Establishment and validation of an eight-gene metabolic–related prognostic signature model for lung adenocarcinoma

Weishuang Ma[1,4,*], Jiaming Liang[2,*], Jinhui Liu[3,*], Dongbo Tian[4], Zisheng Chen[4]

[1]Zhouxin Community Health Service, Qingcheng District, Qingyuan, China
[2]State Key Laboratory of Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, National Clinical Research Center for Respiratory Disease, Guangzhou, China
[3]The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China
[4]Department of Respiratory Medicine, The Sixth Affiliated Hospital of Guangzhou Medical University, Qingyuan People's Hospital, Qingyuan, China
[*]Equal contribution

Correspondence to: Zisheng Chen; **email:** 502463784@qq.com, https://orcid.org/0000-0002-7466-5289
Keywords: lung adenocarcinoma, metabolism reprogramming, tumor microenvironment, The Cancer Genome Atlas, prognosis
Received: October 31, 2020   **Accepted:** February 3, 2021   **Published:** February 22, 2021

ABSTRACT

In this study, we constructed an eight-gene metabolic related signature for LUAD. The eight-gene prognostic signature (including PLAUR, F2, UGT2B17, GNG7, IDO2, ST3GAL6, PIK3CG, and GLS2) exhibited a good prognostic value in the TCGA LUAD training dataset and testing dataset. In addition, the risk score based on the signature model was significantly correlated with immune cell infiltration and expression levels of immune markers in LUAD patients. LUAD cohorts from GEO were used to validate the model, indicating the usefulness of the model. In summary, we developed and validated an eight-gene signature model for LUAD, which can reflect the immune microenvironment characteristics and predict the prognostic outcomes for LUAD patients.

INTRODUCTION

Globally, lung cancer is the leading cause of cancer-associated mortalities, and lung adenocarcinoma (LUAD) is the most common type of lung cancer accounting for 40% of all cases [1]. In recent years, immune therapy has significantly improved the prognosis of non-small-cell lung carcinoma (NSCLC) patients [2–4]. The use of traditional biomarkers such as tumor mutation burden (TMB), blood TMB, expression levels of PD-1/PD-L1 and circulating tumor DNA (ctDNA) for predicting immunotherapeutic responses is inhibited by several limitations [5]. Therefore, identification of new biomarkers for predicting immune responses and prognosis is urgently needed. There has been a growing awareness on the importance of the tumor microenvironment (TME). The TME is composed of different types of immune cells, extracellular matrix, blood and lymphatic vessels, which exhibit complex interactions with tumor cells and have the ability to influence tumor survival and progression in a beneficial or harmful way [6]. Among the TME components, immune cells are of great importance as they can directly kill the tumor cells, and there are many drug targets which improve or reinvigorate their functions [7].

Studies have documented the importance of metabolic reprogramming in immune cell functions [8]. Tumor metabolism can transform the TME by providing a favorable environment for tumor growth [9]. The transformed TME can enhance or decrease immune cell functions, thereby inhibiting or promoting tumor progression. Metabolic reprogramming in the immune cells inhibits their anti-tumor activities, thereby influencing tumor progress and immunotherapeutic efficacy.

The LUAD is a unique lung cancer subtype with a complex TME [10], and the complex TME can impact on

the progression of LUAD. This study evaluated the immune cell and stromal cell landscape of TME in LUAD samples obtained from the TCGA database using algorithms based on the bulk mRNA expression of the tumor samples to determine the prognosis of LUAD. Metabolic associated differentially expressed genes from the two groups were identified based on the median of the estimated score. The relationship between the metabolic genes and the immune cells was then explored. Finally, we identified a metabolic gene set associated with a higher immune cell infiltration that can be used to predict the survival outcomes of LUAD patients. A flow chart of the study design used is shown in Figure 1.

RESULTS

Higher estimate score is correlated with better prognosis

We first assigned the TCGA LUAD cohort into high and low groups according to their median immune score, stromal score, and estimate score, respectively. Then, we compared the differences in the distribution of clinical characteristics including gender, age, smoking status, TNM stage and survival outcomes between the two groups (Figure 2, Supplementary Figure 1 and Supplementary Figure 2). Clinical characteristics including gender and clinical stage were found to be significantly associated with immune score (Supplementary Figure 1A–1G), stromal score (Supplementary Figure 2A–2G), and estimate score (Figure 2A–2G). The tumor size was significantly associated with the estimate score (Figure 2E). The differences in overall survival time and progression-free survival time were then compared between the two groups. We observed statistically significant differences between the estimate score (Figure 2H) and immune score (Supplementary Figure 1H) and overall survival. However, there were no statistical differences between the estimate score (Figure 2I), stromal score (Supplementary Figure 2I), and immune score (Supplementary Figure 1I) and progression-free survival.

GO function enrichment analysis of the DEGs

The relative immune cell infiltration level in the TCGA LUAD cohort was estimated using the ssGSEA algorithm. Then, the relative quantity of the 28 immune
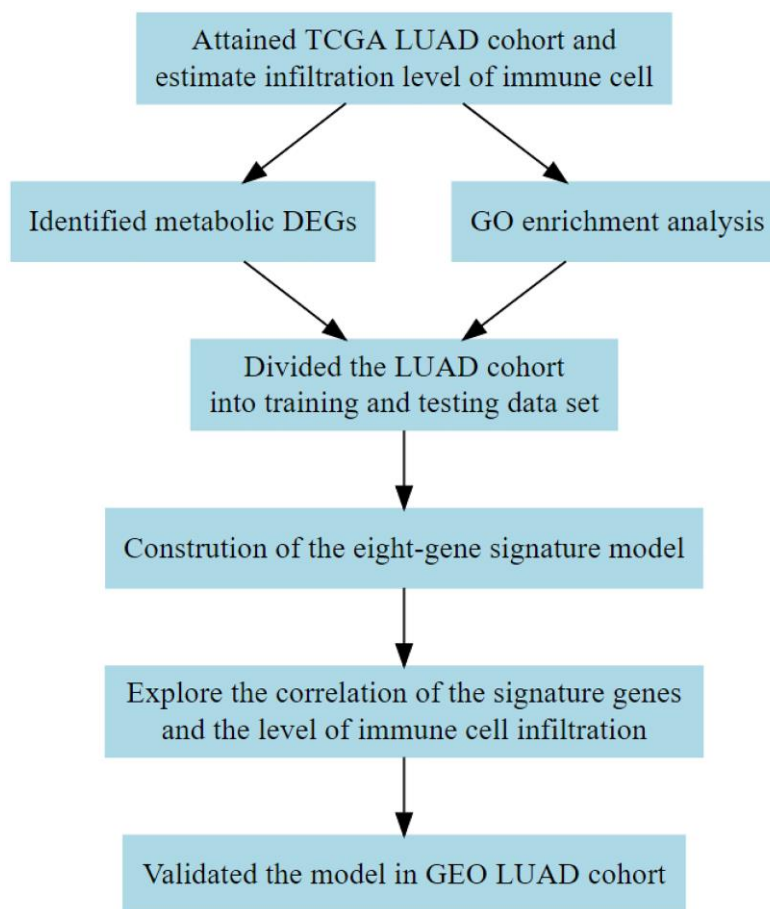
Figure 1. Flow diagram showing the design of the study. TCGA, The Cancer Genome Atlas; DEG, differentially expressed gene; GEO, Gene Expression Omnibus; GO, Gene Ontology.

cells among the two estimate groups were compared. It was found that for most immune cell types, the estimate scores between the two groups were significantly different and the relative quantity of immune cells in the high estimate score group was significantly higher than that in the low estimate score group (Figure 3A). Figure 3A shows that the activated B cell, activated CD4+ T cell, and activated CD8+ T cell exhibited a significant higher score in the high estimate score group than in the lower estimate score group.

