Research Paper

A novel classification method for NSCLC based on the background interaction network and the edge-perturbation matrix

Yuan Tian[1,*], Caiqing Zhang[2,*], Wanru Ma[3,*], Alan Huang[4], Mei Tian[5], Junyan Zhao[6], Qi Dang[7], Yuping Sun[7]

[1]Somatic Radiotherapy Department, Shandong Second Provincial General Hospital, Shandong Provincial ENT Hospital, Jinan, Shandong 250023, PR China
[2]Department of Respiratory and Critical Care Medicine, Shandong Second Provincial General Hospital, Shandong Provincial ENT Hospital, Shandong University, Jinan, Shandong 250023, PR China
[3]Department of Blood Transfusion, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, PR China
[4]Department of Oncology, Jinan Central Hospital, The Hospital Affiliated with Shandong First Medical University, Jinan, Shandong 250013, PR China
[5]Respiratory Department, Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, Shandong 250014, PR China
[6]Nursing Department, The First Affiliated Hospital of Shandong First Medical University and Shandong Provincial Qianfoshan Hospital, Jinan, Shandong 250014, PR China
[7]Phase I Clinical Trial Center, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, Shandong 250012, PR China
[*]Equal contribution

Correspondence to: Yuping Sun, Yuan Tian; **email:** 199057020185@email.sdu.edu.cn, tytytianyuan@bjmu.edu.cn
Keywords: multi-dimensional, multi-omics, gene interaction, perturbation subtype, NSCLC
Received: January 19, 2022 **Accepted:** March 28, 2022 **Published:** April 9, 2022

ABSTRACT

The biological functional network of tumor tissues is relatively stable for a period of time and under different conditions, so the impact of tumor heterogeneity is effectively avoided. Based on edge perturbation, functional gene interaction networks were used to reveal the pathological environment of patients with non-small cell carcinoma at the individual level, and to identify cancer subtypes with the same or similar status, and then a multi-dimensional and multi-omics comprehensive analysis was put into practice. Two edge perturbation subtypes were identified through the construction of the background interaction network and the edge-perturbation matrix (EPM). Further analyses revealed clear differences between those two clusters in terms of prognostic survival, stemness indices, immune cell infiltration, immune checkpoint molecular expression, copy number alterations, mutation load, homologous recombination defects (HRD), neoantigen load, and chromosomal instability. Additionally, a risk prediction model based on TCGA for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) was successfully constructed and validated using the independent data set (GSE50081).

INTRODUCTION

Smoking cessation and advancements in early detection and treatment [1], have resulted in a continuous decline in cancer death rates worldwide from 1991 to 2018, a total decline of 31%, particularly for lung cancer [2]. Lung cancer, on the other hand, remained the leading cause of cancer

death in 2020, with an estimated 1.8 million deaths (18%) [3]. As is well known, NSCLC is the leading cause of cancer death in patients with lung cancer [4]. Thus, further investigation of disease biology, pathological type, genotype, application of predictive biomarkers, and treatment improvements for NSCLC are required.

Tumor heterogeneity is prevalent and plays a significant role in the progression of the disease [5, 6]. This heterogeneity can manifest itself in the uneven distribution of tumor cell subpopulations across the tumor (spatial heterogeneity) [7], or in the temporal variation in the molecular composition of cancer cells (temporal heterogeneity) [5, 8], both of which present difficulties for clinical research [9]. It has been suggested that expression profiles measured at different time points or under different conditions may exhibit significant temporal heterogeneity [5, 8, 9]. However, the biological functional network of tumor tissues remains relatively stable over time and under a variety of conditions, effectively avoiding the effects of tumor heterogeneity [10–14].

Based on edge perturbation [15–18], functional gene interaction networks were used to deduce the pathological environment of NSCLC patients at the individual level [19–22], and to identify cancer subtypes with the same or similar status, followed by a multi-dimensional and multi-omics comprehensive analysis for validation [23–26].

MATERIALS AND METHODS

Published data sets

Lung adenocarcinoma (LUAD) and Lung squamous cell carcinoma (LUSC) expression data of the TCGA database, including clinical characteristics and survival information, were downloaded from the UCSC Xena website (https://xenabrowser.net/datapages/). Copy number variation (CNV) information was downloaded by R package RTCGA, while mutation data was downloaded by R package TCGAbiolinks. 986 (497/489) cancer samples with expression data, clinical characteristics, and survival information were prepared for the subsequent analysis. The basic clinical information of cancer samples from the TCGA dataset was displayed in Table 1. The flow diagram was provided in (Supplementary Figure 1).

The expression data of 288 lung control samples from the GTEX database were gotten from the UCSC Xena website (https://xenabrowser.net/datapages/). 181 NSCLC samples with survival information of the GSE50081 data cohort were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) and used as the validation set for model verification.

Data preprocessing

In order to keep the data consistency, 70% of the genes with 0 expression level in the three data cohorts (TCGA-LUAD, TCGA-LUSC, GTEX-LUNG) were filtered out first, and then the R package sva was used for batch calibration.

Construction of the background interaction network and edge-perturbation matrix

The Cytoscape plug-in ReactomeFIPlugIn was used to download the gene affiliation data of pathway coding in Reactome, and the functional gene interaction network in the database was constructed based on the existing gene or protein interaction information. Specifically, the background interaction network is a gene interaction network based on the pathway in Reactome, including protein-protein interaction, gene co-expression, protein domain interaction, GO (Gene Ontology) annotation, and proteins interaction data obtained from the text data mining analyses. The ReactomeFIPlugIn plug-in was used to download all genetic interaction pairs in the pathway and then merged into a large background network. The construction process of the edge disturbance matrix mainly included the following three steps [15].

First, according to the inconsistency of the expression levels of genes in the background interaction network between cancer and normal samples, the ranks of gene expression in cancer and normal samples were obtained separately (ranked from small to large, the smaller value of the expression level ranked in the front of the queue, while the bigger value of the expression level ranked in the back of the queue), and then the gene expression matrix was converted into a gene expression rank matrix.

Second, according to the interaction relationship of gene pairs in the background of interaction network, if two genes interaction existed, there would be an edge in the network that connected the two genes, then the difference between the ranks of this edge in the two genes was calculated. In this way, the delta rank matrix of each edge among all the samples were obtained. The calculation formula was displayed in the followings:

$$\delta_{e,s} = r_{i,s} - r_{j,s}$$

In the formula, "$r_{i,s}$" represents the rank of gene "i" in sample "s", "$\delta_{e,s}$" represents the delta rank of edge "e" in sample "s", and gene "i" and "j" are connected by the edge "e".

Table 1. Clinical characteristic of TCGA-NSCLC patients.

Parameter	Subtype	TCGA (LUAD + LUSC)
Age	≥60	240
	<60	746
Stage	I	506
	II	274
	III	163
	IV	32
	Unknown	11
Gender	Female	396
	Male	590
Smoke_year	1	89
	2	249
	3	210
	4	403
	5	9
	Unknown	26
EGFR_mutation	NO	458
	YES	100
	Unknown	428
M	M0	733
	M1	31
	MX	214
	Unknown	8
N	N0	634
	N1	220
	N2	109
	N3	7
	NX	15
	Unknown	1
T	T1	277
	T2	553
	T3	112
	T4	41
	TX	3

Third, the average expression level of genes in normal samples was converted into a normal sample gene expression rank matrix, and the average delta rank matrix (Benchmark delta rank vector) of normal samples was established according to the background interaction network. Then, the delta rank matrix and the average delta rank matrix of the normal samples were used to construct the edge-perturbation matrix. The calculation formula was displayed in the followings:

$$\Delta_{e,s} = \delta_{e,s} - \overline{\delta_e}$$

In the formula, "$\delta_{e,s}$" represents the delta rank value of the edge "e" in the sample "s", and $\overline{\delta}_e$ represents the average delta rank of the delta rank matrix and the normal sample, which was the eigenvalue of the edge in the edge disturbance matrix.

Finally, the specific edge perturbation matrix in the cancer sample was screened. According to the Kruskal-Wallis test, the difference of the edge disturbance matrix between the normal and the cancer sample was calculated, and the top 30,000 different edges would be selected according to the level of significant difference; At the same time, the standard deviations (SDs) of the

edge disturbance matrix in the cancer samples were calculated, and the top 30,000 different edges according to the SDs were selected. The intersection of the above two edges were considered as the specific edge perturbation matrix of the cancer samples, which was named as the characteristic edge. The feature logarithm conversion formula between cancer sample and feature sample was listed as follows:

$$features = \log_2(\Delta_{es} + 1)$$

In the formula, "Δ_{es}" is the eigenvalue of the edge in the edge disturbance matrix.

Clustering and survival analysis of features of edge disturbance matrix

According to the feature edge of the edge perturbation matrix specific to the cancer sample, the R package ConsensusClusterPlus was used to perform the consistent clustering analysis. The distance used for clustering is spearman, the clustering method is pam, and 100 repetitions are performed to ensure the stability of the classification.

Log-rank test was used to explore the difference in survival time among subtypes, and R package survival was used to draw the KM survival curve of patients subtypes. The R package clusterRepro and independent data sets were used to verify the efficacy of clustering of the feature edge perturbation matrices. Then, the intra-group proportion (IGP) of each subtype would be calculated, which the larger IGP indicated the better the consistency in the clustering group.

Feature analysis of edge disturbance feature subtypes

Based on the known literature or calculated various feature indicators, statistical tests were used to explore the correlation between edge disturbance feature subtypes and known feature indicators. Tumor purity and ploidy were derived from TCGA data pan-cancer analysis [27].

The homologous recombination defect score (HRD score), neoantigen load, and genome alteration frequency were derived from previous studies on the analysis of immune characteristics of TCGA data [28]. Based on previous studies, the patient's stemness index (mRNAsi) and epigenetically regulated-mRNAsi (EREG-mRNAsi) were obtained [11].

The three indicators of chromosomal instability (LST score, TAI score, and LOH score) were derived from previous studies based on the correlation analyses between genome damage and homologous recombination defects [29]. The R package CIBERSORT was used to calculate the infiltration scores of 22 immune cell types. The expression level of immune checkpoint molecules is the RNAseq level of cancer samples currently used.

In the significance analysis between various values (expression, infiltration ratio, mutation count, etc.), the Wilcoxon test was used to compare the differences between the two sets of samples. In the graphical display, ns (no significance) represents $p > 0.05$, *represents $p \leq 0.05$, **represents $p \leq 0.01$, ***represents $p \leq 0.001$, and ****represents $p \leq 0.0001$.

Copy number variant (CNV) analyses

The GISTIC method was used to detect the common copy number alteration area in all samples based on the SNP6 CopyNumber segment data. The parameters of the GISTIC method were set as follows: Q ≤ 0.05 was taken as the significance standard of the alteration. 0.90 was adopted as the confidence level, when the peak interval was determined. The analysis was performed through the corresponding MutSigCV module of the online analysis tool GenePattern (https://cloud.genepattern.org/gp/pages/index.jsf) developed by Broad Research Institute.

