World Journal of
Clinical Oncology

Contents

ABOUT COVER

Editorial Board Member of *World Journal of Clinical Oncology*, Mohamed Elbadawy, BSc, MD, Ph.D., (Med), Associate Professor, Department of Pharmacology, Faculty of Veterinary Medicine, Benha University, 13736 Moshtohor, Toukh, Elqaliobiya, Egypt. mohamed.elbadawy@fvtm.bu.edu.eg

AIMS AND SCOPE

The primary aim of *World Journal of Clinical Oncology* (*WJCO*, *World J Clin Oncol*) is to provide scholars and readers from various fields of oncology with a platform to publish high-quality basic and clinical research articles and communicate their research findings online.

   WJCO mainly publishes articles reporting research results and findings obtained in the field of oncology and covering a wide range of topics including art of oncology, biology of neoplasia, breast cancer, cancer prevention and control, cancer-related complications, diagnosis in oncology, gastrointestinal cancer, genetic testing for cancer, gynecologic cancer, head and neck cancer, hematologic malignancy, lung cancer, melanoma, molecular oncology, neurooncology, palliative and supportive care, pediatric oncology, surgical oncology, translational oncology, and urologic oncology.

INDEXING/ABSTRACTING

The *WJCO* is now abstracted and indexed in PubMed, PubMed Central, Emerging Sources Citation Index (Web of Science), Reference Citation Analysis, China National Knowledge Infrastructure, China Science and Technology Journal Database, and Superstar Journals Database. The 2022 edition of Journal Citation Reports® cites the 2021 Journal Citation Indicator (JCI) for *WJCO* as 0.35.

RESPONSIBLE EDITORS FOR THIS ISSUE

ORIGINAL ARTICLE

Basic Study

5-mRNA-based prognostic signature of survival in lung adenocarcinoma

Qian-Lin Xia, Xiao-Meng He, Yan Ma, Qiu-Yue Li, Yu-Zhen Du, Jin Wang

Qian-Lin Xia, Yu-Zhen Du, Laboratory Medicine, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200233, China

Xiao-Meng He, Yan Ma, Qiu-Yue Li, Jin Wang, Scientific Research Center, Shanghai Public Health Clinical Center, Fudan University, Shanghai 201508, China

Corresponding author: Jin Wang, PhD, Professor, Scientific Research Center, Shanghai Public Health Clinical Center, Fudan University, No 2901 Caolang Road, Jinshan District, Shanghai 201508, China. wjincityu@yahoo.com

Abstract

BACKGROUND
Lung adenocarcinoma (LUAD) is the most common non-small-cell lung cancer, with a high incidence and a poor prognosis.

AIM
To construct effective predictive models to evaluate the prognosis of LUAD patients.

METHODS
In this study, we thoroughly mined LUAD genomic data from the Gene Expression Omnibus (GEO) (GSE43458, GSE32863, and GSE27262) and the Cancer Genome Atlas (TCGA) datasets, including 698 LUAD and 172 healthy (or adjacent normal) lung tissue samples. Univariate regression and LASSO regression analyses were used to screen differentially expressed genes (DEGs) related to patient prognosis, and multivariate Cox regression analysis was applied to establish the risk score equation and construct the survival prognosis model. Receiver operating characteristic curve and Kaplan-Meier survival analyses with clinically independent prognostic parameters were performed to verify the predictive power of the model and further establish a prognostic nomogram.

RESULTS
A total of 380 DEGs were identified in LUAD tissues through GEO and TCGA datasets, and 5 DEGs (TCN1, CENPF, MAOB, CRTAC1 and PLEK2) were screened out by multivariate Cox regression analysis, indicating that the prognostic risk model could be used as an independent prognostic factor (Hazard ratio = 1.520, $P < 0.001$). Internal and external validation of the model confirmed that the prediction model had good sensitivity and specificity (Area under the

curve = 0.754, 0.737). Combining genetic models and clinical prognostic factors, nomograms can also predict overall survival more effectively.