Metabolic reprogramming in the tumor environment can impact on the function and population of the infiltrating immune cells [11]. Therefore, the cancer cell metabolic gene set was downloaded from ccmGDB (https://bioinfo.uth.edu/ccmGDB/), and a total of 2,072 metabolic genes were used to determine the association between metabolic gene expression and immune cell infiltration, as well as the relationship between the immune landscape and metabolic genes in the metabolic pathway.
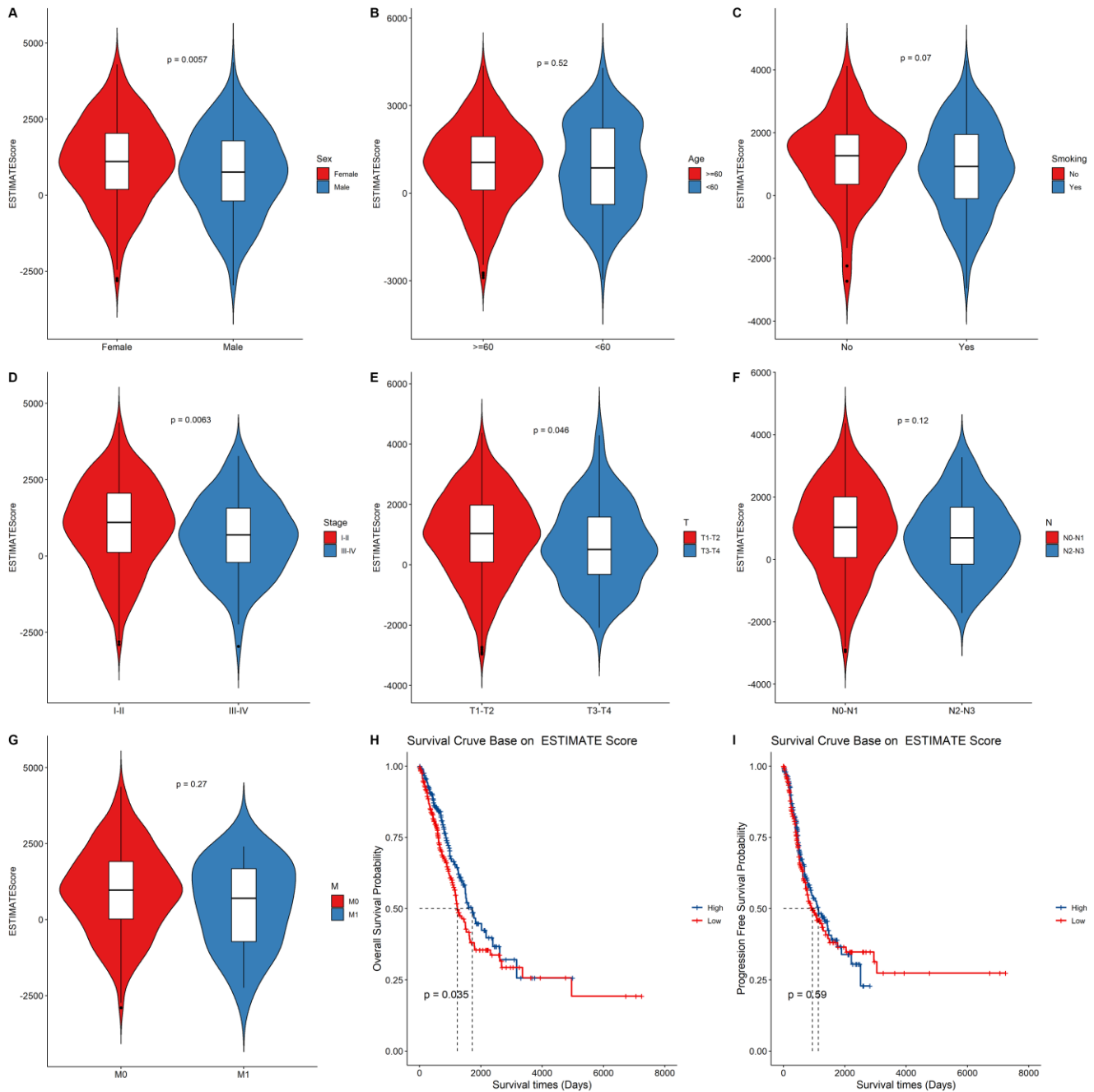
Figure 2. The relationship between the ESTIMATE score and clinical status, survival outcomes. (**A–G**) Boxplot showing the difference between the ESTIMATE score and the clinical characteristic. *p* value above the boxplot indicates the difference between the two groups. (**H–I**) The Kaplan–Meier curves for overall survival and progression free survival of LUAD risk groups divided using the median cutoff point of ESTIMATE score.

The samples were assigned into high estimate and low estimate groups based on the median of the estimate score. R package limma was then used to determine the differentially expressed genes (DEGs) of the two groups using |logFC| > 0.5 and adjusted $p < 0.05$ criteria. A total of 4,331 up regulated DEGs and 1,113 down regulated DEGs were identified (Figure 3B). As shown in Figure 3C–3D, of the differentially up-regulated and differentially down-regulated genes, 201 and 106 were metabolic-related genes, respectively. GO analysis

showed that the upregulated metabolic DEGs were significantly associated with sulfur compound metabolic processes, glycoprotein metabolic processes, and aminoglycan metabolic processes (Figure 3E). However, mRNA catabolic processes, RNA catabolic processes, and cellular amino acid metabolic processes were the enriched pathways among the down-regulated metabolic DEGs (Figure 3F). These findings imply that the two estimate groups exhibited different metabolic phenotypes as well as different infiltration levels of immune cell types.
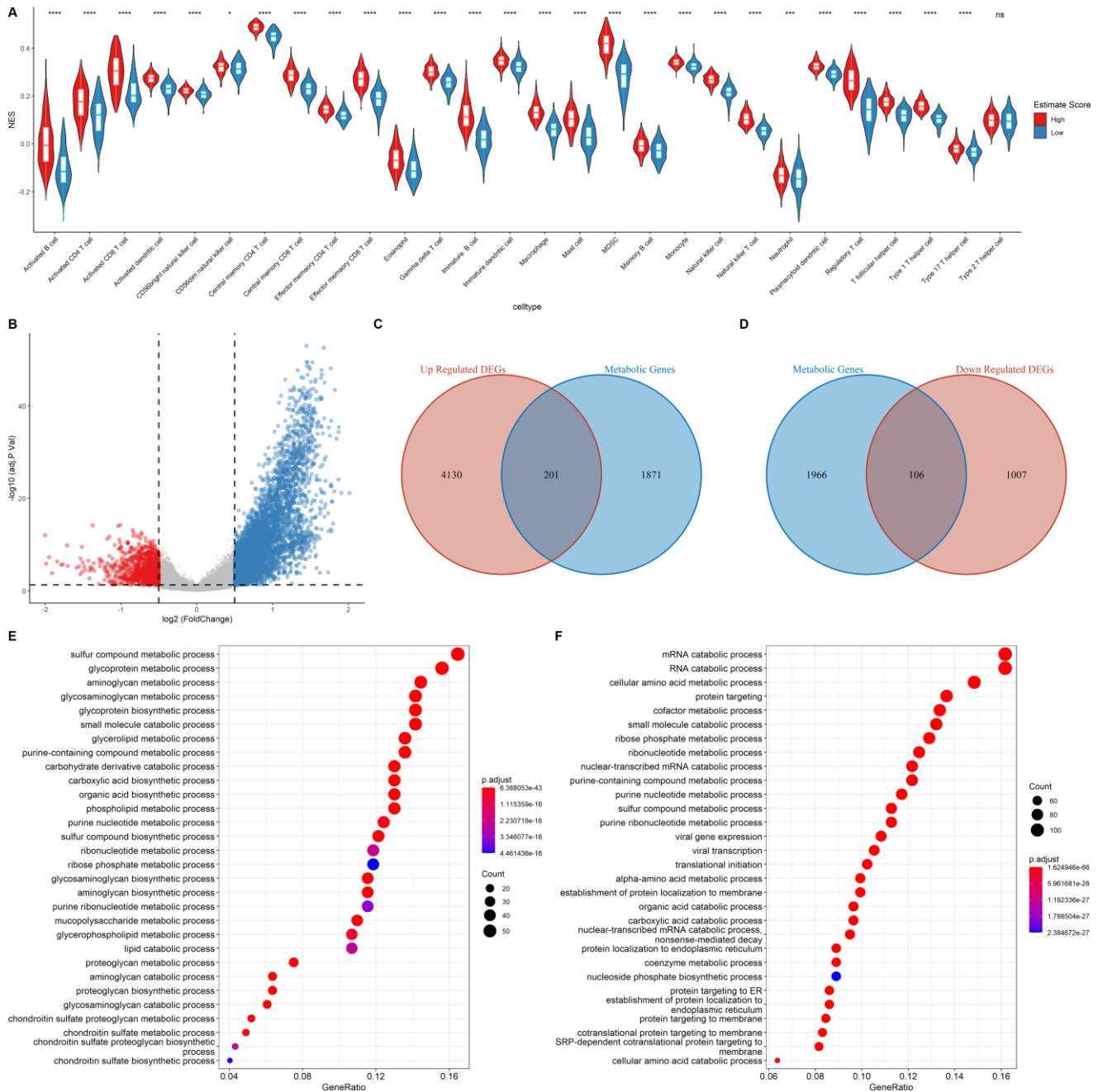

Figure 3. Functional annotation of the DEGs. (**A**) Correlation of Immune score and Immune cell score based on the ssGSEA algorithm (**B**) Volcano plot showing the DEGs. The criteria of the DEGs were set as |logFC| > 0.05 and adjusted $p < 0.05$. Red dots and blue dots represent genes that are significantly downregulated or upregulated, respectively. (**C–D**) Venn diagram showing the up regulated and down regulated metabolic DEGs. (**E–F**) GO analysis of up regulated genes and down regulated metabolic genes.

