lung cancer, *Clin. Cancer Res.*, **19** (2013), 4273–4281. https://doi.org/10.1158/1078-0432.CCR-13-0318
10. Z. Wang, K. S. Embaye, Q. Yang, L. Qin, C. Zhang, L. Liu, et al., Establishment and validation of a prognostic signature for lung adenocarcinoma based on metabolism-related genes, *Cancer Cell Int.*, **21** (2021). https://doi.org/10.1186/s12935-021-01915-x
11. P. E. Serrano, S. Parpia, L. A. Linkins, L. Elit, M. Simunovic, L. Ruo, et al., Venous thromboembolic events following major pelvic and abdominal surgeries for cancer: A prospective cohort study, *Ann. Surg. Oncol.*, **25** (2018), 3214–3221. https://doi.org/10.1245/s10434-018-6671-7
12. A. Falanga, M. Marchetti, L. Russo, The mechanisms of cancer-associated thrombosis, *Thromb. Res.*, **135** (2015), 8–11. https://doi.org/10.1016/S0049-3848(15)50432-5
13. L. Bao, S. Zhang, X. Gong, G. Cui, Trousseau syndrome related cerebral infarction: Clinical manifestations, laboratory findings and radiological features, *J. Stroke Cerebrovasc. Dis.*, **29** (2020), 104891. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.104891

14. Y. Li, S. Wei, J. Wang, L. Hong, L. Cui, C. Wang, Analysis of the factors associated with abnormal coagulation and prognosis in patients with non-small cell lung cancer (in Chinese), *Zhongguo fei ai za zhi*, **17** (2014), 789–796. https://doi.org/10.3779/j.issn.1009-3419.2014.11.04
15. M. J. Goldman, M. J. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, et al., Visualizing and interpreting cancer genomics data via the Xena platform, *Nat. Biotechnol.*, **38** (2020), 675–678. https://doi.org/10.1038/s41587-020-0546-8
16. Q. He, J. Yang, Y. Jin, Immune infiltration and clinical significance analyses of the coagulation-related genes in hepatocellular carcinoma, *Briefings Bioinf.*, **23** (2022). https://doi.org/10.1093/bib/bbac291
17. C. Ren, J. Li, Y. Zhou, S. Zhang, Q. Wang, Typical tumor immune microenvironment status determine prognosis in lung adenocarcinoma, *Transl. Oncol.*, **18** (2022), 101367. https://doi.org/10.1016/j.tranon.2022.101367
18. A. Mayakonda, D. C. Lin, Y. Assenov, C. Plass, H. P. Koeffler, Maftools: efficient and comprehensive analysis of somatic variants in cancer, *Genome Res.*, **28** (2018), 1747–1756. https://doi.org/10.1101/gr.239244.118
19. Y. Zhou, T. O. Sharpee, Using global t-SNE to preserve intercluster data structure, *Neural Comput.*, **34** (2022), 1637–1651. https://doi.org/10.1162/neco_a_01504
20. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biol.*, **15** (2014), 550. https://doi.org/10.1186/s13059-014-0550-8
21. The Gene Ontology Consortium, Gene Ontology Consortium: going forward, *Nucleic Acids Res.*, **43** (2015), 1049–1056. https://doi.org/10.1093/nar/gku1179
22. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, *OMICS: J. Integr. Biol.*, **16** (2012), 284–287. https://doi.org/10.1089/omi.2011.0118
23. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *J. Stat. Software*, **33** (2010), 1–22.
24. P. J. Heagerty, T. Lumley, M. S. Pepe, Time-dependent ROC curves for censored survival data and a diagnostic marker, *Biometrics*, **56** (2000), 337–344. https://doi.org/10.1111/j.0006-341x.2000.00337.x
25. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data, *Bioinformatics*, **32** (2016), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313
26. B. Chen, M. S. Khodadoust, C. L. Liu, A. M. Newman, A. A. Alizadeh, Profiling tumor infiltrating immune cells with CIBERSORT, in *Cancer Systems Biology*, Springer Nature, **1711** (2018), 243–259. https://doi.org/10.1007/978-1-4939-7493-1_12
27. H. T. Sørensen, L. Mellemkjaer, J. H. Olsen, J. A. Baron, Prognosis of cancers associated with venous thromboembolism, *N. Engl. J. Med.*, **343** (2000), 1846–1850. https://doi.org/10.1056/NEJM200012213432504
28. Y. B. Yu, J. P. Gau, C. Y. Liu, M. Yang, S. Chiang, H. Hsu, et al., A nation-wide analysis of venous thromboembolism in 497,180 cancer patients with the development and validation of a risk-stratification scoring system, *Thromb. Haemostasis*, **108** (2012), 225–235. https://doi.org/10.1160/TH12-01-0010