CONCLUSION
A 5-mRNA-based model was constructed to predict the prognosis of lung adenocarcinoma, which may provide clinicians with reliable prognostic assessment tools and help clinical treatment decisions.

Key Words: Lung adenocarcinoma; Differentially expressed genes; Prognostic signature; Risk score; Nomogram

Core Tip: Five differentially expressed genes (DEGs) (TCN1, CENPF, MAOB, CRTAC1, and PLEK2) selected by multiple Cox regression analysis in the prognostic risk models could be considered as independent prognostic factors for lung adenocarcinoma.

INTRODUCTION

Lung adenocarcinoma (LUAD) is a common histological type of lung cancer that is a malignant tumor that seriously threatens human health, accounting for approximately 40% of lung cancers[1]. In recent years, some progress has been made in diagnostic and treatment strategies of clinical and experimental oncology for lung cancer. However, LUAD patients with localized or locally advanced disease still have a high risk of death, and their 5-year overall survival rate is still less than 15%[2]. Assessing the patient's prognosis can help choose effective treatments to balance side effects with treatment benefits and decide whether to give more aggressive treatment. Although tumor–node–metastasis (TNM) classification plays an important role in the prognosis assessment of LUAD patients, the prognosis of some patients is significantly different even if the stages are similar. Therefore, the identification of reliable prognostic biomarkers to predict clinical outcomes and help make accurate clinical treatment decisions is clearly critical. The rapid development of gene chips and high-throughput sequencing have facilitated the development of new predictive tools based on prognostic genes for lung cancer. These relevant studies involved in prognostic genes of lung cancer have identified several prognostic models that have predicted the overall survival rate of LUAD patients (Table 1)[3-14]. For example, a six-gene model (RRAGB, RSPH9, RPS6KL1, RXFP1, RTL1 and RRM2) based on the weighted gene coexpression network predicted the overall survival rate of non-small-cell lung cancer patients[13]. A 3-gene prognostic model (URKA, CDC20, and TPX2A) also accurately predicted overall survival in smoking-related lung adenocarcinoma[14]. In addition, through analysis of TCGA data, the risk score of the 12-mRNA signature was correlated with poor prognosis in patients with lung adenocarcinoma[3]. Therefore, the in-depth exploration of public databases such as the Gene Expression Omnibus (GEO) and the Cancer Genome Atlas (TCGA) databases, discovery of other genes related to the prognosis of LUAD and development of a comprehensive prognosis assessment system including multiple biomarkers may be effective ways to predict the prognosis of lung adenocarcinoma and individual treatment.

Here, we first integrated three lung adenocarcinoma datasets from the GEO database to screen for differentially expressed genes (DEGs). Then, the TCGA-LUAD data set was used to identify DEGs. Univariate Cox and LASSO regression analyses were further used to determine the DEGs associated with overall survival. The risk score was calculated by multiplying multiple Cox coefficients by gene expression. The prognostic model was also combined with clinical parameters to construct a prognostic nomogram to predict overall survival. Finally, Gene set enrichment analysis (GSEA) was performed to identify the potential biological pathways of the five genes in the model.

Table 1 Data and studies involved several prediction models for the prognostic signature of non-small cell lung cancer/ lung adenocarcinoma