Construction of the metabolic gene signature

Univariate COX regression using the survival and gene expression data of LUAD patients was used to evaluate the overall survival prognostic value of the metabolic related DEGs. The 19 most significant genes were selected based on the $p < 0.01$ criteria. The hazard ratio of each gene is shown in the forest plot (Figure 4A). A heatmap representing the expression profiles of the 19 genes is shown in Figure 4B. The dataset was then divided into the training and testing data set at a ratio of 3:1 to construct a prognostic model (training data set $n = 385$; testing data set $n = 128$). LASSO regression was the performed in the training dataset to integrate the roles of these key molecules and to determine the genes that exhibited the greatest importance on the survival outcomes. A signature model of eight genes (PLAUR, F2, UGT2B17, GNG7, IDO2, ST3GAL6, PIK3CG, and GLS2) was constructed based on the LASSO regression results (Figure 4C–4E). The relationships among the relative expression levels of the eight signature genes in the TCGA LUAD cohort were then evaluated. The correlation analysis showed that F2 was significantly negatively correlated with ST3GAL6, GNG7, PIK3CG;
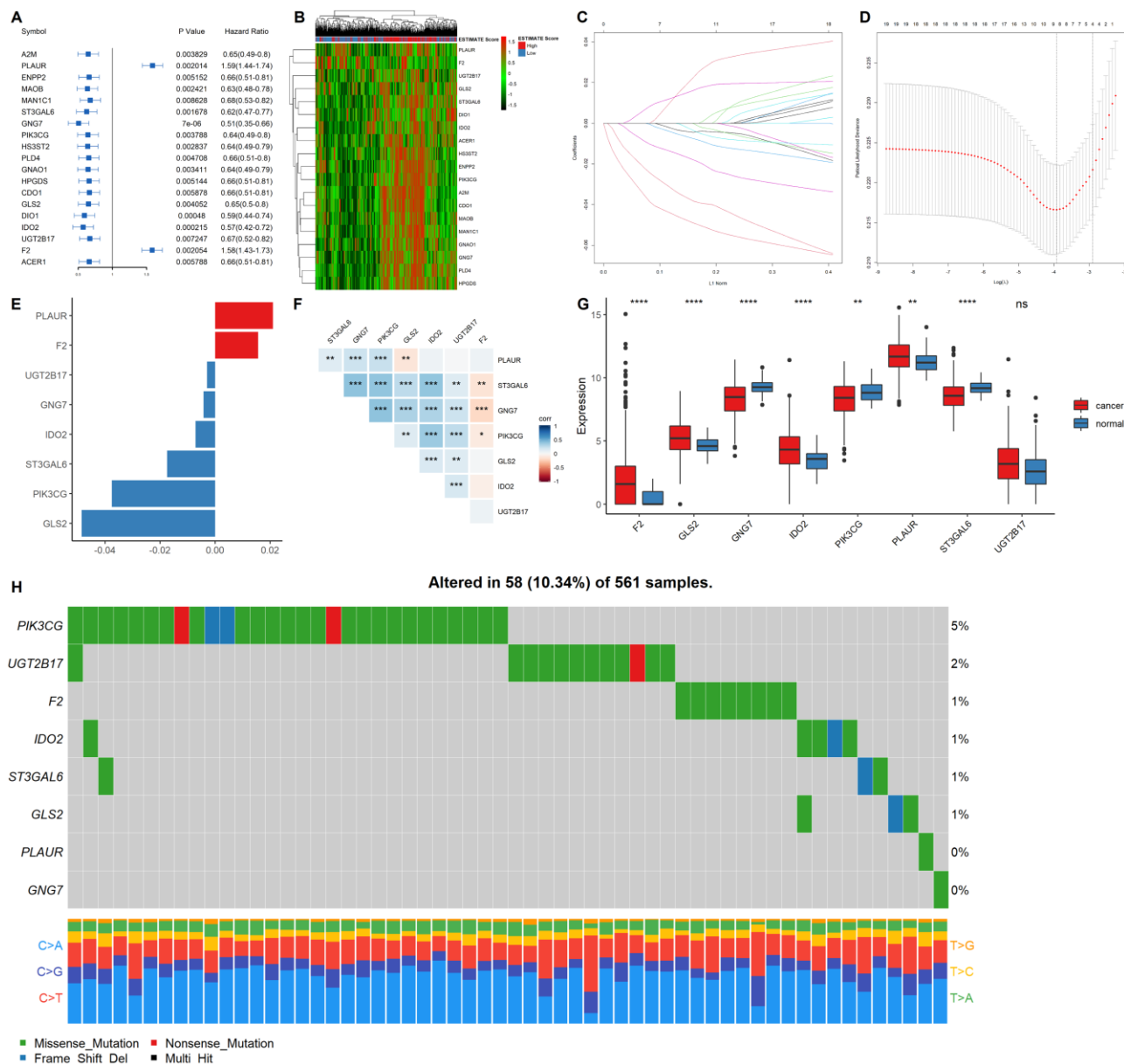
Figure 4. Construction of a prognostic gene signature. (**A**) Hazard ratio and p values of the selected candidate genes. (**B**) Heatmap showing the expression profiles of the selected candidate genes. (**C–D**) LASSO cox regression identified eight signature genes that were most correlated with OS. These genes were used to construct a signature model. (**E**) The coefficient value of the selected eight signature genes. (**F**) Expression correlation analysis of the eight signature genes. (**G**) The expression profiles of the eight signature genes in tumor and cancer samples. (**H**) Mutation landscape of the eight signature genes in LUAD.

GLS2 was significantly negatively correlated with PLAUR, while UGT2B17 was significantly positively correlated with ST3GAL6, GNG7, PIK3CG, GLS2, and IDO2. However, PLAUR did not exhibit any significant correlation with IDO2, UGT2B17, and F2 (Figure 4F). Analysis of this eight-gene signature using the STRING-DB database revealed that GNG7 interacts with F2 and PIK3CG (Supplementary Figure 3A). Next, we evaluated the expression profiles of these signature genes. In the tumor samples, F2, GLS2, IDO2, and PLAUR were shown to be significantly upregulated, while GNG7, PIK3CG, and ST3GAL6 were significantly down-regulated (Figure 4G).

EGFR mutations are highly prevalent in LUAD, especially among East Asian populations. The other main mechanisms of bypass signaling activation include IGF1R, MET, FGFR3, NTRK1, BRAF, ALK, RET, ROS1, and AXL [12]. Then, we elucidated on the correlation between these signature genes and other major drivers or bypass pathways in LUAD. It was found that PLAUR, ST3GAL6, GNG7, PIK3CG, GLS and IDO2 were significantly and positively correlated with most of the oncogenic drivers and bypass signaling, while F2 was negatively correlated with most of them (Supplementary Figure 3B). We also evaluated the mutation landscape of these genes in LUAD. Among the 561 samples, 10.34% of patients exhibited mutations in at least one signature gene. The PIK3CG exhibited the highest mutation frequency followed by UGT2B17, while GNG7 exhibited the fewest mutations in LUAD samples. The waterfall plot presentation of the mutation landscape of the eight signature genes showed that the mutation types were mainly missense mutation (Figure 4H). Most of the eight signature genes were differentially expressed in the tumor and normal tissues of LUAD and exhibited a certain rate of mutation in LUAD.

Low risk score correlated with better LUAD outcomes

The prognostic value of the eight signature genes was evaluated in the training and testing data sets. First, the risk score of each patient was calculated and ranked based on the risk score in the training data set (Figure 5A). The scatter plot was used to present the overall survival status of LUAD patients based on the risk score. Samples in the high-risk group were correlated with a higher mortality rate than those in the low-risk group (Figure 5B). A heatmap presenting the expression profiles of the signature genes showed that tumors with higher risk scores tended to exhibit elevated F2 and PLAUR levels, while those with lower risk scores tended to exhibit elevated UGT2B17, GNG7, IDO2, ST3GAL6, PIK3CG, and GLS2 levels (Figure 5C). This analysis was also performed on the testing dataset which showed consistent results with the training dataset (Figure 5D–5F).

The correlation between the eight-gene signature and the estimate score in LUAD was then evaluated. The risk score was found to be significantly higher in the low estimate score group than in the high estimate score group in both the training (Figure 6A) and testing data sets (Figure 6D). Estimations of the overall survival and progression free survival using the signature score was performed in the training and testing datasets. The low risk score group was correlated with better OS compared to the high-risk score group in both the training (Figure 6B, $p = 0.00014$) and testing datasets (Figure 6E, $p = 0.0082$). However, PFS was significantly different only in the training set (Figure 6C, $p = 0.014$) and not in the testing set (Figure 6F, $p = 0.51$). Collectively, these findings imply that the eight-gene signature model exhibits a good predictive prognostic power in LUAD, and the risk score of the signature model correlated with the estimate score of LUAD.

The signature score is associated with immune infiltration in LUAD

Since the metabolic eight-gene signature correlated with the estimate score, the correlation between the risk score and immune cell scores was then determined. A higher risk score was correlated with a lower abundance of activated CD4+ T cells, natural killer cells, and other immune cell types (Figure 7A). The relationship between the expression of the eight signature genes and the expression of the immune checkpoint molecules was also determined. The expression of PLAUR, GNG7, IDO2, ST3GAL6, and PIK3CG was significantly and positively correlated with the expression of the four checkpoint markers, PD-1, PD-L1, PD-L2, and CTLA-4 (Figure 7B). We also evaluated the correlation between the eight signature genes and immune cell infiltration. The expression levels of PLAUR, UGT2B17, GNG7, IDO2, ST3GAL6, and PIK3CG were significantly and positively correlated with the infiltration of most immune cell types (Figure 7C). Finally, the relationship between risk score and the expression levels of the four immune checkpoint molecules was determined. Risk score was found to be negatively correlated with the expression levels of the four immune checkpoint markers (Figure 7D–7G). Together, these results indicate that the metabolic eight-gene signature is correlated with the expression of immune checkpoint molecules and with the infiltration level of immune cells.

### Verification of the prognostic value of the signature genes using the GEO data set

Independent validation of the signature model was performed using the GEO LUAD cohort to verify the ability of the metabolic eight-gene signature model to predict the prognosis. The risk score for each sample of

the GEO LUAD cohort was determined using the signature model. The cohort was then divided into two groups based on the median of the risk score. Survival analysis of the four cohorts was then done to validate the prognostic value of the eight gene signature. A lower risk score was correlated with a better overall survival in all the four GEO datasets, including GSE31210 ($p = 0.0064$), GSE30219 ($p = 0.01$), GSE13213 ($p = 0.057$), and GSE50081 ($p = 0.014$) (Figure 8A–8D). Taken together, these results indicate

that the eight-gene signature model has a good predictive power for the NSCLC cohort.