TIDE (tumor immune dysfunction and exclusion) prediction

The TIDE (http://tide.dfci.harvard.edu/) analysis tool was developed by researchers from Harvard University and used to evaluate the clinical effects of immune checkpoint suppression therapy. A higher tumor TIDE prediction score was corresponded to a lower immune checkpoint suppression therapy efficacy and poor prognosis. The prediction for the prognosis of immune checkpoint inhibitor (ICI) treatment in this analysis was completed by the TIDE score.

Subtype-specific characteristic clusters and pathway enrichment of feature clusters

The unique biological functions and pathways of the subtype were analyzed by identifying the unique characteristic clusters of the subtype, and Z-score was used to standardize the characteristic values of the characteristic edges, and then the characteristic clusters were screened through the following steps.

The first step is to perform hierarchical clustering based on the normalized characteristic values of the characteristic edges, and the clustering method was "complete linkage". The number of clusters is 100, and

clusters with fewer than 30 characteristic edges would be filtered out. Second, for each remaining cluster, the percentage of "characteristic edges", whose "absolute value" of the perturbation mean was greater than 0.5 to all the "characteristic edges", would be calculated. Third, the cluster with an average percentage greater than 0.7 in a subtype would be considered as the characteristic cluster of that subtype, and all genes in the characteristic edge corresponding to the characteristic cluster would be used for pathway enrichment analysis. Online software Metascape (http://metascape.org) would be used for enrichment analysis of KEGG and Reactome pathways, setting the parameter $P < 0.01$.

Differential expression analysis and differential methylation site recognition

The R software package limma would be used for analyzing the expression profile and methylation level of the subtype samples. According to the fold change (|logFC|) and significant False Discovery Rate (FDR), the genes and methylation sites that were differentially expressed in the subtype samples were screened.

Prognostic analysis

Univariate Cox regression analysis would be used for determining the hazard ratio (HR) and prognostic significance of different expressed genes, and genes with $p < 0.01$ would be screened as prognostic-related genes. The Kaplan-Meier method was used to generate the survival curve for prognostic analysis, and the log-rank test was used to determine the significance of the difference. The receiver operating characteristic curve (ROC) was used to evaluate the risk model's prediction of the score, and the area under the curve (AUC) was quantified by the R package survivalROC.

Ethics statement

No interaction with human subjects of the study was involved, no ethical issues were encountered, and no ethical approval was needed.

RESULTS

Background interaction network construction and gene expression extraction

The Cytoscape plug-in ReactomeFIPlugIn was used to download the gene affiliation data of pathway encoding in Reactome. According to the existing gene or protein interaction information, including protein-protein interaction, gene co-expression, protein domain interaction, GO annotation and protein interaction data based on text mining, integrating all interaction information, the functional background interaction network would be constructed. A total of 7,360 nodes and 169,710 edges were included in the background interaction network.

In the three data sets (TCGA-LUAD, TCGA-LUSC, and GTEX-LUNG), 70% of the genes whose expression levels were 0 were filtered out, and then the R package sva was used for batch correction. The two datasets TCGA-LUAD and TCGA-LUSC were merged as the expression profile of cancer samples (26209 genes ×986 samples), and the GTEX-LUNG dataset was used as the expression profile of normal samples (26209 genes ×288 samples).

After filtering out the nodes in the background interaction network that were not within the range of 26209 genes, a new background interaction network was constructed, which included 6327 nodes and 153314 edges, which would be used for subsequent calculation of edge disturbances Feature matrix. The interaction network was closely connected, and most of the nodes in the network had a high degree. According to the degree of nodes in the background interaction network, the nodes were sorted from large to small.

Perturbation matrix (edge-perturbation matrix) construction and feature extraction

The perturbation of the edge in the background interaction network would inevitably lead to the alteration of the interaction relationship in the network, and the perturbation of the gene in the network to the edge could be reasonably used for simulating and revealing the pathological environment at the individual level. In order to measure the degree of disturbance of the background interaction network at the level of a single sample, based on the difference in expression fluctuations between cancer and normal samples, the Edge-Perturbation Matrix (EPM) was constructed separately. In order to compare the difference in EPM between cancer and normal samples, 1000 features were randomly selected, and logarithmic transformation was performed. It was found that the feature values of cancer samples were significantly higher than those of normal samples (Supplementary Figure 2). This showed that the degree of edge perturbation in cancer samples was greater, indicating that the degree of perturbation in the background network of cancer samples was much more obvious than that of normal samples. This provided reliable evidence for the subsequent use of EPM to reveal the heterogeneity among NSCLC samples.

In order to further perform feature extraction in this study, we used the Kruskal-Wallis test to calculate the difference of edge perturbations between cancer and

normal samples, and calculated the Standard Deviations (SDs) of the edge perturbation matrices in cancer samples. The top 30,000 different edges of the above two methods were selected respectively, and the intersection edges (N = 5468) were selected as the feature edges of the edge perturbation matrix in the cancer sample for subsequent analyses.

Clustering and survival analyses of edge disturbance matrix features

Based on the 5468 feature edges extracted in the previous step, consistent clustering analyses were put into practice based on their feature values. 986 cancer samples were divided into two different subtypes (Figure 1A and 1B). These two subtypes were named cluster 1 (N = 406) and cluster 2 (N = 580), which displayed significant differences in prognosis from each other (Figure 1C). The distance used for clustering was spearman, the clustering method was pam, and 100 repetitions were performed. The hierarchical clustering method was used to perform cluster analysis on the extracted features, and it was found that there were obvious specific feature in subtypes (Figure 1D).

In order to verify the clustering performance of gene interaction perturbation, we collected a set of independent expression data sets (GSE50081) from reported studies, and used the R package clusterRepro to calculate the intra-group proportion (IGP) of each subtype. The results showed that both cluster 1 and cluster 2 had higher IGP values (Figure 1E), which indicated that the clustering consistency of the cluster analysis based on the edge perturbation matrix in this study was better.

Comparative analysis of edge perturbation feature subtypes

Genomic heterogeneity indicators were obtained from reported studies, and statistical tests were used to explore their differences in edge perturbation feature subtypes. The results showed that the cluster 2 subtype with poor prognosis displayed higher tumor purity and genome ploidy compared with cluster 1 (Supplementary Figure 3A and 3B).

Based on previous analyses of TCGA samples, the transcriptome (mRNAsi) and epigenetic regulatory (EREG-mRNAsi) index of NSCLC samples were obtained. Through the evaluation of the stemness index in subtypes, it was found that the cluster 2 displayed a higher stemness index (Supplementary Figure 3C and 3D). The clinical characteristics of the two subtype samples were statistically analyzed. Fisher's exact test results were displayed in the form of Stage, T, N staging and Age ($p < 0.05$, Supplementary Figure 3E–3J).

The infiltration scores of 22 immune cell types were calculated by the R package CIBERSORT, and the differences among subtype samples were further explored. The results showed that immune cells such as B cells naive, Plasma cells and Mast cells resting were significantly different in the two subtypes ($p < 0.05$, Figure 2A and 2B). On the other hand, the expression differences of important immune checkpoint molecules of the two subtypes were also analyzed, and it was found that multiple immune checkpoint molecules, such as PDCD1, CD4, and LAG3, were significantly different between the two subtypes (Figure 2C). Furthermore, based on online verification (http://gepia.cancer-pku.cn/detail.php?gene), CD276, CXCR4, and BTLA were found to be significantly related to the prognosis of LUAD, while CCL2 was significantly related to the prognosis of LUSC.

TIDE was used to evaluate the clinical efficiency of two subtypes receiving immune checkpoint inhibitors. There was no significant difference in TIDE scores between the two subtypes, but there were significant differences for the expression of IFNG immune checkpoints (Figure 2D and 2E).

Comparing the edge perturbation feature subtypes with the reported subtypes of NSCLC, the fish results showed that the distribution of cancer subtypes corresponding to the edge perturbation feature subtypes were different (Figure 2F and 2G). To further explore the molecular differences between the two sample sets, the R package maftools was used for somatic mutations analysis. Among the top 20 genes with mutation frequencies in the two groups, the distribution of the 15 common mutation genes that appeared in both of them were shown in (Figure 3A and 3B). In order to observe the proportion of high-frequency mutation genes in the subtypes in detail, a line chart of the proportions of these 15 genes between the two subtypes was drawn and displayed in (Figure 3C). The distribution of CNV in the two sets of samples was shown in (Figure 3D and 3E), which indicated that the CNV frequency of the two sets was significantly different ($p < 0.05$, Figure 3F).

In order to further investigate the correlation between the innate immune escape mechanism and subtypes, we compared some potential factors that determine tumor immunogenicity of the two subtypes, including tumour mutational load (TML), homologous recombination deficiency (HRD), and neoantigen load, chromosome instability (Supplementary Figure 4). The functional enrichment results of the two subtype-specific clusters were further analyzed. The cluster 1 was found to be significantly enriched in those pathways such as Peptide chain elongation and Translation initiation complex formation, while the cluster 2 was significantly enriched