29. N. S. Kwon, K. J. Baek, D. S. Kim, H. Y. Yun, Leucine-rich glioma inactivated 3: Integrative analyses reveal its potential prognostic role in cancer, *Mol. Med. Rep.*, **17** (2018), 3993–4002. https://doi.org/10.3892/mmr.2017.8279
30. A. P. Wolffe, Architectural transcription factors, *Science*, **264** (1994), 1100–1101. https://doi.org/10.1126/science.8178167
31. Z. Shang, X. Niu, K. Zhang, Z. Qiao, S. Liu, X. Jiang, et al., FGA isoform as an indicator of targeted therapy for EGFR mutated lung adenocarcinoma, *J. Mol. Med.*, **97** (2019), 1657–1668. https://doi.org/10.1007/s00109-019-01848-z
32. M. Majesky, Vascular development, *Arterioscler. Thromb. Vasc. Biol.*, **38** (2018), 17–24. https://doi.org/10.1161/ATVBAHA.118.310223

Appendix

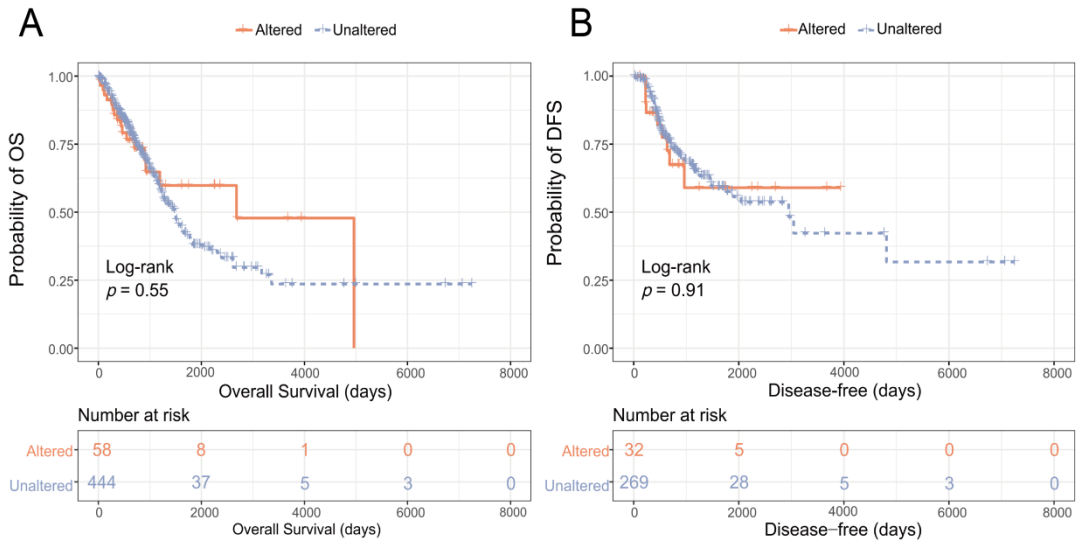

Figure A1. (A) Kaplan-Meier curve for disease-free survival analysis; (B) Kaplan-Meier Curve for Overall Survival Analysis.

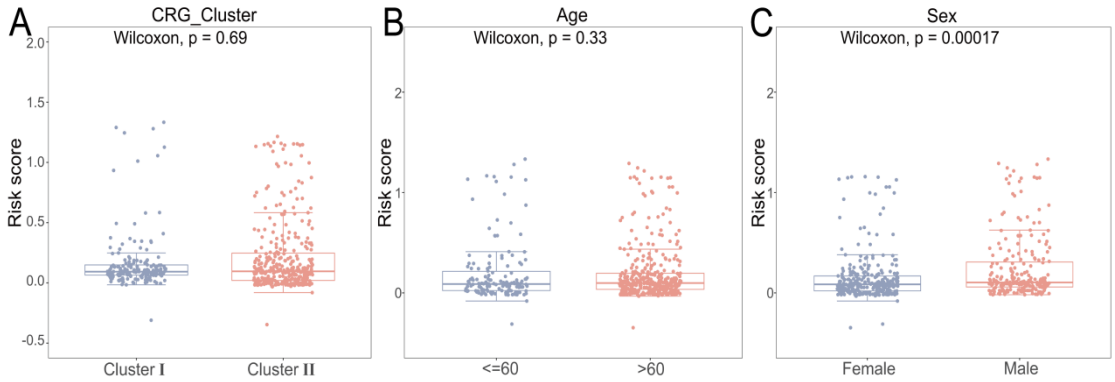

Figure A2. Correlation analysis of different clinical characteristics.