DISCUSSION

The application of RNA-seq and the rapid development of bioinformatics tools have enhanced our understanding of tumors through research. Bianchi et al. constructed a ten-gene predictive model for stage I LUAD [13]; Chen et al. constructed a five-gene
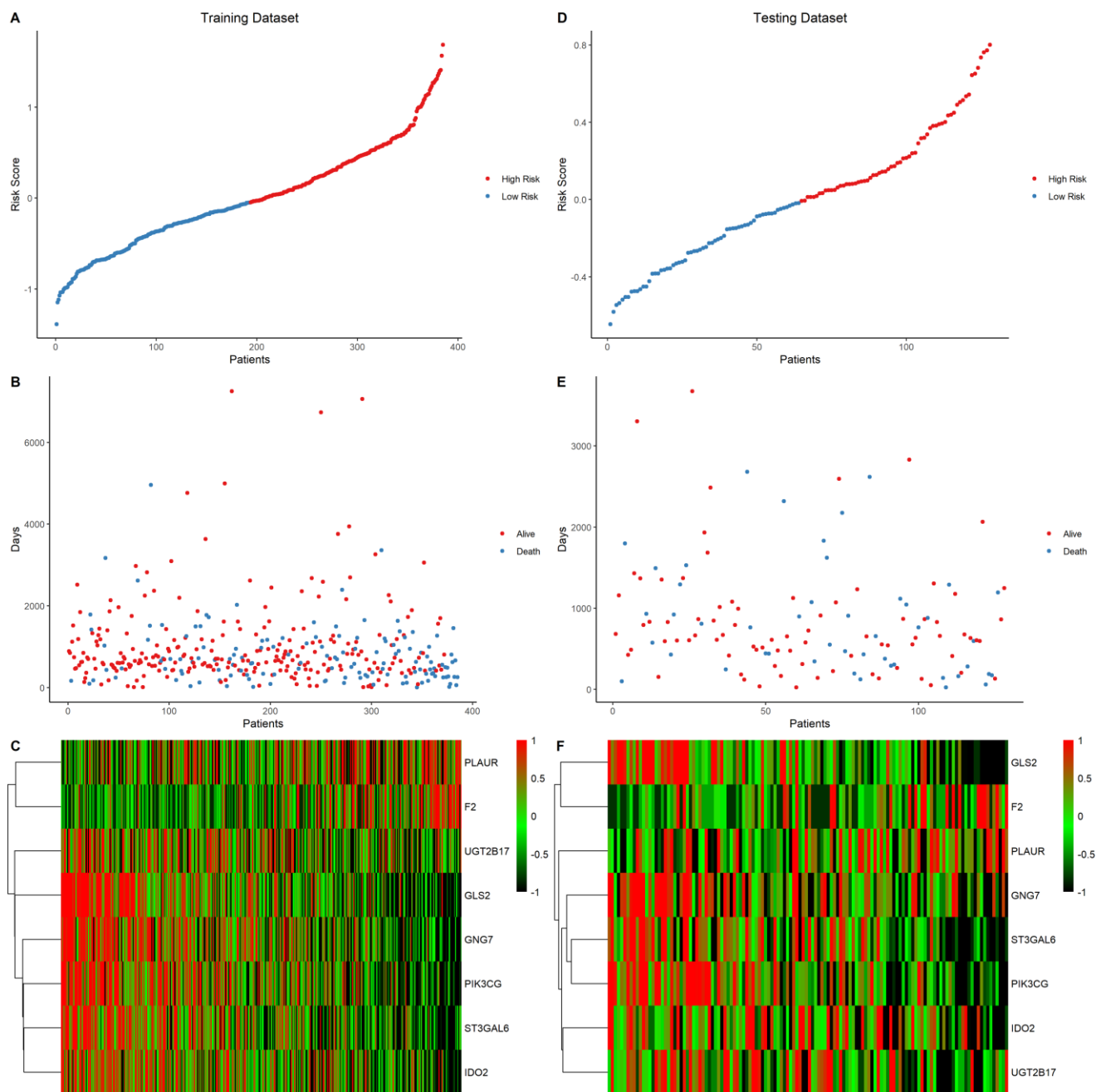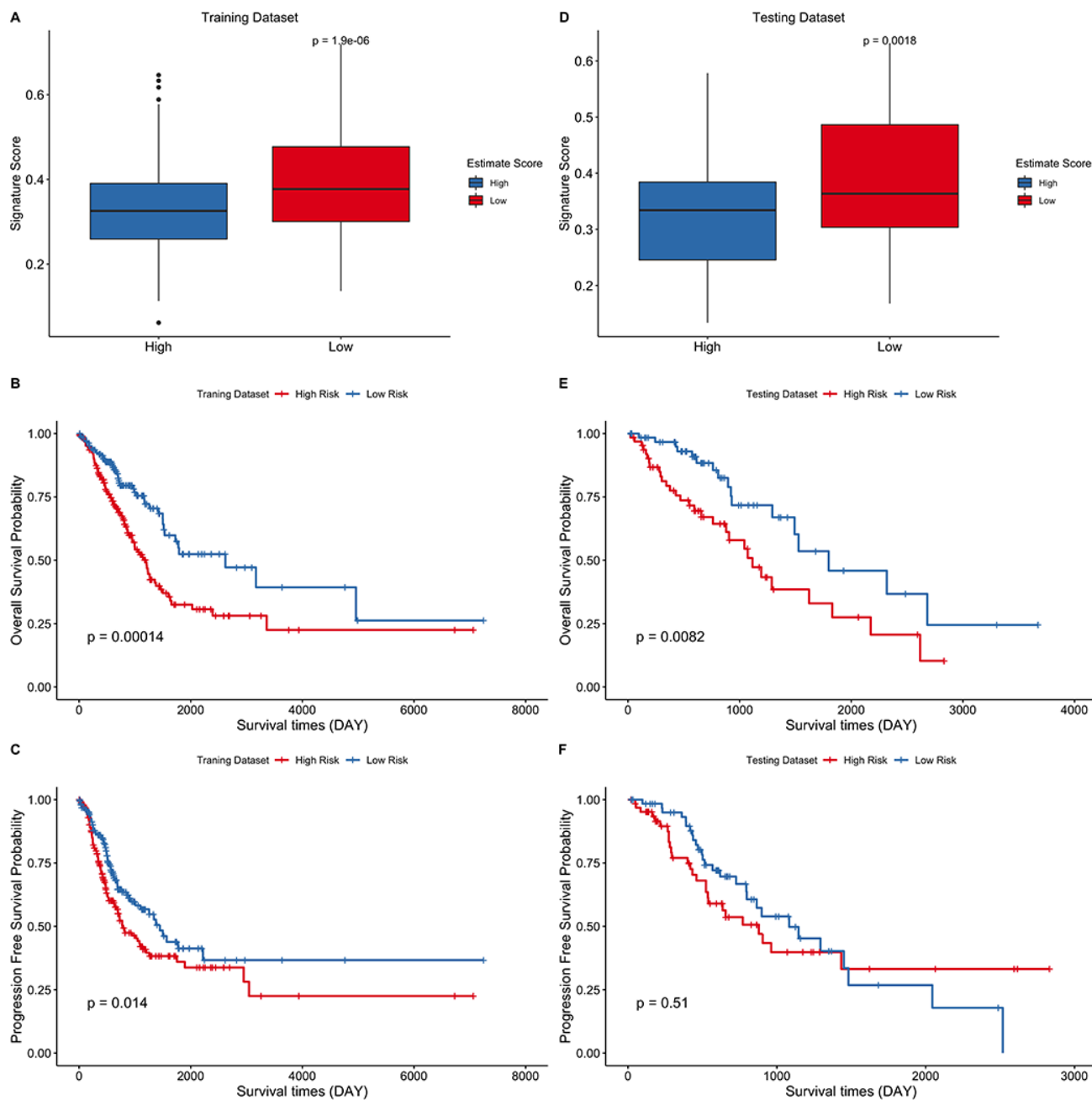

Figure 5. Eight-gene signature predictor score analysis in training and testing data set. (**A**–**C**) Training data set, (**D**–**F**) Testing data set. The ranked dot plot illustrated the predictor-score distribution of the training data set (A) and testing data set (D). A scatter plot presenting the patients' overall survival status from training data set (B) and testing data set (E). A heatmap showing the expression profile of the eight signature genes of LUAD patients from training data set (C) and testing data set (F).

signature model for the prediction of relapse-free and overall survival in NSCLC [14]; Shukla et al. constructed a four-gene prognostic signature for LUAD based on TCGA cohort [15]; Boutros et al. constructed a six-gene prognostic signature for NSCLC and validated it using other independent public microarray datasets that include multiple histology types and different stages [16]. However, few of these studies focused on the metabolic genes and their association with TME. Therefore, integrated analysis of metabolic

genes in LUAD based on data from GEO and TCGA databases were performed and a metabolic eight-gene signature for predicting the prognosis of LUAD patients was constructed.

We first evaluated LUAD immune infiltration using the TCGA LUAD cohort and identified the differentially expressed metabolic genes between the high estimate score group and the low estimate score group. The data set was then divided into training and testing data set.


Figure 6. Eight-gene signature predict survival outcomes in training and testing data sets. (**A–C**) Training dataset (**D–F**) Testing dataset. LUAD cohort was divided into two groups using the median of estimate score and the risk score of the two groups was then compared. Kaplan–Meier curves for LUAD risk groups divided using the median cutoff point. Patients with higher risk score exhibited significantly poor OS outcomes.

A prognostic gene signature was then constructed and used to predict the survival outcomes of the LUAD patients using the training data set. The prognostic value of the model was evaluated in the training and testing data sets. A higher risk score was correlated with poor overall survival outcomes. Correlation of the risk scores and immune cells revealed that a lower risk score was associated with increased monocyte infiltration score and T cell infiltration score, which has recently been reported as a predictor for poor prognosis in pancreatic cancer [17]. Immune markers, particularly PD-L1, are used as predictive biomarkers for immunotherapy [18].
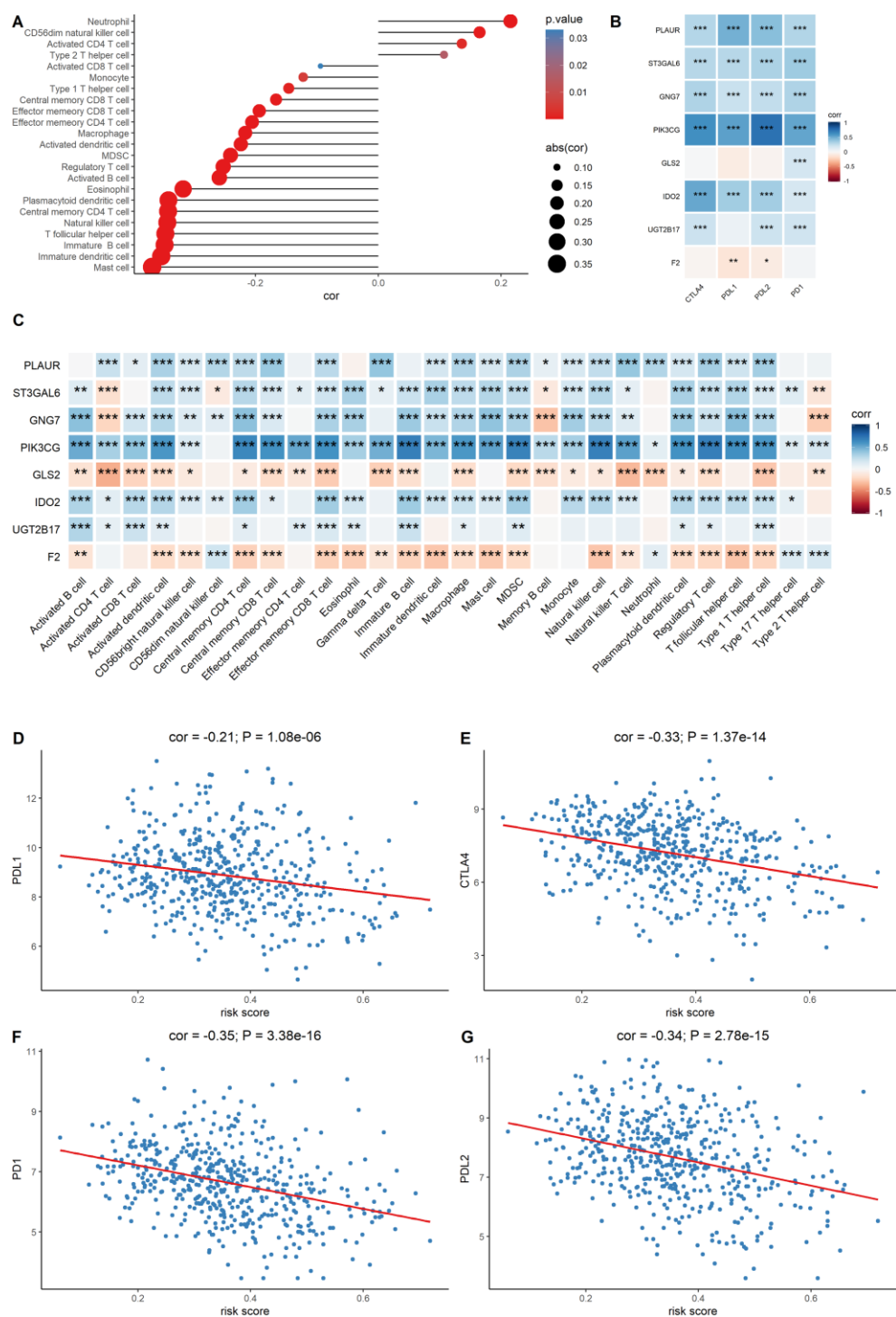