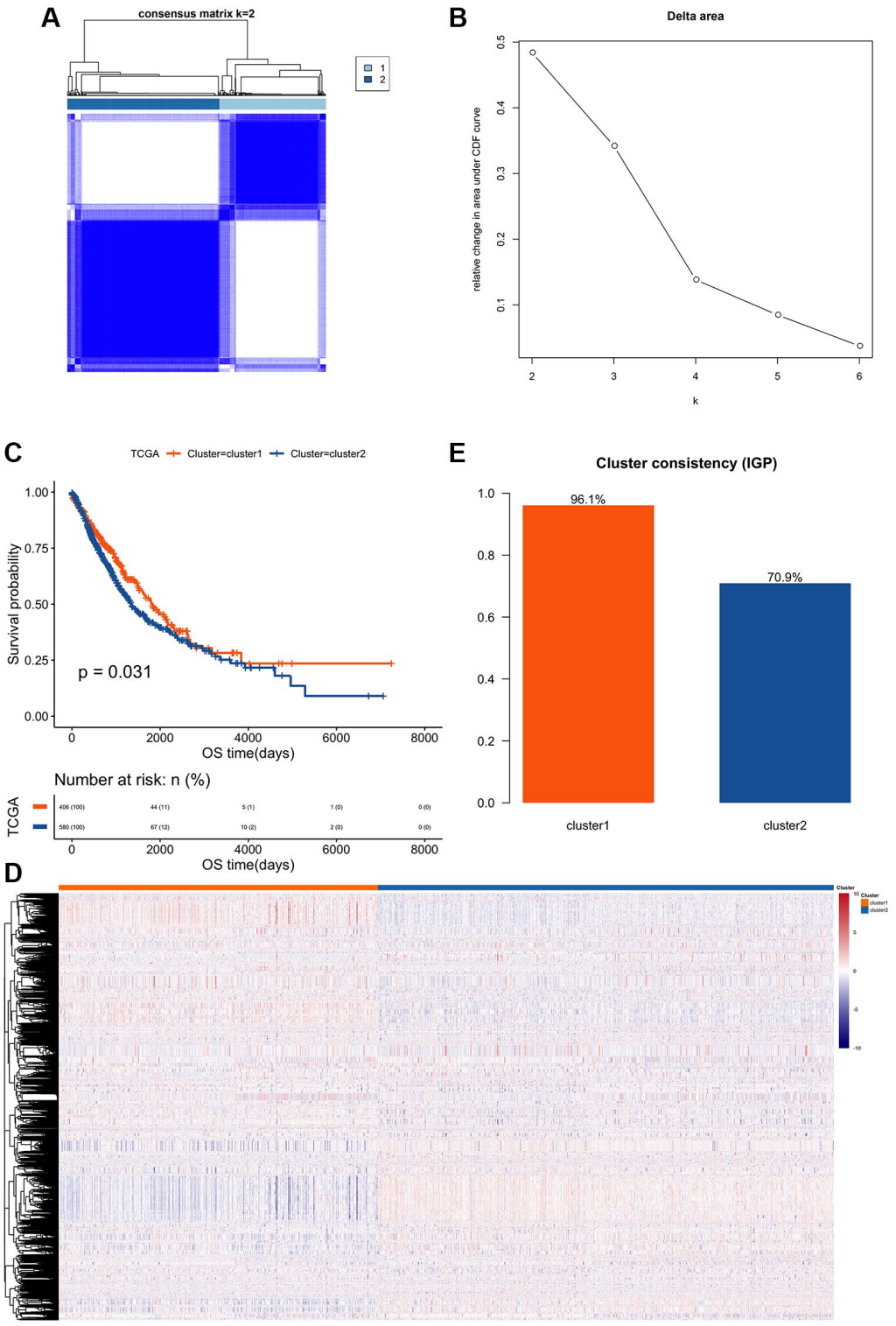


Figure 1. Clustering and survival analysis of feature edge perturbation matrix. (**A**) Based on the extracted 5468 feature edges, a consistent cluster analysis was performed according to their feature values, and 986 cancer samples were divided into two different subtypes. (**B**) Cumulative Distribution Function of Consistent Cluster Analysis. The abscissa axis represents the K value; the ordinate axis represents the relative change in area under CDF curve. (**C**) The prognostic survival curves of the two cluster subtypes. The abscissa axis represents the overall survival time; the ordinate axis represents the survival probability corresponding to different survival time points. (**D**) Z-score heatmap of eigenvalues of feature edges: Using hierarchical clustering method to perform cluster analyses on the extracted feature values, it is found that there are obvious specific feature differences between the two subtypes. (**E**) Validation of clustering performance of gene interaction perturbation: Using an independent data set (GSE50081) to verify the clustering performance of edge perturbation features. The larger IGP corresponds to the better consistency of the clustering group.

in Fc epsilon receptor (FCERI) signaling, Fcgamma receptor (FCGR) dependent phagocytosis and other pathways (Supplementary Figure 5A and 5B).

Identification of differential methylation sites among characteristic subtypes

The TCGA-LUAD and TCGA-LUSC methylation value matrices were merged, and the samples (N = 797) that appeared in the two subtypes would be retained, and then more than 70% of the samples with NA sites were filtered out and filled with 0. Finally, after filtering out and filling, 395786 methylation sites were collected for differential analyses.

According to the fold change (|logFC|>0.1) and the significance threshold (FDR<0.01), a total of 7304 differentially methylated sites were screened in the two subtype samples by the R package limma. The |logFC| values of the differentially methylated sites were sorted from big to small, and the top 200 differentially methylated sites were plotted (Supplementary Figure 5C).

Construction and verification of risk scoring models based on differentially expressed prognostic-related genes

According to the fold change (|logFC|>0.585) and the significance threshold (FDR<0.01), a total of 945 differentially expressed genes were screened from the two subtype samples by the R package limma (Supplementary Figure 5D and 5E). Univariate cox regression analysis was performed on these 945 differentially expressed genes. When P value was less than 0.01, 9 differentially expressed genes were found

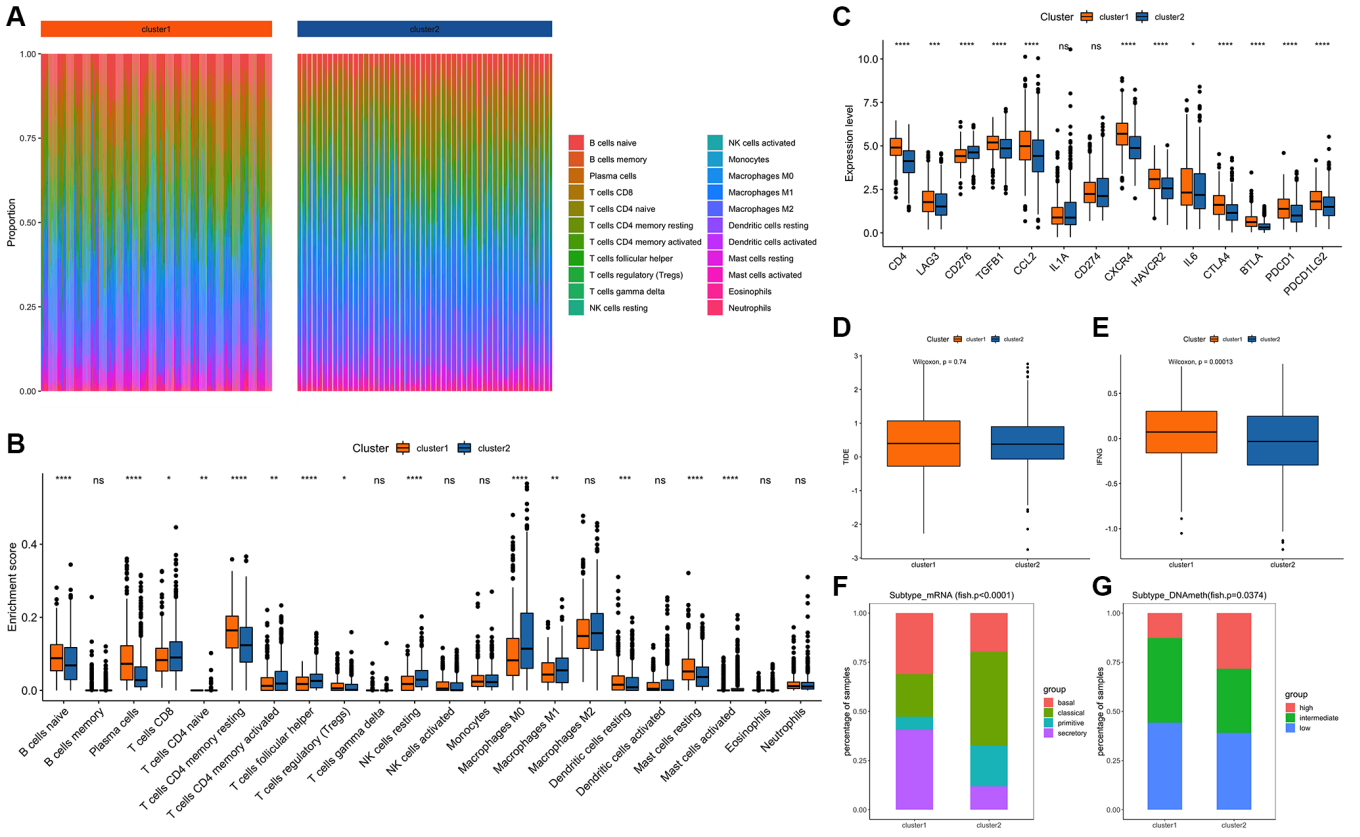


Figure 2. Comparison and analysis of feature subtypes of edge perturbation. (**A**) Infiltration scores of 22 immune cells in different samples: Different colors represent different immune infiltrating cells. The abscissa axis represents different samples; the ordinate axis represents the proportion of different immune cells. (**B**) Comparison of the differences of 22 immune cell infiltration scores among different cluster subtypes: The abscissa axis represents different types of immune cells; the ordinate represents the proportion of infiltrated immune cells. (**C**) Comparison of the expression levels of important immune checkpoint molecules among different cluster subtypes. (**D**) TIDE comparison between cluster 1 and cluster 2: The abscissa axis represents different clusters; the ordinate axis represents TIDE. (**E**) IFNG comparison between cluster 1 and cluster 2: The abscissa axis represents different clusters; the ordinate axis represents IFNG. (**F**) Comparison of mRNA expression levels among different subtypes: The different colors represent the known pathological types of NSCLC. The abscissa axis represents different clusters, and the ordinate axis represents the expression level of mRNA. (**G**) Comparison of DNA methylation levels among different subtypes: The different colors represent the levels of DNA methylation. The abscissa axis represents different clusters, and the ordinate axis represents the proportion of samples with different levels of DNA methylation.

to be related to the prognosis, and a forest plot was shown in (Figure 4A).

The LASSO method was further used to screen out 8 key prognostic genes (Figure 4B–4E), and after weighting the expression of these 8 genes by the LASSO regression coefficient, a risk scoring model for predicting the survival of the sample was constructed ("exp" represents gene expression level, "coef" represents LASSO regression coefficient).

$$\text{RiskScore} = \sum \exp \times coef$$

The risk score of each cancer sample was calculated based on the risk model. The surv_cutpoint function in the R package survminer was used to determine the classification threshold (1.1254), and further divide the cancer samples into high and low risk groups (N = 193/793), while significant prognosis differences were found in the two groups (Figure 5A). ROC was used to evaluate the predictive efficiency of the model, the AUC of cancer samples at 1, 3, and 5 years were 0.635, 0.659, and 0.605 (Figure 5B–5D).

It was verified in the independent data set GSE50081, and a similar analysis result trend was obtained (Figure 6A). The ROC predictive efficiency AUC values at 1, 3 and 5 years were 0.659, 0.554 and 0.555 (Figure 6B–6D).

DISCUSSION

Our understanding of cancer has promoted further with the advancement of sequencing technology and advances in cancer genome research [27, 28], particularly for lung cancers [27, 29–31]. These advances in genome research also enabled us to gain a comprehensive understanding of the temporal and spatial heterogeneity of lung cancer cells [29, 32], indicating that cancer's heterogeneity was unavoidable. However, recent research on biological functional networks has discovered that heterogeneity can be avoided or minimized to a certain extent through the use of sequencing and bioinformatics analysis technology [10–14]. Thus, using the R language-based bioinformatics analysis technology, we designed the following research to re-evaluate the alterations in the NSCLC genome after excluding tumor heterogeneity: Based on edge perturbation [15–18], functional gene interaction networks were used to deduce the pathological environment of individual patients with NSCLC [19–22], and to identify cancer subtypes with the same or similar status, followed by a multi-dimensional and multi-omics comprehensive analysis for validation [23–26].