Datasets	Model-related genes	Application	Ref.
GEO (GSE19188, GSE33532)	DLGAP5, KIF11, RAD51AP1, CCNB1, AURKA, CDC6, OIP5, NCAPG	Prognostic signature for predicting overall survival in lung adenocarcinoma	Li *et al*[3], 2018
GEO (GSE31210, GSE37745, GSE50081) + TCGA	PLEKHH2, ISCU, CLUL1, CHRDL1, PAIP2B, CDCP1	Prognostic signature for predicting both disease-free and overall survival in non-small cell lung cancer	Zuo *et al*[4], 2019
GEO (GSE31210, GSE37745, GSE50081) + TCGA	CDCP1, HMMR, TPX2, CIRBP, HLF, KBTBD7, SEC24B-AS1, SH2B1	Prognostic Signature for predicting overall survival of early-stage non-small cell lung cancer	He *et al*[5], 2019
TCGA	RHOV, CD109, LINC00941, FRRS1	Prognostic signature for predicting overall survival in lung adenocarcinoma	Shukla *et al*[6], 2016
GEO (GSE3141, GSE30219, GSE50081) + TCGA	ADAM12, BTK, ERG	Prognostic gene signature associated with the microenvironment of lung adenocarcinoma	Yue *et al*[7], 2019
GEO (GSE50081, GSE30219, GSE31210, GSE19188, GSE37745, GSE3141, GSE31908)	ABCC4, ADRBK2, KLHL23, PDS5A, UHRF1, ZNF551	Prognostic signature for predicting overall survival in non-small cell lung cancer	Huang *et al*[8], 2016
GEO (GSE50081, GSE31210, GSE30219, GSE29013, GSE68465, GSE42127) + E-MTAB-923	STAT1, CLU, GTSE1, NUSAP1, ABCA8, TNNT1, ENTPD3, CPA3	Prognostic signature for predicting overall survival in non-small cell lung cancer	Shahid *et al*[9], 2016
GEO (GSE8894, GSE14814, GSE30219, GSE31210, GSE37745, GSE50081)	KIF15, DLGAP5, ASPM, ADAM10, RAD51AP1, FGFR10P, NCGAP	Prognostic gene expression signature for early stage lung adenocarcinoma	Krzystanek *et al* [10], 2016
TCGA	BCHE, CCNA1, CYP24A1, DEPTOR, MASP2, MGLL, MYO1A, PODXL2, RAPGEF3, SGK2, TNNI2, ZBTB16	Prognostic signature for predicting overall survival in lung adenocarcinoma	Zengin *et al* [11], 2020
TCGA	PTPRH, OGFRP1, LDHA, AL365203.1, LINC02178, AL512488.1, LINC01312, AL353746.1, DRAXINP1, LINC02310	Prognostic signature for predicting overall survival in lung adenocarcinoma	Li *et al*[12], 2018
14 GEO datasets	ABCC4, ADRBK2, KLHL23, PDS5A, UHRF1, ZNF551	Prognostic signature for predicting overall survival in non-small cell lung cancer	Xie *et al*[13], 2019
GEO (GSE31210, GSE32863, GSE40791, GSE43458, GSE75037) + TCGA	AURKA, CDC20, TPX2	Prognostic signature for predicting overall survival in smoking-related lung adenocarcinoma	Zhang *et al*[14], 2019

GEO: the Gene Expression Omnibus; TCGA: the Cancer Genome Atlas.

MATERIALS AND METHODS

Gene expression profile data collection from the GEO and TCGA databases

The GEO database (https://www.ncbi.nlm.nih.gov/geo/) was used for the mRNA expression and clinical data of lung adenocarcinoma that needed to meet the following criteria: (1) Human lung adenocarcinoma tissue samples; (2) tumor and nontumor lung control tissue samples; and (3) ≥ 50 samples. Finally, three gene expression microarray data sets (GSE43458, GSE32863 and GSE27262), which included 163 LUAD tumor tissue samples and 113 adjacent normal tissue samples, were downloaded for DEG analysis. On the other hand, the original count data and corresponding clinical data of LUAD patients in the training set and test set, which includes 535 LUAD patient samples and 59 control samples, were downloaded from TCGA project (https://tcga-data.nci.nih.gov/tcga/). Complete survival information and gene expression profile data of 494 patients were obtained from the TCGA database after excluding samples that could not be assessed for tumor histological grade or had no overall survival (OS) information. The model was validated using transcriptome analysis of 90 LUAD patients from the GSE11969 dataset. The workflow of our LUAD biomarker analysis process is shown in Supplementary Figure 1.