Figure 7. Evaluation of the correlation between the signature genes and immune characteristics. (**A**) Correlation between the risk score and immune cell infiltration score. (**B**) Correlation between the expression level of immune markers and the eight signature genes. (**C**) Correlation between each signature gene of the model and each immune cell type. (**D–G**) Correlation between the four immune checkpoint markers and the risk score.

We found that the expression levels of immune markers (PD-1, PD-L1, PD-L2, and CTLA-4) are associated with a lower risk score. This implies that the eight-gene signature model could be used to predict immunotherapeutic responses. Finally, we validated the gene signature using a GEO data set. A higher risk score was correlated with poor survival outcomes, which proved the prognostic value of the model (Figure 8).

In this study, we identified the differentially expressed genes between the high estimate score group and the low estimate score group based on the expression profile of LUAD patients. The metabolic related genes were then selected from the DEGs. Nineteen candidate genes were obtained after evaluating the prognostic value of the metabolic DEGs. These genes were then reduced to eight potential predictor genes using the LASSO algorithm. Finally, eight genes (PLAUR, F2, UGT2B17, GNG7, IDO2, ST3GAL6, PIK3CG, and GLS2) were included in the signature model. These

signature genes have been reported in previous studies. Some of them function in tumor progression or as prognostic markers for tumor patients.

For example, PIK3CG encodes the phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit gamma isoform (PI3Kγ) enzyme. PI3Kγ is involved in Akt/mTOR signaling, inhibition of NF-κB activation, and in the regulation of tumor immune inhibition by promoting MDSC migration to the tumor environment as well as by stimulating the immunosuppressive transformation of MDSCs [19]. GLS2 is a glutaminase (GLS) isoform, which converts glutamine into glutamate and provides nitrogen for nucleotide and protein synthesis. GLS2 knockdown was shown to inhibit cell proliferation by down-regulating the mTORC1 signaling and inducing autophagy in Gln-dependent lung squamous cell carcinoma cell lines [20]. UGT2B17 is a member of the UDP-glucuronosyltransferases (UGTs) family and is involved
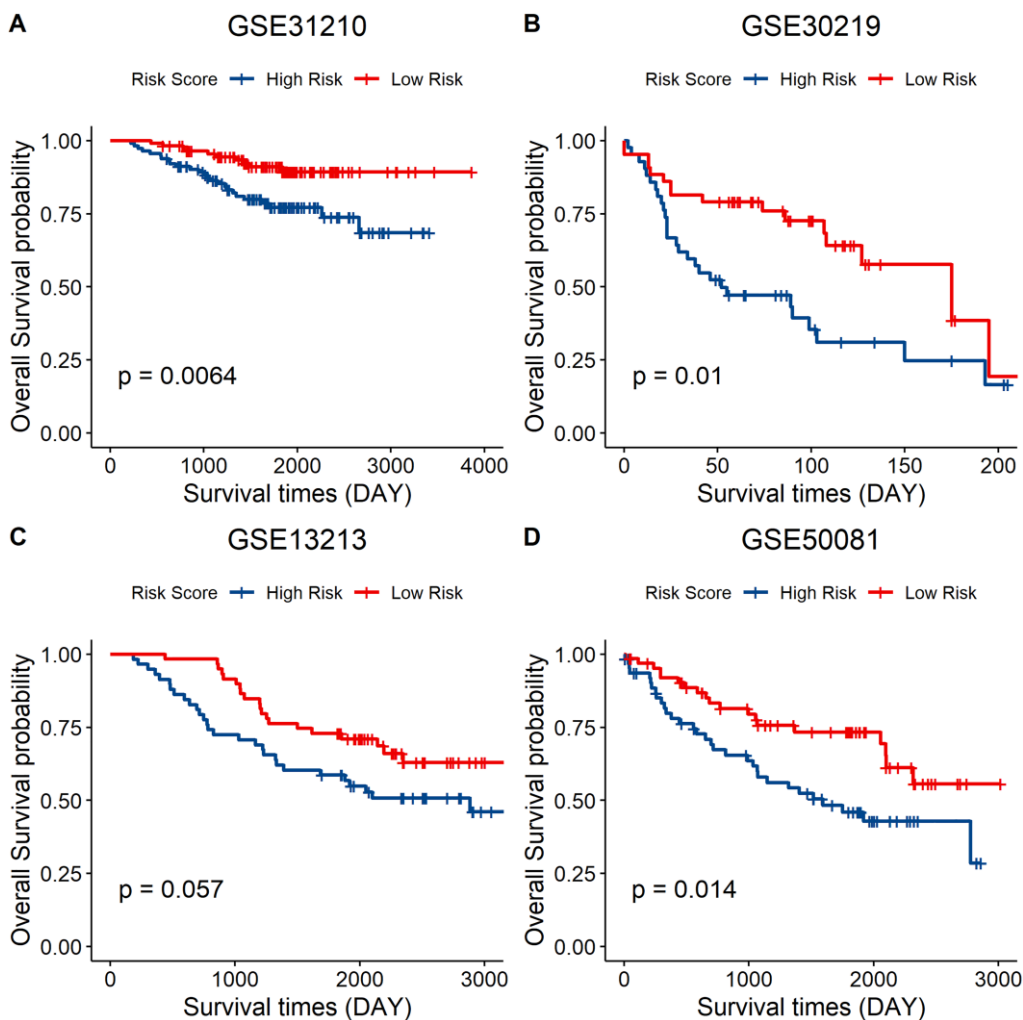
Figure 8. External validation of the eight-gene signature model. (**A–D**) Survival curve of GSE31210 (*p* = 0.0064), GSE30219 (*p* = 0.01), GSE13213 (*p* = 0.057), and GSE50081 (*p* = 0.014) indicating that a lower risk score was associated with better overall survival outcomes. Red, low‑risk group; blue, high‑risk group.

in the formation of β-D-glucuronides. The function of UGT2B17 is to maintain androgen homeostasis in the prostate. Elevated UGT2B17 expression levels enhances the progression of castration-resistant prostate cancer by promoting independent androgen receptor signaling and cancer cell mitosis [21]. The cell maintains its cellular homeostasis by autophagy, whereby it consumes its organelles using lysosomes to generate energy for protein synthesis [22]. G protein γ 7 (GNG7) is a heterotrimeric G protein subunit. It has been reported that GNG7 inhibits cell division and induces cell autophagy by inhibiting the mTOR pathway [23]. Moreover, it has been reported that GNG7 is a tumor suppressor gene in clear cell renal cell carcinoma and a lower expression of GNG7 predicts poor overall survival outcomes [24]. ST3GAL6 is a member of the sialyltransferase (STs) family. Overexpression of STs contribute to hypersialylation on the cell surface. Hypersialylation plays an important role in tumor progression. Cancer cells can recruit Siglec-7 and increase sialylated glycans on its surface, thereby protecting it from being killed by natural killer cells [25]. ST3GAL6 expression is significantly suppressed in hepatocellular carcinoma patients [26], and it mediates colorectal cancer progression through the PI3K/Akt signaling [27, 28]. Indoleamine 2,3-dioxygenase 2 (IDO2) is a member of the tryptophan catabolic enzyme. Studies have shown that IDO2 modulates dendritic cell functioning [28], and contributes to immune tolerance by controlling the Treg population [28]. However, the relationship between the IDO2 associated immune response and tumor progression has not been elucidated [29].

The aim of immunotherapy is to reactivate T cells by blocking the interaction of PD-1/PD-L1, thereby, inhibiting PD-1 signaling [30]. In this study, the risk score calculated from the eight-gene signature model was negatively correlated with the expression level of the commonly used immune markers such as PD-L1, PD-1, and CTLA-4, indicating that high risk score patients would not benefit from immune therapy. Correlation analysis showed that the expression level of the eight genes was significantly correlated with the level of immune checkpoint molecules. This finding shows that these genes function in immune responses. PIK3CG is highly expressed in TME and it prevents T cell mediated tumor elimination [31]. Therefore, the eight signature genes may provide clues on the different immunotherapeutic responses. Moreover, a lower risk score was correlated with a higher ssGSEA score of effector memory CD8+ T cell, natural killer cell, and macrophages. This implies that lower risk patients have a higher infiltration of immune cells and, therefore, have a higher probability of benefiting from immune

therapy. Therefore, the eight-gene based signature score has the potential for predicting immunotherapeutic responses and LUAD prognosis.