After successfully constructing a background interaction network and extracting gene expression

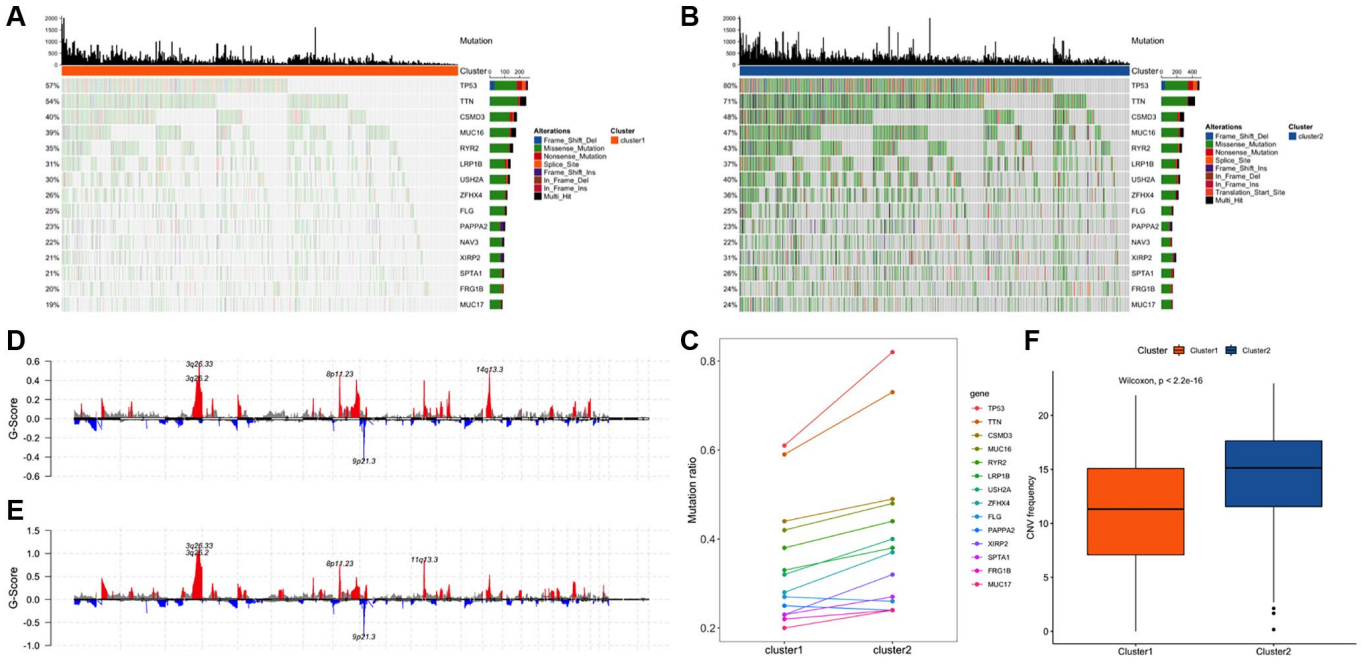


Figure 3. Comparison of genome alterations between cluster 1 and cluster 2. (**A**) The distribution of 15 mutated genes in cluster 1 among the co-occurring genes with mutation frequencies in the top 20 of the two clusters. (**B**) The distribution of 15 mutated genes in cluster 2 among the co-occurring genes with mutation frequencies in the top 20 of the two clusters. (**C**) Comparison of the mutation ratios of 14 genes in the two clusters. (**D**) Distribution of concentrated copy number amplification and deletion regions in the cluster 1. (**E**) Distribution of concentrated copy number amplification and deletion regions in the cluster 2. (**F**) Frequency distribution of copy number variation among subtype samples.

(Figure 1A and 1B), it was found that the characteristic value of the cancer sample was significantly higher than that of the normal sample (Figure 2A and 2B). This indicated that the degree of edge perturbation in cancer samples was greater, implying that the degree of perturbation in the background network of cancer samples was significantly greater than in normal samples [15], providing reliable evidence for the

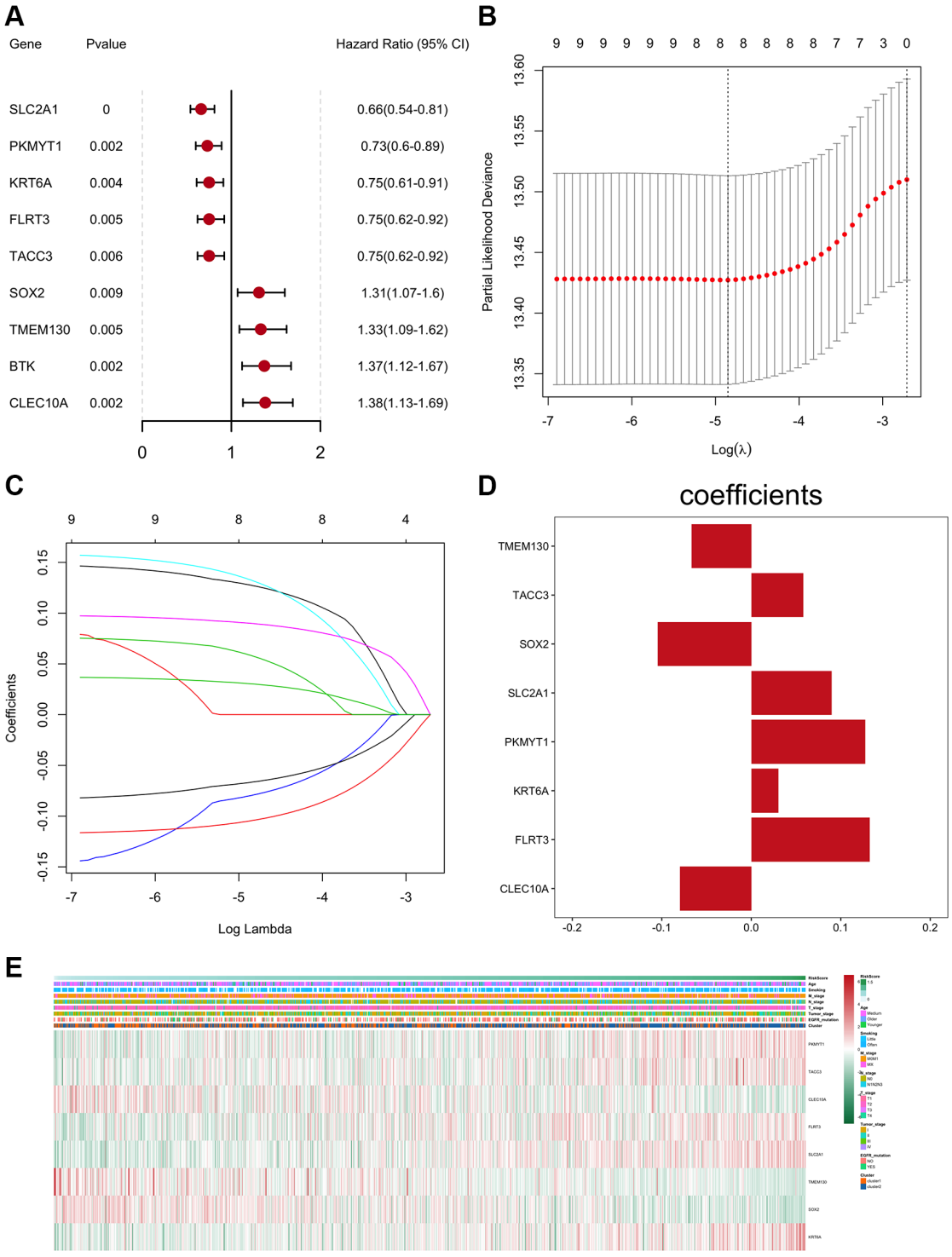


Figure 4. Construction and verification of the prognostic-related gene risk scoring model. (**A**) The forest plot of prognostic genes (univariate cox analysis result). (**B**) Confidence interval of each Lambda corresponding to LASSO regression. (**C**) The change trajectory of the independent variable in LASSO regression; the abscissa represents the logarithm of the independent variable Lambda, and the ordinate represents the coefficient of the independent variable. (**D**) LASSO regression coefficient of key prognostic genes. (**E**) Z-score heatmap of key prognostic genes expression levels.

subsequent use of EPM to reveal the heterogeneity of NSCLC samples.

Cluster analysis was performed using the R package ConsensusClusterPlus, as described previously [33, 34], and two distinct cluster subtypes were identified: cluster 1 (N = 406) and cluster 2 (N = 580) (Figure 3A–3D). Additional validation using the data set GSE50081 revealed that the clustering consistency based on the characteristic clustering analysis of the edge perturbation matrix was improved (Figure 3E). In comparison to cluster 1, cluster 2 had a higher tumor purity, ploidy, and stemness index (Supplementary Figure 3A–3D). Additionally, significant differences between clusters 1 and 2 were observed in stage, T, N, age, 22 immune cell infiltration scores, and differential expression of immune checkpoint molecules and IFNG (Supplementary Figure 3E–3J, Figure 2A–2E), which largely correlated with patient prognosis (Supplementary Figure 3C) [35, 36]. When the edge perturbation characteristic subtypes were compared to the reported NSCLC subtypes, the distribution of cancer subtypes corresponding to the edge perturbation characteristic subtypes remained clearly distinct (Figure 2F and 2G). Further examination of the molecular differences between the two clusters revealed obvious differences in either the mutation frequency (Figure 3A–3C), or the copy number variation frequency and distribution for the 15 genes co-occurring in the top 20 (Figure 3D–3F). This confirmed the molecular distinctions between these two cluster subtypes. Based on the above findings, we hypothesized that EPM-based clustering subtype classification could be a novel classification method for adenosquamous carcinoma