Screening and verification of DEGs in lung adenocarcinoma tissue

To identify DEGs between LUAD and lung tissues, GE02R was used for differential expression analysis of the GSE43458, GSE32863, and GSE27262 data sets. The DEGs of the TCGA-LUAD dataset were analyzed using the "limma" software package of R software, and the threshold of DEG screening was |log FC| > 2 and $P < 0.05$ according to our previous study[15]. Human protein mapping (https://www.proteinatlas.org/) evaluates lung adenocarcinoma and DEG protein expression in normal lung tissue[16]. Mutation data in lung adenocarcinoma patients were obtained from cBioPortal (

https://www.cbioportal.org/)[17].

Identification of prognostic differential genes and establishment of prognostic models

TCGA-LUAD data were randomly divided into a training set (n = 346) and a test set (n = 148). In the test set, we performed univariate Cox regression analysis for DEGs determined by a comprehensive analysis of the GEO data set to determine the relationship between patient survival and gene expression. $P < 0.01$ was considered statistically significant and was included in subsequent analysis. Next, we applied LASSO regression to further reduce the number of DEGs in the selected panel with the best predictive performance by 10-fold cross-validation of the R-based glmnet package. Finally, multivariate Cox regression analysis was performed to obtain the five optimal prognostic gene regression coefficients from the multivariate Cox proportional hazard regression model. A prognostic risk score for the five genes was then established based on the multivariate Cox regression model regression coefficient multiplied by a linear combination of its mRNA expression level.

Identification of prognostic models and related genes

The Lung Adenocarcinoma (TCGA, PanCancer Atlas) database in cBioPortal was used to analyze the genetic mutation model. We used data from the TCGA to analyze model-related gene expression. The THPA (http://www.proteinatlas.org) database was used to analyze the protein expression of model-related genes[16]. Patients in the training set were divided into high-risk and low-risk groups according to the median risk score as the cutoff point. Kaplan-Meier (KM) survival curves and Wilcoxon tests combined with the R package "survival" were used to compare the survival differences between the high-risk and low-risk groups. Time-dependent receiver operating characteristic (ROC) curve analysis was conducted using the R software package "survivalROC" to assess the prediction model's forecasting capacity. $P < 0.05$ was considered statistically significant. The test cohort and the entire cohort were used for internal validation, the GSE11969 dataset was downloaded from the GEO database for external validation, and the risk score of each patient was calculated using the same model based on the prognostic gene signature to further verify the predictive value of the prognostic gene signature.

Establish and verify the forecast nomograms

To provide clinicians with a quantitative method for predicting 1-year, 3-year, and 5-year overall survival in LUAD patients, we used a combined model of all independent prognostic factors selected by multivariate Cox regression analysis to construct a nomogram. KM analysis, area under the curve (AUC), consistency index (C-index), and comparison of predicted and observed overall survival were used to evaluate the prognostic nomographs' performance[18].