However, this study is associated with several limitations. First, it is a retrospective study. Further prospective studies are required to validate our findings. Secondly, the study did not verify the ability of the eight genes signature to predict immune responses in LUAD patients because we lacked the clinical data on patients receiving immunotherapy. Finally, other patient characteristics such as age, tumor size, and lymph node status were not included in our prognostic analysis.

MATERIALS AND METHODS

Data collection and pre-processing

Clinical data, TCGA RNA-seq data, and probe annotation files of the LUAD patients were downloaded from the UCSC xena browser (https://xenabrowser.net/) and used to obtain the gene expression profiles of the human LUAD patients. Data for the normal tissue was discarded while samples with no clinical data were excluded. Finally, a total of 513 tumor samples were retained and the TCGA LUAD cohort was randomly divided into training and testing data set at a ratio of 3:1 (training data set $n = 385$; testing data set $n = 128$). The two groups exhibited a similar estimate score distribution and other clinical characteristics. GSE31210, GSE30219, GSE13213, and GSE50081 datasets were downloaded from Gene Expression Omnibus (GEO) database in R using R package "GEOquery" [32]. The probe IDs were then transformed into gene symbols according to the annotation files, and the cancer metabolism gene set was obtained from ccmGDB (https://bioinfo.uth.edu/ccmGDB/).

Evaluation of the immune score and stromal score

The R package "ESTIMATE" was used to infer the fraction of immune cells and stromal cells in the patient's tumor samples [33]. The ESTIMATE algorithm was designed to calculate the immune and stromal scores of each sample based on the expression of certain stromal cell and immune cell genes. The results obtained from the ESTIMATE algorithm are presented in three categories; where immune score represents the score of immune cell infiltration, stromal score represents stromal cell infiltration, and estimate score represents the sum of both the immune score and stromal score. The LUAD samples were divided into two groups based on the expression level of the estimate score of each sample.

Determination of the infiltration level of immune cells in LUAD

The R package "GSVA" uses a method called Gene Set Variation Analysis (GSVA), which implements a non-parametric unsupervised method to assess the underlying pathway activity using gene expression microarray and RNA-seq data. By applying this approach, we can use the traditional analytical methods such as correlation analysis and survival analysis in a pathway focused manner. A set of immune cells' gene marker, consisting of 782 genes representing 28 immune cell types from innate and adaptive immunity, was obtained from a previous study and used to evaluate the infiltration of different immune cell types in the tumor microenvironment [34]. The immune cell types included dendritic cells, B cells, NK cells, MDSC, neutrophils, T cells, among others. Subsequently, the ssGSEA algorithm from R package "GSVA" was used to determine the infiltration level of each immune cell type in LUAD using the expression profiles [35]. Each sample was evaluated using the gene signature expressed by the immune cell and the calculation was then performed using the ssGSEA algorithm.

Identification of Differentially Expressed Genes (DEGs) and functional enrichment analysis

The LUAD cohort was divided into two groups based on the median of the estimate score. R package "limma" was used to determine the DEGs between the two groups [36]. Adjusted $p < 0.05$ and $|\log FC| > 0.5$ were set as the cutoff criteria to determine the significantly differentially expressed genes. The gene Ontology system of classification was used to classify genes into different gene sets based on their functions. The genes were then assigned with their GO terms. Gene Ontology (GO) term enrichment can be used to interpret which GO terms are over-expressed or under-expressed when given a set of up-regulated or down-regulated genes. The metabolic genes with a criteria of adjusted $p < 0.05$ were used to explore the metabolic landscape between the two groups. The metabolic genes were then used in the GO analysis using R package "ClusterProfiler" to identify the enriched biological pathways [37]. A cutoff value of 0.05 was set to obtain significant results for biological process (BP), cellular components (CC), and molecular functions (MF). R package "ClusterProfiler" was then used to visualize the GO enrichment results.

Generation of a prognostic model using LASSO Regularization to evaluate the mutation profile of the signature genes

Least absolute shrinkage and selection operator (LASSO) is a type of linear regression. Adding a penalty equal to the absolute value of the magnitude of some coefficients can result in the coefficients becoming zero, and thus they can be removed from the model. Therefore, a model with few coefficients can then be created. The candidate gene expression profiles were obtained from the training data set ($n = 385$) and R package "glmnet" was used to perform LASSO regularization to reduce the coefficients. This was followed by selection of the most robust markers to construct the risk score signature, which included eight genes. The following formula was used:

Risk Score = 0.021 × PLAUR – 0.017 × ST3GAL6 – 0.004 × GNG7 – 0.038 × PIK3CG – 0.048 × GLS2 – 0.007 × IDO2 – 0.004 × UGT2B17 + 0.016 × F2.

"Maftools" was selected for mutation analysis based on the model constructed by LASSO regularization to explore the mutation frequency of the signature genes in LUAD [38].

Testing data set validation

The expression profiles of the signature genes were extracted from the testing data set, then used in the prognosis model for calculation. The predicted risk score was calculated and its association with survival outcomes was further analyzed.

Survival analysis

Univariate Cox proportional hazards regression analysis was used to evaluate the association between the expression level of the metabolic DEGs and the overall survival (OS) of LUAD patients. Metabolic DEGs with $p < 0.05$ based on the log-rank test were selected as candidate genes for construction of the prognosis model. The risk score for each sample was calculated based on the signature model to evaluate the association between the gene signature and the prognosis of LUAD patients. The samples were classified into either high risk or low risk groups depending on the median risk score. Kaplan-Meier curve and log-rank test were used to compare the differences in overall survival and progression free survival outcomes between the predicted high risk and low risk groups. $p \leq 0.05$ was set as the significant level. All the survival analyses and log-rank tests were performed using R package survival, while the R package "surviminer" was used to plot the Kaplan-Meier curve.

Statistical analysis

Univariate analysis of survival outcomes was performed using the log-rank test. The correlation relationships

between the risk score and immune markers, the risk score and signature gene expression, the signature gene expression and immune cell infiltration score, and the risk score and immune cell infiltration score were determined by Pearson correlation. A two-tailed student t-test was used to compare the two groups. $p \leq 0.05$ was set as the threshold for statistical significance. All statistical analyses were performed in R version 4.0.2.

Data availability statement

The data that supports this study is available in GSE31210, GSE30219, GSE13213, and GSE50081 at https://www.ncbi.nlm.nih.gov/gds. TCGA RNA-seq and clinical data of the LUAD patients was downloaded from the UCSCxena browser (https://xenabrowser.net/), while the cancer cell metabolism gene set was downloaded from ccmGDB (https://bioinfo.uth.edu/ccmGDB/).

AUTHOR CONTRIBUTIONS

C. Z. made substantial contributions in conception, design, interpretation, and preparation of the final manuscript. L. J., M.W., L. J., C. Z. and T.D. participated in the coordination of data acquisition, data analysis, and reviewed the manuscript.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interests.

FUNDING

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020; 70:7–30.
   https://doi.org/10.3322/caac.21590
   PMID:31912902

2. Remon J, Caramella C, Jovelet C, Lacroix L, Lawson A, Smalley S, Howarth K, Gale D, Green E, Plagnol V, Rosenfeld N, Planchard D, Bluthgen MV, et al. Osimertinib benefit in EGFR-mutant NSCLC patients with T790M-mutation detected by circulating tumour DNA. Ann Oncol. 2017; 28:784–790.
   https://doi.org/10.1093/annonc/mdx017
   PMID:28104619

3. Ramalingam SS, Vansteenkiste J, Planchard D, Cho BC, Gray JE, Ohe Y, Zhou C, Reungwetwattana T, Cheng Y, Chewaskulyong B, Shah R, Cobo M, Lee KH, et al. Overall Survival with Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC. N Engl J Med. 2020; 382:41–50.
   https://doi.org/10.1056/nejmoa1913662
   PMID:31751012

4. Gandhi L, Rodriguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, Domine M, Clingan P, Hochmair MJ, Powell SF, Cheng SY, Bischoff HG, Peled N, et al. Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. N Engl J Med. 2018; 378:2078–2092.
   https://doi.org/10.1056/nejmoa1801005
   PMID:29658856

5. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer. 2019; 19:133–150.
   https://doi.org/10.1038/s41568-019-0116-x
   PMID:30755690

6. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–674.
   https://doi.org/10.1016/j.cell.2011.02.013
   PMID:21376230

7. Fridman WH, Zitvogel L, Sautes-Fridman C, Kroemer G. The immune contexture in cancer prognosis and treatment. Nat Rev Clin Oncol. 2017; 14:717–734.
   https://doi.org/10.1038/nrclinonc.2017.101
   PMID:28741618

8. O'Sullivan D, Sanin DE, Pearce EJ, Pearce EL. Metabolic interventions in the immune response to cancer. Nat Rev Immunol. 2019; 19:324–335.
   https://doi.org/10.1038/s41577-019-0140-9
   PMID:30820043

9. Chang CH, Qiu J, O'Sullivan D, Buck Michael D, Noguchi T, Curtis Jonathan D, Chen Q, Gindin M, Gubin Matthew M, van der Windt Gerritje JW, Tonc E, Schreiber Robert D, Pearce Edward J, et al. Metabolic Competition in the Tumor Microenvironment Is a Driver of Cancer Progression. Cell. 2015; 162:1229–1241.
   https://doi.org/10.1016/j.cell.2015.08.016
   PMID:26321679

10. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Rev Cancer. 2014; 14:535–546.
    https://doi.org/10.1038/nrc3775
    PMID:25056707

11. Wegiel B, Vuerich M, Daneshmandi S, Seth P. Metabolic Switch in the Tumor Microenvironment Determines Immune Responses to Anti-cancer Therapy. Front Oncol. 2018; 8:284.
    https://doi.org/10.3389/fonc.2018.00284
    PMID:30151352