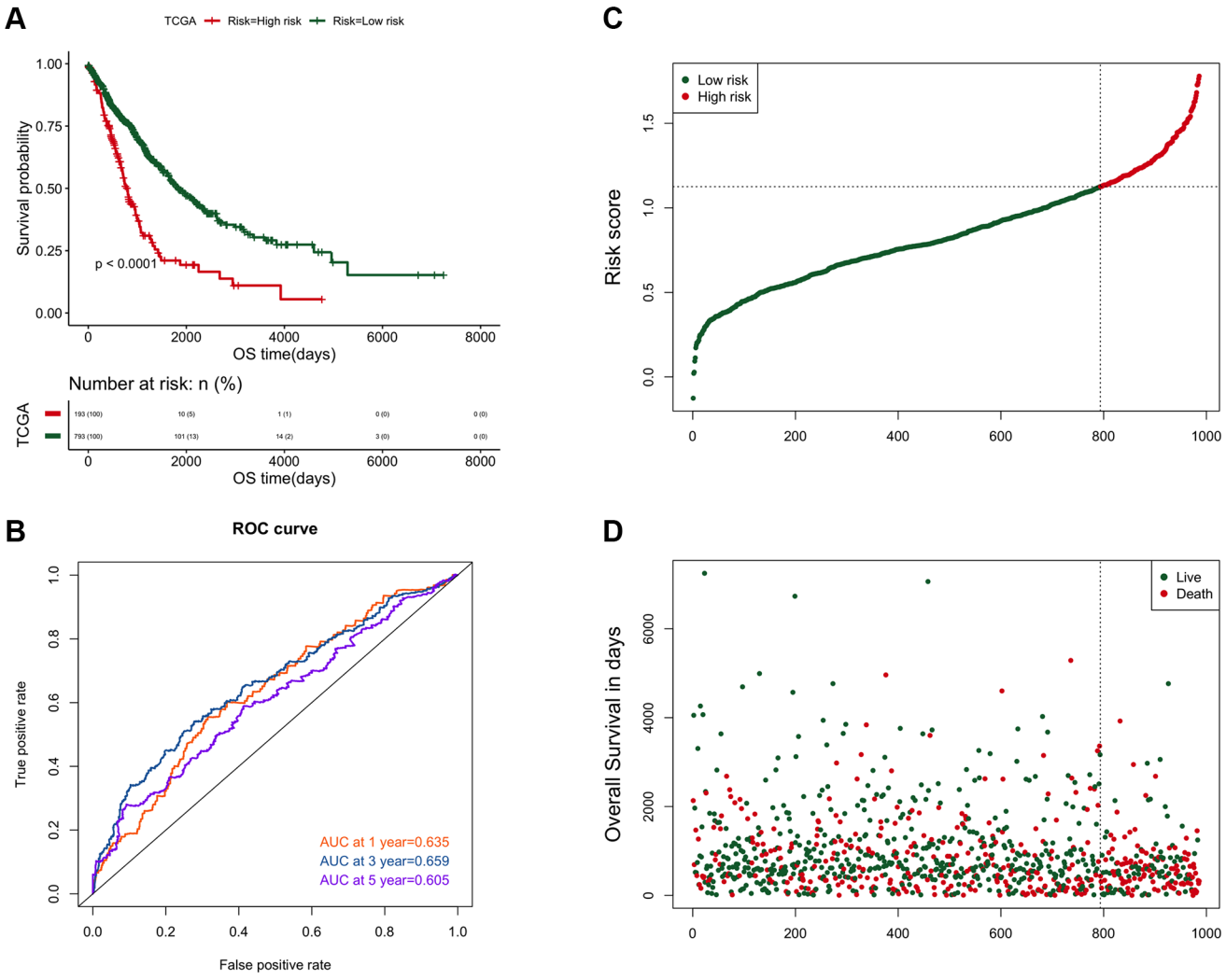


Figure 5. The efficiency analysis results of the risk prediction model. (**A**) Comparison of the prognosis between high and low risk groups. The abscissa axis represents the survival time (days); the ordinate axis represents the survival probability. (**B**) AUC curve of risk score on 1, 3, and 5-year survival prediction efficiency. (**C**) Risk score curve of the two groups. The abscissa axis represents the number of samples; the ordinate axis represents the risk score. (**D**) The ranking results of risk scores from small to large. The abscissa represents the number of samples, and the ordinate represents the survival time.

that was not inferior to the traditional pathological classification [37, 38].

TML, HRD, neoantigen load, and chromosome instability all played significant roles in the development of cancer [39–44]. As a result, these potential tumor immunogenicity-related factors were compared between the two cluster subtypes, and it was discovered that these indicators demonstrated statistically significant differences between the two clusters (Supplementary Figure 4). Similar differences were observed in analyses of the two subtype-specific clusters' KEGG pathway enrichment and methylation site recognition (Supplementary Figure 5A–5C). As a result of the above findings, we hypothesized that the innate immune escape mechanism between the two subtypes may be significantly different, confirming the importance of classifying NSCLC using this method.

As previously reported [45–47], LASSO was used to construct a risk scoring model containing eight genes based on the differential expression of prognostic-related genes between the two clusters (Figure 4) [48]. Regardless of whether it was verified in those two clusters (Figure 5) or in external data (GSE50081) (Figure 6), the risk scoring model demonstrated high prediction efficiency. Due to the fact that the data and validation data for this prediction model were derived from clinical sequencing results, it was believed that the prediction model would have a high probability of clinical applicability. In other words, this further demonstrated the feasibility and rationality of the

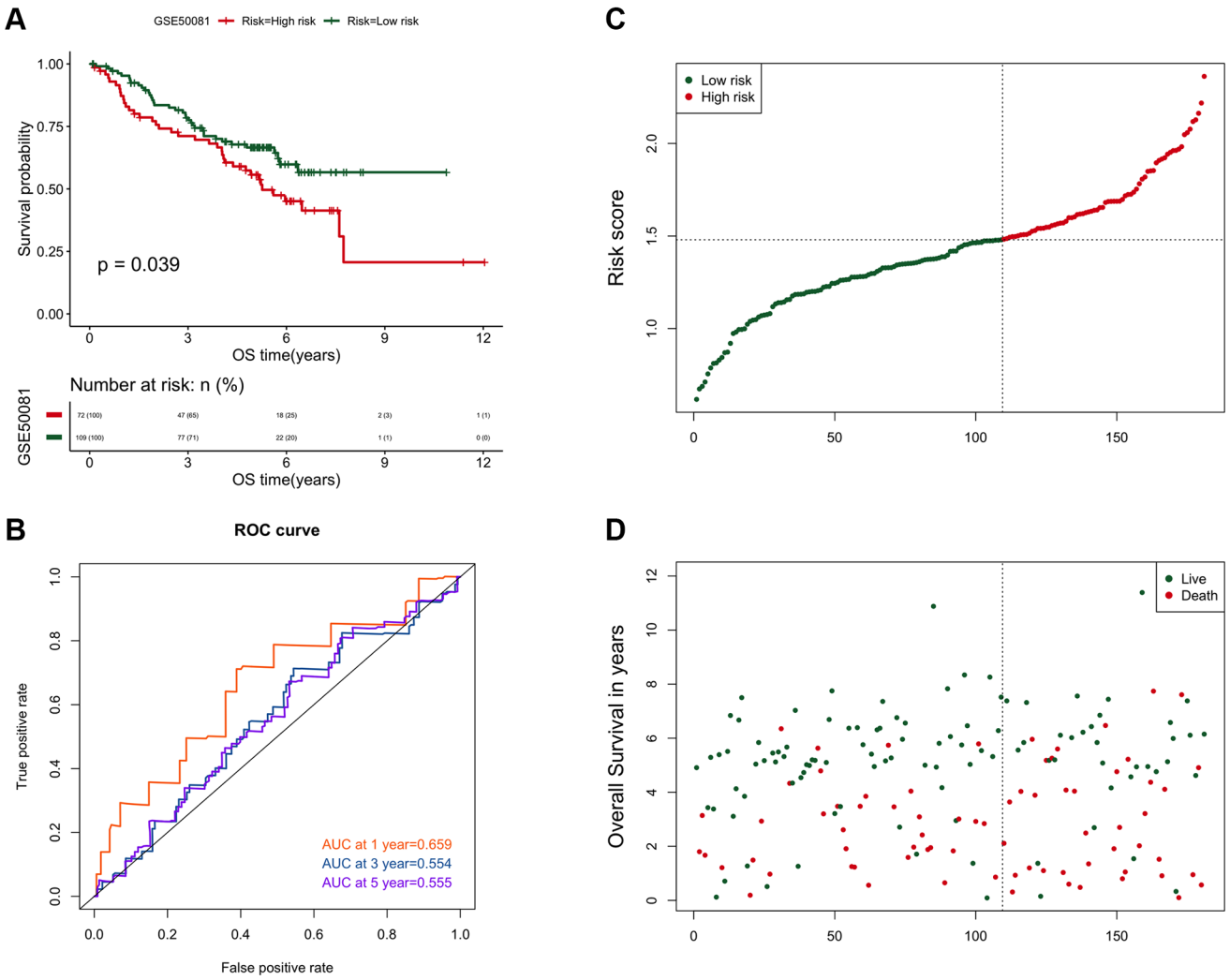


Figure 6. The verification result of the risk scoring model in the independent data set GSE50081. (**A**) Comparison of the prognosis between high and low risk groups in GSE50081. The abscissa axis represents the survival time (days); the ordinate axis represents the survival probability. (**B**) AUC curve of risk score on 1, 3, and 5-year survival prediction efficiency in GSE50081. (**C**) Risk score curve of the two groups in GSE50081. The abscissa axis represents the number of samples; the ordinate axis represents the risk score. (**D**) The ranking results of risk scores from small to large in GSE50081. The abscissa represents the number of samples, and the ordinate represents the survival time.

clustering analysis method based on EPM for the classification of NSCLC.

The EPM was constructed using the differences in expression fluctuations between cancer and normal samples, and then features for cluster analyses were extracted to accomplish the purpose of NSCLC classification. The advantage of this method is that it attempts to minimize the influence of tumor sample heterogeneity as much as possible, demonstrating some degree of innovation. With the rapid development of next-generation sequencing, for elderly patients with lung cancer, their blood samples can be used for next-generation sequencing, and then this method can be used for pathological typing, which can effectively avoid the damage and risks caused by pathological biopsy. Regardless of the method used to classify NSCLC or the risk scoring model developed, additional clinical samples and related basic experiments are still need to be conducted.

CONCLUSION

Clustering analysis using EPM for NSCLC classification is a feasible typing method that minimizes the effect of cancer sample heterogeneity. The risk scoring model constructed using those two clusters and involving eight genes has a high prediction efficiency.

Abbreviations

UCSC: University of California Santa Cruz; EPM: Edge-Perturbation Matrix; NA: No Available; TCGA: The Cancer Genome Atlas; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; TIDE: Tumour Immune Dysfunction and Exclusion; CNV: Copy Number Variation; LUAD: Lung Adenocarcinoma; LUSC: Lung squamous cell carcinoma; LASSO: Least Absolute Shrinkage and Selection Operator; TML: Tumour Mutation Load; FCGR: Fcgamma Receptor; FDR: False Discovery Rate; ROC curve: Receiver Operating Characteristic curve; AUC: Area Under Curve; GTEx: Genotype-Tissue Expression Program; ICI: Immune Checkpoint Inhibitor; HRD: Homologous Recombination Deficiency; IGP: Intra-Group Proportion; FCERI: Fc epsilon receptor; FCGR: Fcgamma receptor.