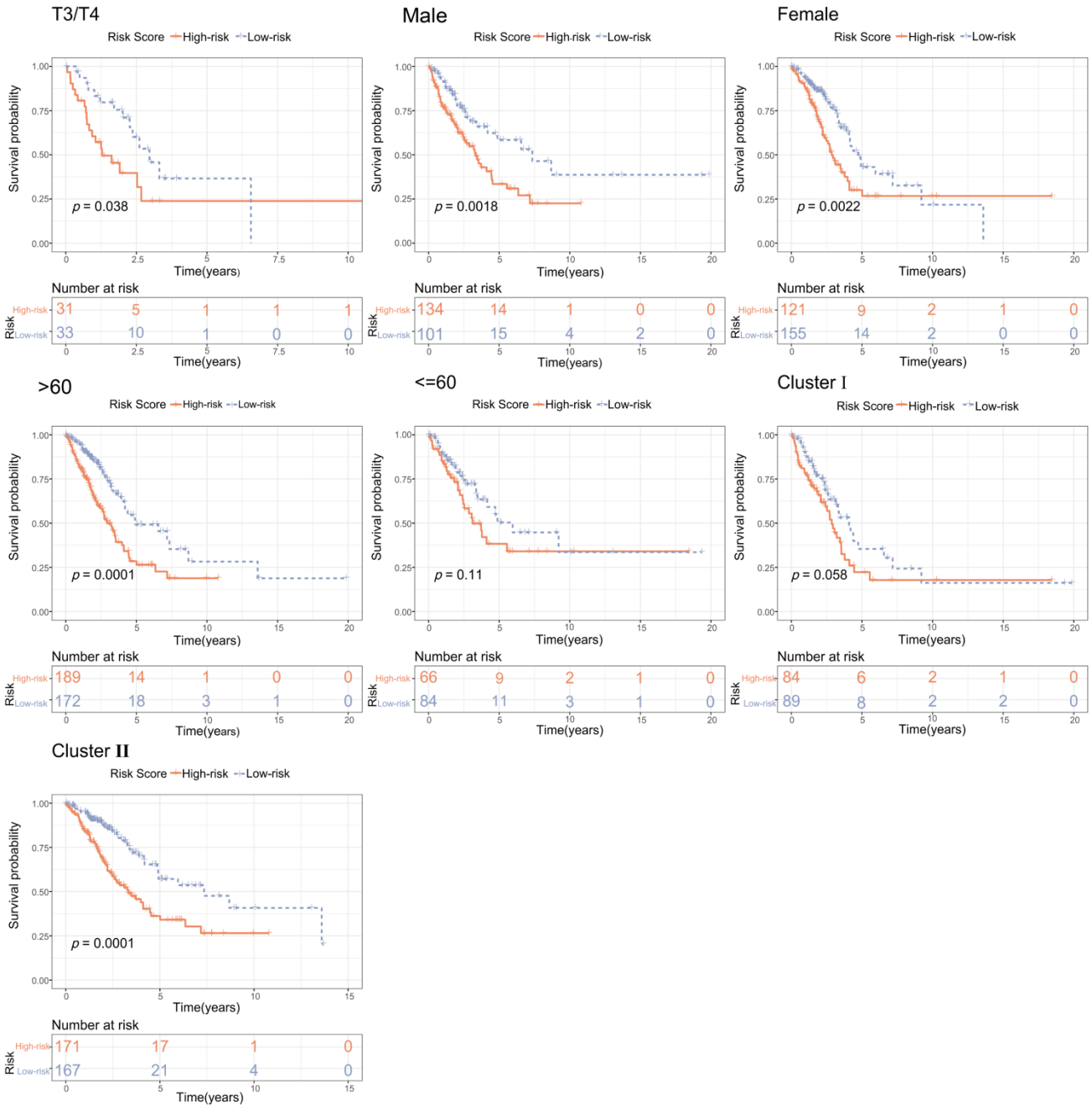

Figure A3. Stratified survival analysis with different clinical characteristics K-M curve.