12. Leonetti A, Sharma S, Minari R, Perego P, Giovannetti E, Tiseo M. Resistance mechanisms to osimertinib in EGFR-mutated non-small cell lung cancer. Br J Cancer. 2019; 121:725–737.
https://doi.org/10.1038/s41416-019-0573-8
PMID:31564718

13. Bianchi F, Nuciforo P, Vecchi M, Bernard L, Tizzoni L, Marchetti A, Buttitta F, Felicioni L, Nicassio F, Di Fiore PP. Survival prediction of stage I lung adenocarcinomas by expression of 10 genes. J Clin Invest. 2007; 117:3436–3444.
https://doi.org/10.1172/jci32007
PMID:17948124

14. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med. 2007; 356:11–20.
https://doi.org/10.1056/nejmoa060096
PMID:17202451

15. Shukla S, Evans JR, Malik R, Feng FY, Dhanasekaran SM, Cao X, Chen G, Beer DG, Jiang H, Chinnaiyan AM. Development of a RNA-Seq Based Prognostic Signature in Lung Adenocarcinoma. J Natl Cancer Inst. 2016; 109:djw200.
https://doi.org/10.1093/jnci/djw200
PMID:27707839

16. Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao MS, Penn LZ, Jurisica I. Prognostic gene signatures for non-small-cell lung cancer. Proc Natl Acad Sci U S A. 2009; 106:2824–2828.
https://doi.org/10.1073/pnas.0809444106
PMID:19196983

17. Hou YC, Chao YJ, Hsieh MH, Tung HL, Wang HC, Shan YS. Low CD8(+) T Cell Infiltration and High PD-L1 Expression Are Associated with Level of CD44(+)/CD133(+) Cancer Stem Cells and Predict an Unfavorable Prognosis in Pancreatic Cancer. Cancers (Basel). 2019; 11:541.
https://doi.org/10.3390/cancers11040541
PMID:30991694

18. Patel SP, Kurzrock R. PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. Mol Cancer Ther. 2015; 14:847–856.
https://doi.org/10.1158/1535-7163.mct-14-0983
PMID:25695955

19. Foubert P, Kaneda MM, Varner JA. PI3Kgamma Activates Integrin alpha4 and Promotes Immune Suppressive Myeloid Cell Polarization during Tumor Progression. Cancer Immunol Res. 2017; 5:957–968.
https://doi.org/10.1158/2326-6066.cir-17-0143
PMID:28963139

20. Ye X, Zhou Q, Matsumoto Y, Moriyama M, Kageyama S, Komatsu M, Satoh S, Tsuchida M, Saijo Y. Inhibition of Glutaminolysis Inhibits Cell Growth via Down-regulating Mtorc1 Signaling in Lung Squamous Cell Carcinoma. Anticancer Res. 2016; 36:6021–6030.
https://doi.org/10.21873/anticanres.11191
PMID:27793929

21. Li H, Xie N, Chen R, Verreault M, Fazli L, Gleave ME, Barbier O, Dong X. UGT2B17 Expedites Progression of Castration-Resistant Prostate Cancers by Promoting Ligand-Independent AR Signaling. Cancer Res. 2016; 76:6701–6711.
https://doi.org/10.1158/0008-5472.can-16-1518
PMID:27659047

22. Rabinowitz JD, White E. Autophagy and metabolism. Science. 2010; 330:1344–1348.
https://doi.org/10.1126/science.1193497
PMID:21127245

23. Liu J, Ji X, Li Z, Yang X, Wang W, Zhang X. G protein gamma subunit 7 induces autophagy and inhibits cell division. Oncotarget. 2016; 7:24832–24847.
https://doi.org/10.18632/oncotarget.8559
PMID:27056891

24. Xu S, Zhang H, Liu T, Chen Y, He D, Li L. G Protein gamma subunit 7 loss contributes to progression of clear cell renal cell carcinoma. J Cell Physiol. 2019; 234:20002–20012.
https://doi.org/10.1002/jcp.28597
PMID:30945310

25. Hudak JE, Canham SM, Bertozzi CR. Glycocalyx engineering reveals a Siglec-based mechanism for NK cell immunoevasion. Nat Chem Biol. 2014; 10:69–75.
https://doi.org/10.1038/nchembio.1388
PMID:24292068

26. Souady J, Hülsewig M, Distler U, Haier J, Denz A, Pilarsky C, Senninger N, Dreisewerd K, Peter-Katalinić J, Müthing J. Differences in CD75s- and iso-CD75s-ganglioside content and altered mRNA expression of sialyltransferases ST6GAL1 and ST3GAL6 in human hepatocellular carcinomas and nontumoral liver tissues. Glycobiology. 2011; 21:584–594.
https://doi.org/10.1093/glycob/cwq200
PMID:21147760

27. Pallotta MT, Orabona C, Volpi C, Vacca C, Belladonna ML, Bianchi R, Servillo G, Brunacci C, Calviti M, Bicciato S, Mazza EM, Boon L, Grassi F, et al. Indoleamine 2,3-dioxygenase is a signaling protein in long-term tolerance by dendritic cells. Nat Immunol. 2011; 12:870–878.
https://doi.org/10.1038/ni.2077
PMID:21804557

28. Trabanelli S, Ocadlikova D, Ciciarello M, Salvestrini V, Lecciso M, Jandus C, Metz R, Evangelisti C, Laury-Kleintop L, Romero P, Prendergast GC, Curti A, Lemoli RM. The SOCS3-independent expression of IDO2 supports the homeostatic generation of T regulatory cells by human dendritic cells. J Immunol. 2014; 192:1231–1240.
https://doi.org/10.4049/jimmunol.1300720
PMID:24391212

29. Merlo LM, Mandik-Nayak L. IDO2: A Pathogenic Mediator of Inflammatory Autoimmunity. Clin Med Insights Pathol. 2016; 9:21–28.
https://doi.org/10.4137/cpath.s39930
PMID:27891058

30. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, Chmielowski B, Spasic M, Henry G, Ciobanu V, West AN, Carmona M, Kivork C, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature. 2014; 515:568–571.
https://doi.org/10.1038/nature13954
PMID:25428505

31. Kaneda MM, Messer KS, Ralainirina N, Li H, Leem CJ, Gorjestani S, Woo G, Nguyen AV, Figueiredo CC, Foubert P, Schmid MC, Pink M, Winkler DG, et al. PI3Kgamma is a molecular switch that controls immune suppression. Nature. 2016; 539:437–442.
https://doi.org/10.1038/nature19834
PMID:27642729

32. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007; 23:1846–1847.
https://doi.org/10.1093/bioinformatics/btm254
PMID:17496320

33. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevino V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013; 4:2612.
https://doi.org/10.1038/ncomms3612
PMID:24113773

34. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H, Trajanoski Z. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Rep. 2017; 18:248–262.
https://doi.org/10.1016/j.celrep.2016.12.019
PMID:28052254

35. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013; 14:7.
https://doi.org/10.1186/1471-2105-14-7
PMID:23323831

36. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47.
https://doi.org/10.1093/nar/gkv007
PMID:25605792

37. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16:284–287.
https://doi.org/10.1089/omi.2011.0118
PMID:22455463

38. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018; 28:1747–1756.
https://doi.org/10.1101/gr.239244.118
PMID:30341162
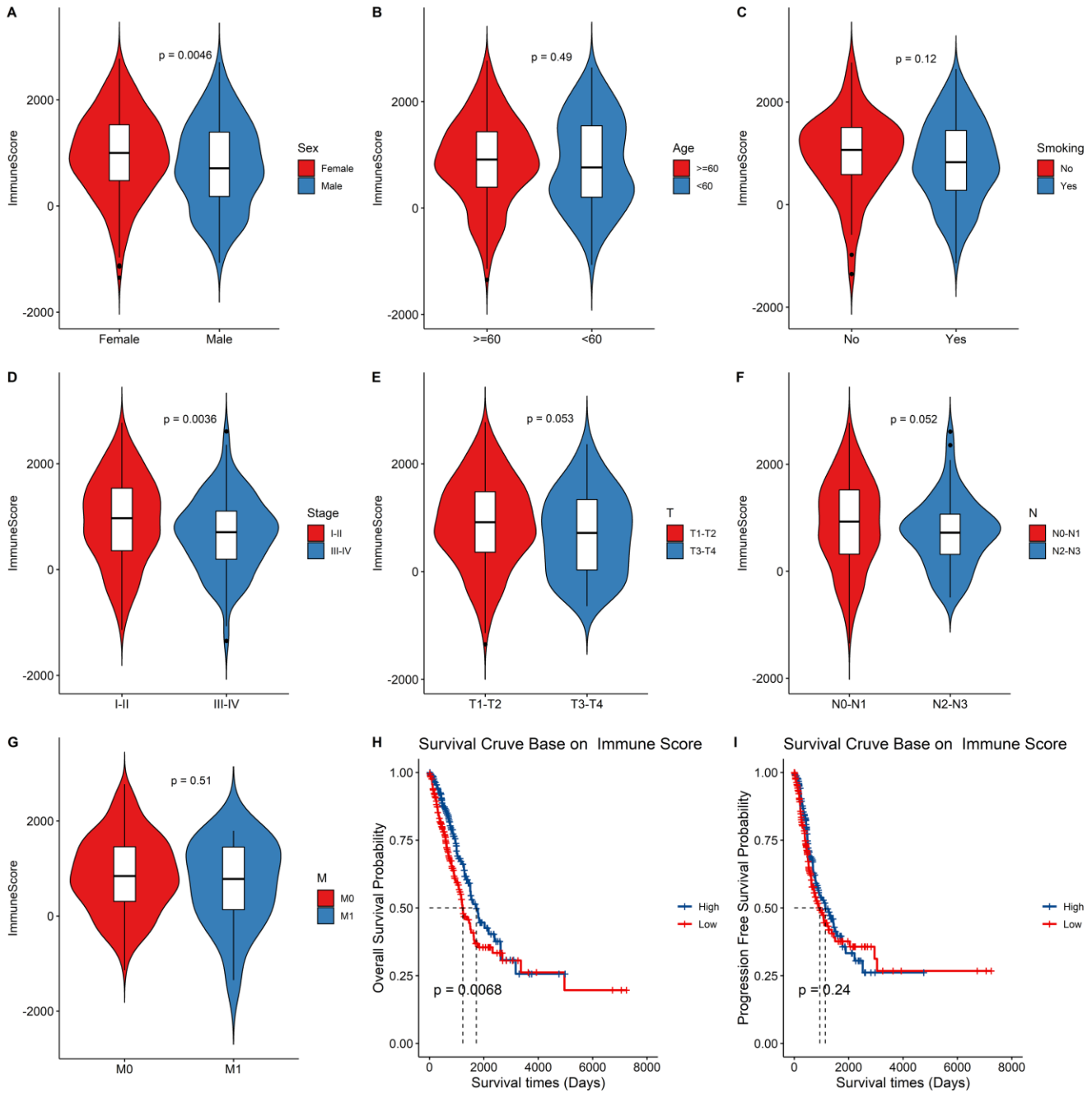
Supplementary Figure 1. Relationship between the Immune score and clinical status, survival outcomes. (**A–G**) Boxplot represent the difference between the immune score and clinical characteristics. *p* value above the boxplot indicates the difference between the two groups. (**H–I**) The Kaplan–Meier curves for overall survival and progression free survival of LUAD risk groups divided using the median cutoff point of immune score.
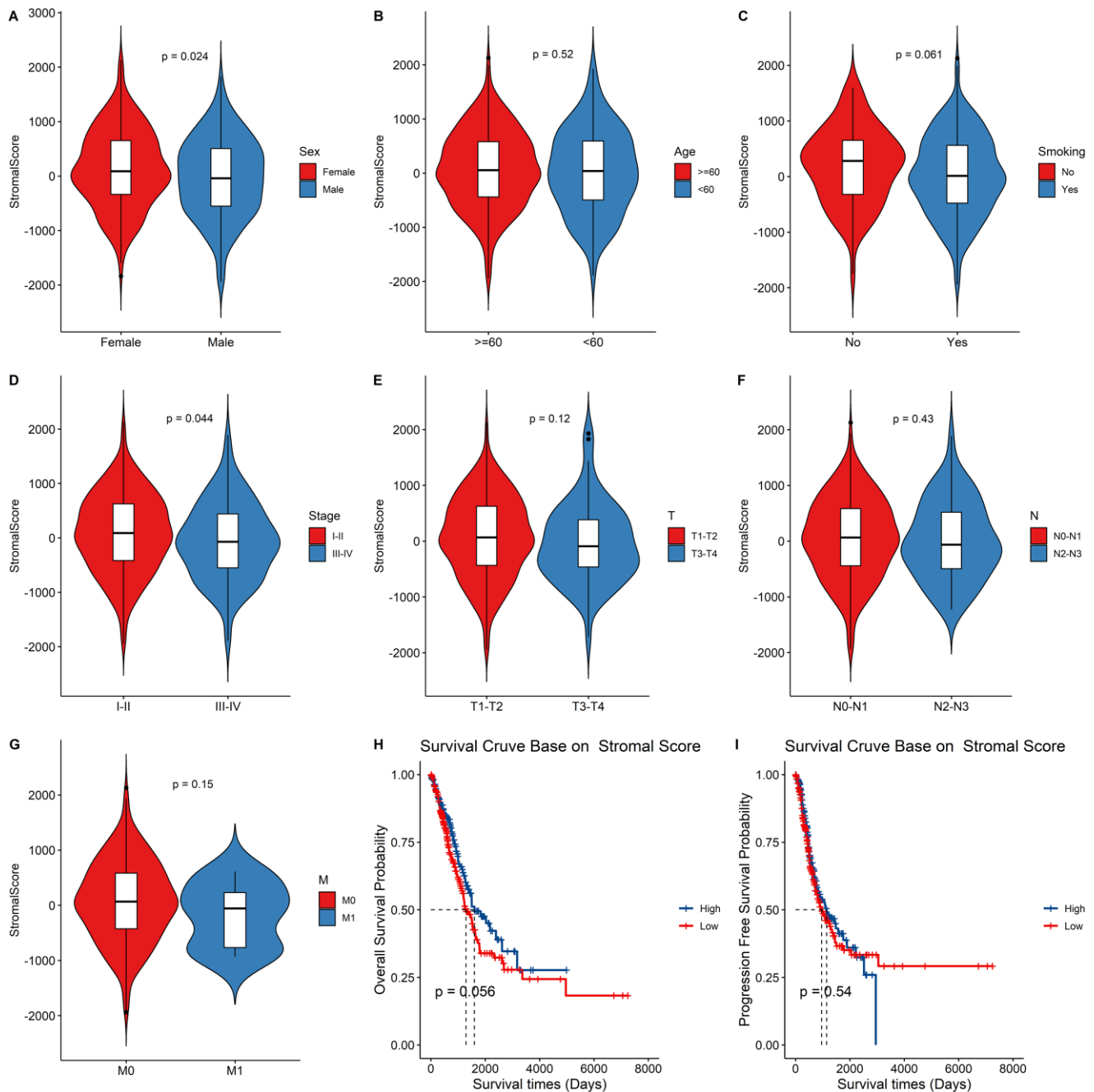
Supplementary Figure 2. Relationship between the Stromal score and clinical status, survival outcome. (**A–G**) Boxplot representing the difference between the stromal score and the clinical characteristic. *p* value above the boxplot indicates the difference between the two groups. (**H–I**) The Kaplan–Meier curves for overall survival and progression free survival of LUAD risk groups divided using the median cutoff point of Stromal score.
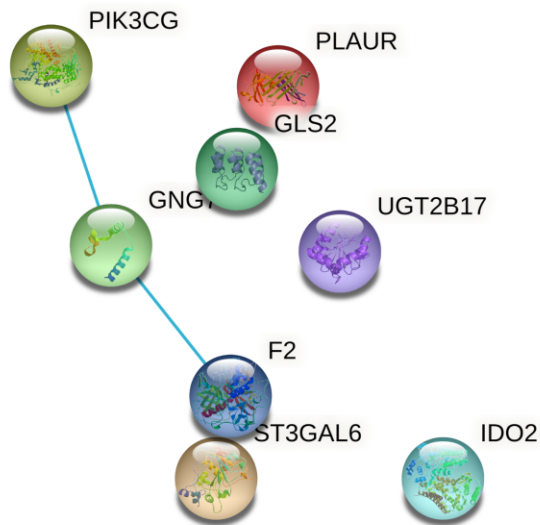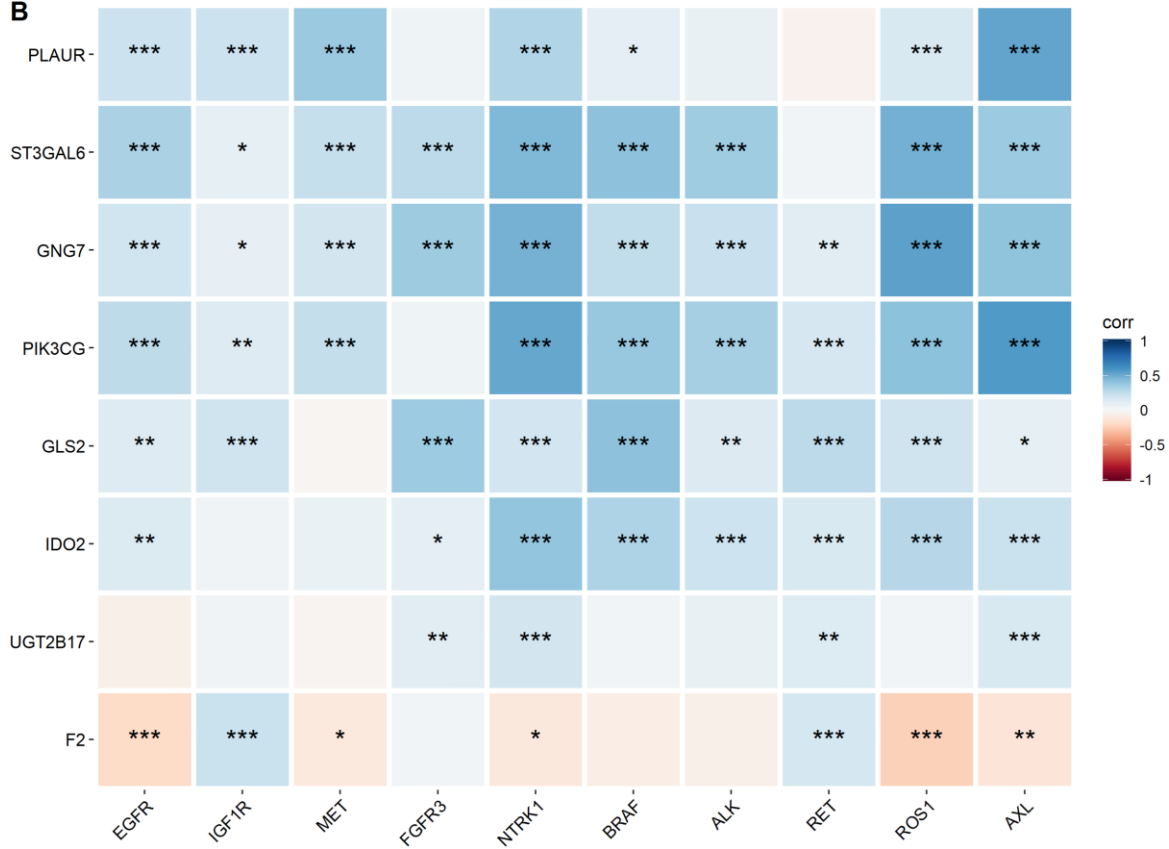
Supplementary Figure 3. (**A**) Analysis of the interactions among the eight signature genes using the STRING-DB database. (**B**) Correlation analysis of expression of the eight signature genes and LUAD oncogenic drivers and bypass signaling.