AUTHOR CONTRIBUTIONS

The study was designed by Yuan Tian and Yuping Sun; Yuan Tian, Alan Huang and Caiqing Zhang were responsible for data collection; All the data cleaning and analyses were put into practice by Yuan Tian, Alan Huang, Mei Tian, Qi Dang and Yumei Wei; The manuscript was drafted by Yuan Tian; Yuan Tian and Yuping Sun reviewed the manuscript for scientific soundness. All authors reviewed the final draft and approved its submission.

ACKNOWLEDGMENTS

The sources of funding for this study are listed as follows: the Academic Promotion Program of Shandong First Medical University (2019QL025; Yuping Sun), Natural Science Foundation of Shandong Province (ZR2019MH042; Yuping Sun), Jinan Science and Technology Program (201805064; Yuping Sun), and Postdoctoral Innovation Project of Jinan (Yuan Tian).

CONFLICTS OF INTEREST

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

REFERENCES

1. Thai AA, Solomon BJ, Sequist LV, Gainor JF, Heist RS. Lung cancer. Lancet. 2021; 398:535–54. https://doi.org/10.1016/S0140-6736(21)00312-3 PMID:34273294

2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022; 72:7–33. https://doi.org/10.3322/caac.21708 PMID:35020204

3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021; 71:209–49. https://doi.org/10.3322/caac.21660 PMID:33538338

4. Sosa E, D'Souza G, Akhtar A, Sur M, Love K, Duffels J, Raz DJ, Kim JY, Sun V, Erhunmwunsee L. Racial and socioeconomic disparities in lung cancer screening in the United States: A systematic review. CA Cancer J Clin. 2021; 71:299–314. https://doi.org/10.3322/caac.21671 PMID:34015860

5. Marte B. Tumour heterogeneity. Nature. 2013; 501:327. https://doi.org/10.1038/501327a PMID:24048064

6. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018; 15:81–94. https://doi.org/10.1038/nrclinonc.2017.166 PMID:29115304

7. Tanaka N, Kanatani S, Tomer R, Sahlgren C, Kronqvist P, Kaczynska D, Louhivuori L, Kis L, Lindh C, Mitura P,

Stepulak A, Corvigno S, Hartman J, et al. Whole-tissue biopsy phenotyping of three-dimensional tumours reveals patterns of cancer heterogeneity. Nat Biomed Eng. 2017; 1:796–806.
https://doi.org/10.1038/s41551-017-0139-0
PMID:31015588

8. Hao JJ, Lin DC, Dinh HQ, Mayakonda A, Jiang YY, Chang C, Jiang Y, Lu CC, Shi ZZ, Xu X, Zhang Y, Cai Y, Wang JW, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. Nat Genet. 2016; 48:1500–7.
https://doi.org/10.1038/ng.3683
PMID:27749841

9. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013; 501: 355–64.
https://doi.org/10.1038/nature12627
PMID:24048068

10. Lee S, Zhang C, Arif M, Liu Z, Benfeitas R, Bidkhori G, Deshmukh S, Al Shobky M, Lovric A, Boren J, Nielsen J, Uhlen M, Mardinoglu A. TCSBN: a database of tissue and cancer specific biological networks. Nucleic Acids Res. 2018; 46:D595–600.
https://doi.org/10.1093/nar/gkx994
PMID:29069445

11. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-Generation Machine Learning for Biological Networks. Cell. 2018; 173:1581–92.
https://doi.org/10.1016/j.cell.2018.05.015
PMID:29887378

12. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004; 5:101–13.
https://doi.org/10.1038/nrg1272
PMID:14735121

13. Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. Nat Rev Genet. 2016; 17:615–29.
https://doi.org/10.1038/nrg.2016.87
PMID:27498692

14. da Rocha EL, Ung CY, McGehee CD, Correia C, Li H. NetDecoder: a network biology platform that decodes context-specific biological networks and gene activities. Nucleic Acids Res. 2016; 44:e100.
https://doi.org/10.1093/nar/gkw166
PMID:26975659

15. Chen Y, Gu Y, Hu Z, Sun X. Sample-specific perturbation of gene interactions identifies breast cancer subtypes. Brief Bioinform. 2021; 22:bbaa268.
https://doi.org/10.1093/bib/bbaa268
PMID:33126248

16. Tegner J, Yeung MK, Hasty J, Collins JJ. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proc Natl Acad Sci U S A. 2003; 100:5944–9.
https://doi.org/10.1073/pnas.0933416100
PMID:12730377

17. Mellis IA, Edelstein HI, Truitt R, Goyal Y, Beck LE, Symmons O, Dunagin MC, Linares Saldana RA, Shah PP, Pérez-Bermejo JA, Padmanabhan A, Yang W, Jain R, Raj A. Responsiveness to perturbations is a hallmark of transcription factors that maintain cell identity in vitro. Cell Syst. 2021; 12:885–99.e8.
https://doi.org/10.1016/j.cels.2021.07.003
PMID:34352221

18. Krouk G, Lingeman J, Colon AM, Coruzzi G, Shasha D. Gene regulatory networks in plants: learning causality from time and perturbation. Genome Biol. 2013; 14:123.
https://doi.org/10.1186/gb-2013-14-6-123
PMID:23805876

19. Adolphe C, Xue A, Fard AT, Genovesi LA, Yang J, Wainwright BJ. Genetic and functional interaction network analysis reveals global enrichment of regulatory T cell genes influencing basal cell carcinoma susceptibility. Genome Med. 2021; 13:19.
https://doi.org/10.1186/s13073-021-00827-9
PMID:33549134

20. Yang YF, Cao W, Wu S, Qian W. Genetic Interaction Network as an Important Determinant of Gene Order in Genome Evolution. Mol Biol Evol. 2017; 34:3254–66.
https://doi.org/10.1093/molbev/msx264
PMID:29029158

21. Wragg D, Liu Q, Lin Z, Riggio V, Pugh CA, Beveridge AJ, Brown H, Hume DA, Harris SE, Deary IJ, Tenesa A, Prendergast JGD. Using regulatory variants to detect gene-gene interactions identifies networks of genes linked to cell immortalisation. Nat Commun. 2020; 11:343.
https://doi.org/10.1038/s41467-019-13762-6
PMID:31953380

22. Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, Pelechano V, Styles EB, Billmann M, et al. A global genetic interaction network maps a wiring diagram of cellular function. Science. 2016; 353:aaf1420.
https://doi.org/10.1126/science.aaf1420
PMID:27708008

23. Lindskrog SV, Prip F, Lamy P, Taber A, Groeneveld CS, Birkenkamp-Demtröder K, Jensen JB, Strandgaard T, Nordentoft I, Christensen E, Sokac M, Birkbak NJ, Maretty L, et al. An integrated multi-omics analysis identifies prognostic molecular subtypes of non-

muscle-invasive bladder cancer. Nat Commun. 2021; 12:2301.
https://doi.org/10.1038/s41467-021-22465-w
PMID:33863885

24. Kim HY, Choi HJ, Lee JY, Kong G. Cancer Target Gene Screening: a web application for breast cancer target gene screening using multi-omics data analysis. Brief Bioinform. 2020; 21:663–75.
https://doi.org/10.1093/bib/bbz003
PMID:30698638

25. Sathyanarayanan A, Gupta R, Thompson EW, Nyholt DR, Bauer DC, Nagaraj SH. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. Brief Bioinform. 2020; 21:1920–36.
https://doi.org/10.1093/bib/bbz121
PMID:31774481

26. Zhang Q, Lou Y, Yang J, Wang J, Feng J, Zhao Y, Wang L, Huang X, Fu Q, Ye M, Zhang X, Chen Y, Ma C, et al. Integrated multiomic analysis reveals comprehensive tumour heterogeneity and novel immunophenotypic classification in hepatocellular carcinomas. Gut. 2019; 68:2019–31.
https://doi.org/10.1136/gutjnl-2019-318912
PMID:31227589

27. Skoulidis F, Heymach JV. Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. Nat Rev Cancer. 2019; 19:495–509.
https://doi.org/10.1038/s41568-019-0179-8
PMID:31406302

28. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, et al, and International Cancer Genome Consortium. International network of cancer genome projects. Nature. 2010; 464:993–8.
https://doi.org/10.1038/nature08987
PMID:20393554

29. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, Shah K, Sato M, Thomas RK, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007; 450:893–8.
https://doi.org/10.1038/nature06358
PMID:17982442

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

31. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–25.
https://doi.org/10.1038/nature11404
PMID:22960745

32. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, Jamal-Hanjani M, Shafi S, Murugaesu N, Rowan AJ, Grönroos E, Muhammad MA, Horswell S, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014; 346:251–6.
https://doi.org/10.1126/science.1253462
PMID:25301630

33. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010; 26:1572–3.
https://doi.org/10.1093/bioinformatics/btq170
PMID:20427518

34. Danziger SA, McConnell M, Gockley J, Young M, Rosenthal A, Schmitz F, Reiss D, Farmer P, Ashby C, Bauer MA, van Rhee F, Davies FE, Zangari M, et al. Baseline and on-Treatment Bone Marrow Microenvironments Predict Myeloma Patient Outcomes and Inform Potential Intervention Strategies. Blood. 2018 (Suppl 1); 132:1882.
https://doi.org/10.1182/blood-2018-99-113169

35. Backman M, La Fleur L, Kurppa P, Djureinovic D, Elfving H, Brunnström H, Mattsson JSM, Pontén V, Eltahir M, Mangsbo S, Isaksson J, Jirström K, Kärre K, et al. WITHDRAWN: Characterization of Patterns of Immune Cell Infiltration in NSCLC. J Thorac Oncol. 2020. [Epub ahead of print].
https://doi.org/10.1016/j.jtho.2019.12.127
PMID:32028050

36. Muppa P, Parrilha Terra SBS, Sharma A, Mansfield AS, Aubry MC, Bhinge K, Asiedu MK, de Andrade M, Janaki N, Murphy SJ, Nasir A, Van Keulen V, Vasmatzis G, et al. Immune Cell Infiltration May Be a Key Determinant of Long-Term Survival in Small Cell Lung Cancer. J Thorac Oncol. 2019; 14:1286–95.
https://doi.org/10.1016/j.jtho.2019.03.028
PMID:31078775

37. Relli V, Trerotola M, Guerra E, Alberti S. Abandoning the Notion of Non-Small Cell Lung Cancer. Trends Mol Med. 2019; 25:585–94.
https://doi.org/10.1016/j.molmed.2019.04.012
PMID:31155338

38. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018; 24:1559–67.
https://doi.org/10.1038/s41591-018-0177-5
PMID:30224757

39. Hoppe MM, Sundar R, Tan DSP, Jeyasekharan AD. Biomarkers for Homologous Recombination Deficiency in Cancer. J Natl Cancer Inst. 2018; 110:704–13. https://doi.org/10.1093/jnci/djy085 PMID:29788099

40. Anagnostou V, Smith KN, Forde PM, Niknafs N, Bhattacharya R, White J, Zhang T, Adleff V, Phallen J, Wali N, Hruban C, Guthrie VB, Rodgers K, et al. Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non-Small Cell Lung Cancer. Cancer Discov. 2017; 7:264–76. https://doi.org/10.1158/2159-8290.CD-16-0828 PMID:28031159

41. Salem ME, Xiu J, Lenz HJ, Atkins MB, Philip PA, Hwang JJ, Gatalica Z, Xiao N, Gibney GT, El-Deiry WS, Tan AR, Kim ES, Shields AF, et al. Characterization of tumor mutation load (TML) in solid tumors. J Clin Oncol. 2017; 35:11517. https://doi.org/10.1200/JCO.2017.35.15_suppl.11517

42. Sansregret L, Vanhaesebroeck B, Swanton C. Determinants and clinical implications of chromosomal instability in cancer. Nat Rev Clin Oncol. 2018; 15:139–50. https://doi.org/10.1038/nrclinonc.2017.198 PMID:29297505

43. Bakhoum SF, Cantley LC. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. Cell. 2018; 174:1347–60. https://doi.org/10.1016/j.cell.2018.08.027 PMID:30193109

44. Watkins TBK, Lim EL, Petkovic M, Elizalde S, Birkbak NJ, Wilson GA, Moore DA, Grönroos E, Rowan A, Dewhurst SM, Demeulemeester J, Dentro SC, Horswell S, et al. Pervasive chromosomal instability and karyotype order in tumour evolution. Nature. 2020; 587:126–32. https://doi.org/10.1038/s41586-020-2698-6 PMID:32879494

45. Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, Lindbohm JV, Ahola-Olli A, Kurki M, Karjalainen J, Palta P, Neale BM, Daly M, Salomaa V, et al, and FinnGen. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med. 2020; 26:549–57. https://doi.org/10.1038/s41591-020-0800-0 PMID:32273609

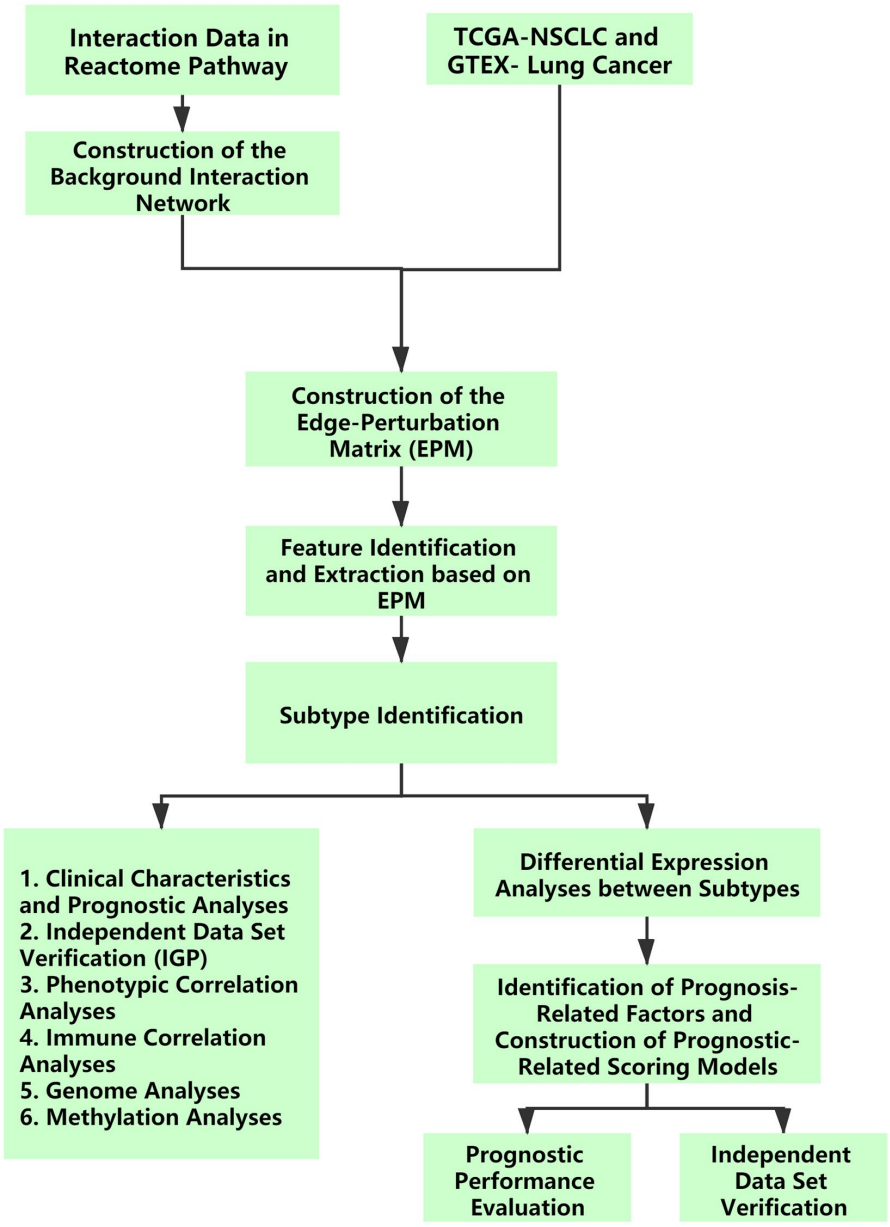
46. Bauer L, Hapfelmeier A, Blank S, Reiche M, Slotta-Huspenina J, Jesinghaus M, Novotny A, Schmidt T, Grosser B, Kohlruss M, Weichert W, Ott K, Keller G. A novel pretherapeutic gene expression-based risk score for treatment guidance in gastric cancer. Ann Oncol. 2018; 29:127–32. https://doi.org/10.1093/annonc/mdx685 PMID:29069277

47. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, Dehghan A, Muller DC, Elliott P, Tzoulaki I. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. JAMA. 2020; 323:636–45. https://doi.org/10.1001/jama.2019.22241 PMID:32068818

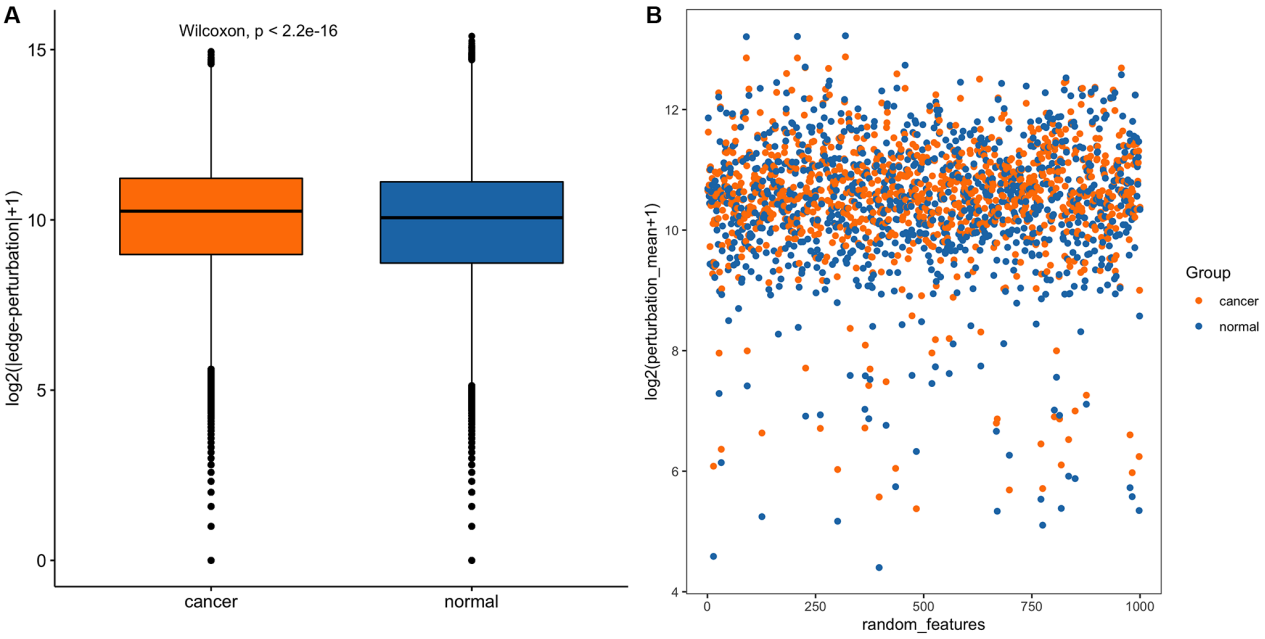
48. Ning Z, Lee Y, Joshi PK, Wilson JF, Pawitan Y, Shen X. A Selection Operator for Summary Association Statistics Reveals Allelic Heterogeneity of Complex Traits. Am J Hum Genet. 2017; 101:903–12. https://doi.org/10.1016/j.ajhg.2017.09.027 PMID:29198721

SUPPLEMENTARY MATERIALS

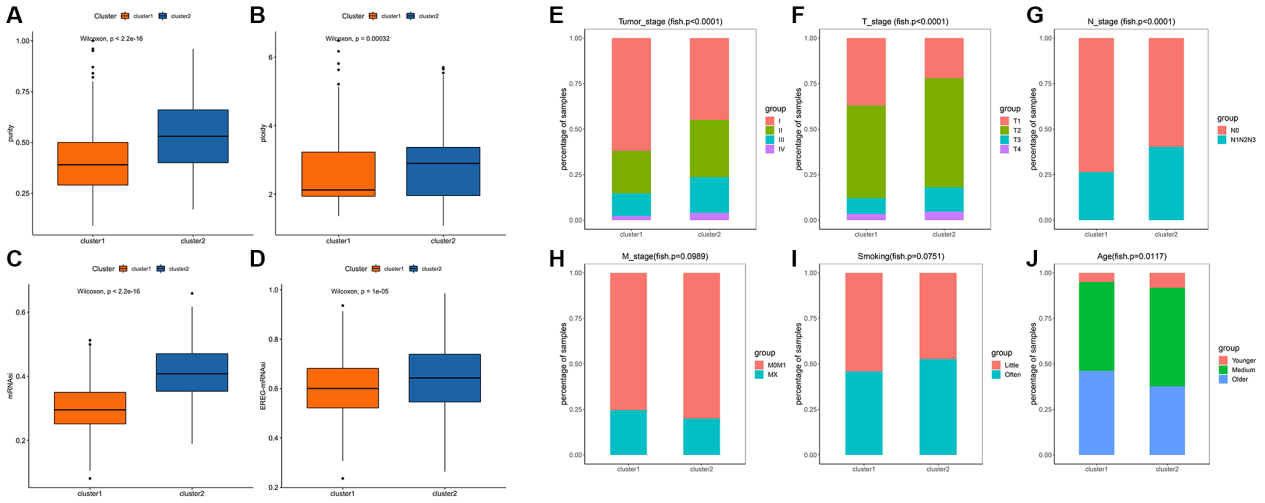
Supplementary Figures



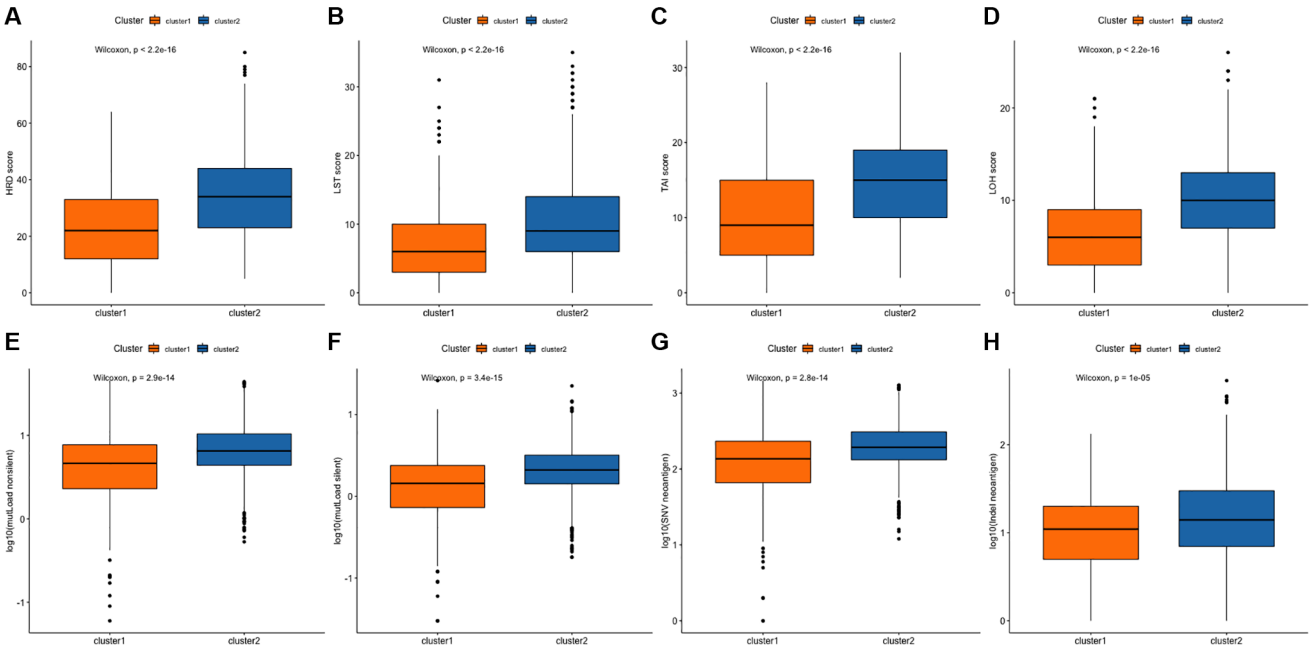
Supplementary Figure 1. The flow diagram of the study.



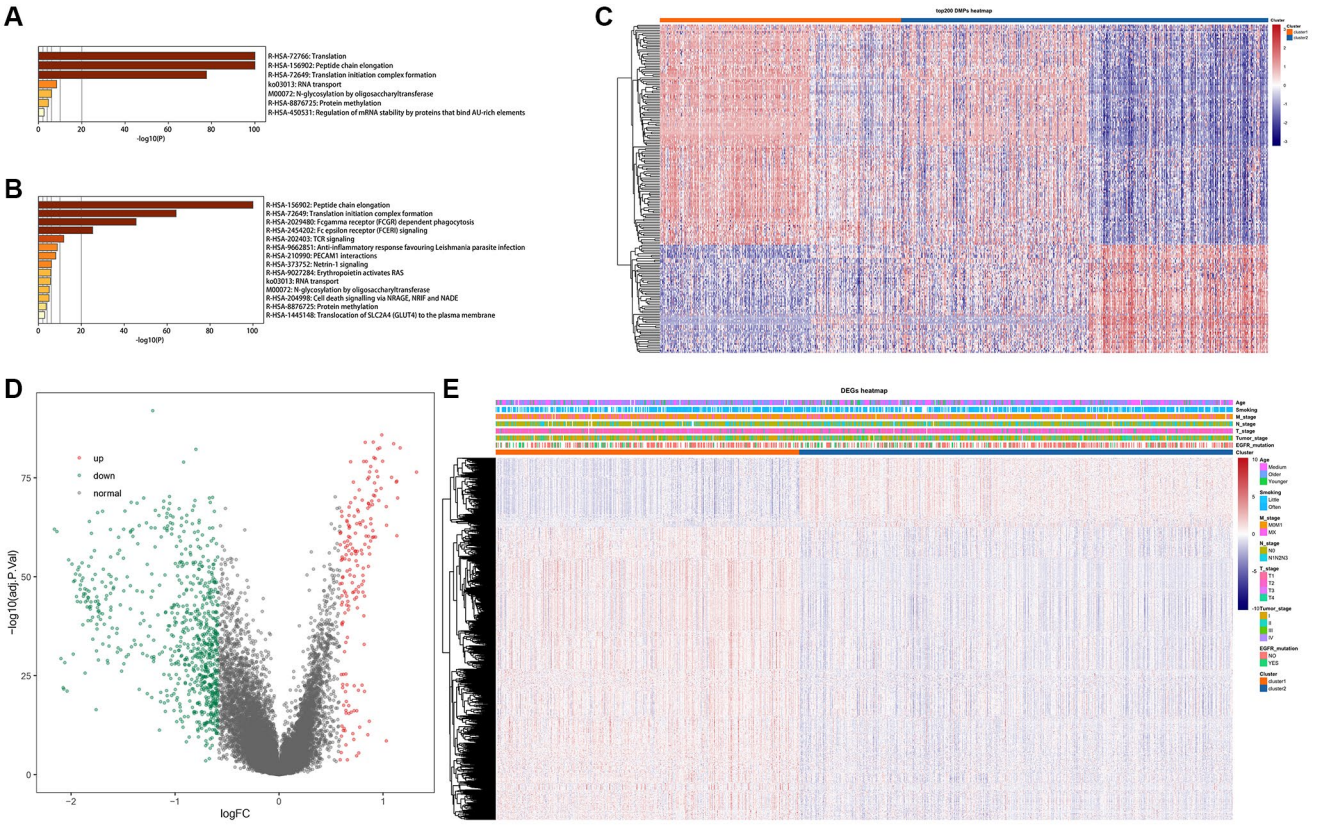
Supplementary Figure 2. Edge-perturbation matrix construction and feature extraction. (**A**) Box-plot diagrams of randomly selected 1000 features in cancer and normal samples. (**B**) Scatter plots of randomly selected 1000 features in cancer and normal samples.



Supplementary Figure 3. (**A–D**) Comparative analysis results of edge perturbation feature subtypes. (**A**) Differences in tumor purity between cluster 1 and cluster 2: The abscissa axis represents the cluster group; the ordinate axis represents the percentage of tumor purity. (**B**) Differences in genome ploidy between cluster 1 and cluster 2: The abscissa axis represents the cluster group; the ordinate axis represents the percentage of tumor genome ploidy. (**C**) Differences in stemness indices of mRNA between cluster 1 and cluster 2: The abscissa axis represents the cluster group; the ordinate axis represents the percentage of stemness indices of mRNA. (**D**) Differences in epigenetic regulated stemness indices of mRNA between cluster 1 and cluster 2: The abscissa axis represents the cluster group; the ordinate axis represents the percentage of epigenetic regulated stemness indices of mRNA. (**E–J**) Comparison of clinical characteristics among different subtypes. (**E**) Comparison of the proportions of different stages (Stage I–IV) in different clusters: The abscissa axis represents different clusters; the ordinate axis represents the proportion of different stages. (**F**) Comparison of the proportions of different *T* stages (T1–T4) in different clusters: The abscissa axis represents different clusters; the ordinate axis represents the proportion of different T stages (T1–T4). (**G**) Comparison of the proportions of different N stages (N1–N3) in different clusters: The abscissa axis represents different clusters; the ordinate axis represents the proportion of different *N* stages (N1–N3). (**H**) Comparison of the proportions of different M stages (M0M1 or Mx) in different clusters: The abscissa axis represents different clusters; the ordinate axis represents the proportion of different M stages (M0M1 or Mx). (**I**) Comparison of the proportions of different smoking status in different clusters: The abscissa axis represents different clusters; the ordinate axis represents the proportion of different smoking status (little or often). (**J**) Comparison of the proportions of different age stages in different clusters: The abscissa axis represents different clusters; the ordinate axis represents the proportion of different age stages (younger, medium or older).



Supplementary Figure 4. Comparison of immune escape mechanism between the two clusters. (**A**) Comparison of homologous recombination deficiency scores between the two clusters: The abscissa axis represents different clusters; the ordinate axis represents the HRD score. (**B**) The level of chromosome instability between the two clusters: The abscissa axis represents different clusters; the ordinate axis represents the LST score. (**C**) The level of chromosome instability between the two clusters: The abscissa axis represents different clusters; the ordinate axis represents the TAI score. (**D**) The level of chromosome instability between the two clusters: The abscissa axis represents different clusters; the ordinate axis represents the LOH score. (**E**) The level of tumor mutation load between the two clusters: The abscissa axis represents different clusters; the ordinate axis represents the value of log10 (mutLoad nonsilent). (**F**) The level of tumor mutation load between the two clusters: The abscissa axis represents different clusters; the ordinate axis represents the value of log10 (mutLoad silent). (**G**) The level of tumor neoantigen load between the two clusters: The abscissa axis represents different clusters; the ordinate axis represents the value of log10 (SNV neoantigen). (**H**) The level of tumor neoantigen load between the two clusters: The abscissa axis represents different clusters; the ordinate axis represents the value of log10 (Indel neoantigen).



Supplementary Figure 5. (**A**) Pathway enrichment analysis results of the cluster 1. (**B**) Pathway enrichment analysis results of the cluster 2. (**C**) Identification results of differential methylation sites among characteristic subtypes: z-score heatmap of the top200 differential methylation sites. (**D**) Volcano map of differentially expressed genes. (**E**) Z-score heatmap of differentially expressed genes.