Functional enrichment analysis of model genes

GSEA was used to analyze the signaling pathways of relevant genes involved in the development of lung adenocarcinoma to clarify the molecular mechanism of the prognostic gene signature. GSEA software (GSEA 4.0.3) was downloaded from the Broad Institute website (http://software.broadinstitute.org/gsea/index.jsp), and the analyzed access was from the c2.cp.kegg.v7.0.symbols.gmt data set in the Molecular Signature Database (MsigDB). The enrichment analysis was carried out by the weighted enrichment method, and the number of random combinations was set as 1000. All other parameters were set as default values. Gene sets with $P < 0.05$ were regarded as significantly enriched gene sets.
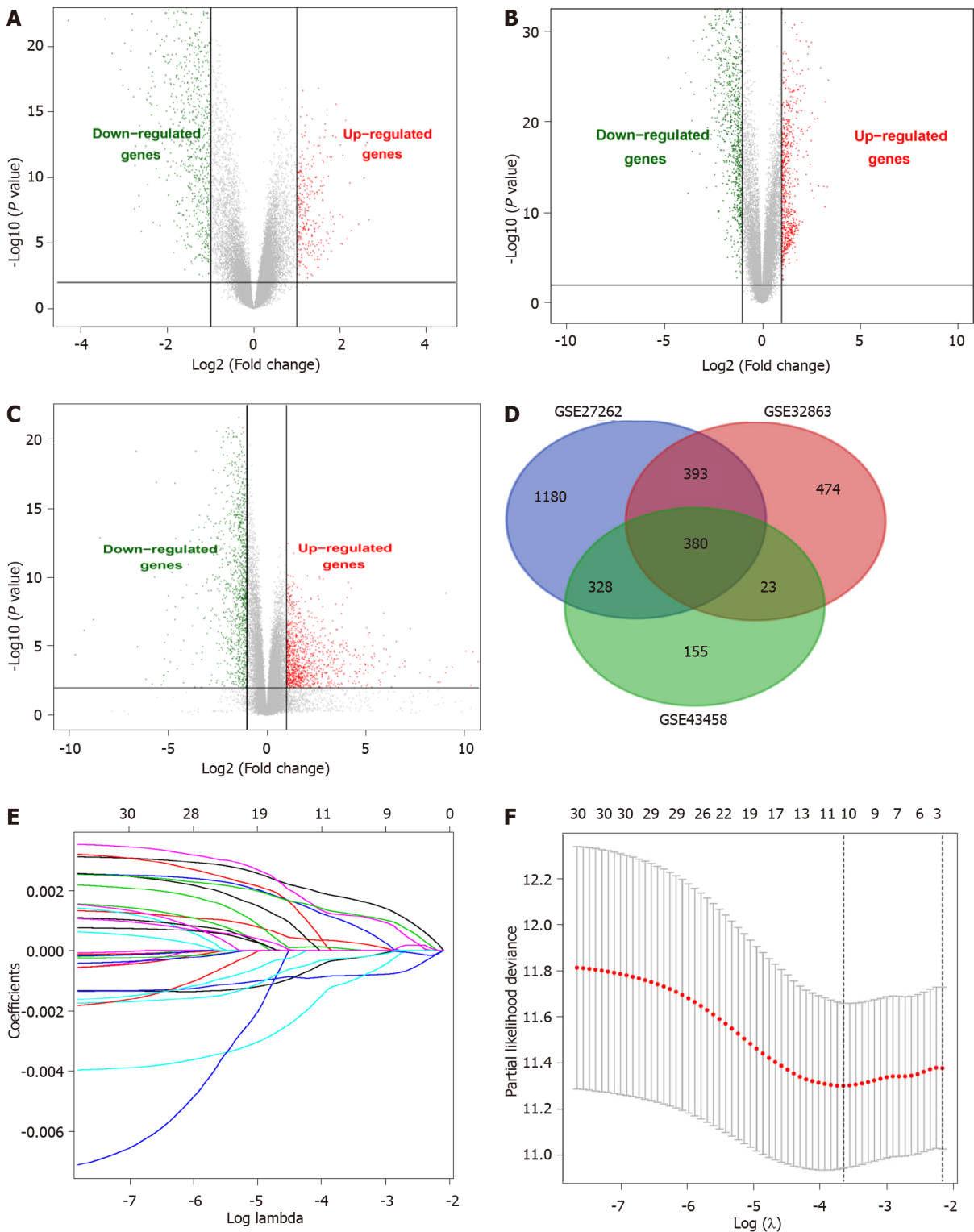
Statistical analysis

Statistical analysis and corresponding graph drawing were performed using R3.6.3 software, and Cox regression analysis of the hazard ratio (HR) and 95% confidence interval (CI) was used to evaluate the association between DEG expression and prognosis. A *t*-test of paired samples or a nonparametric Wilcoxon rank sum test of unpaired samples was used for analysis of continuous variables, and categorical variables were tested by the chi-square test or Fisher's exact test. $P < 0.05$ indicated a significant difference.

RESULTS

Screening and identification of differentially expressed genes in the TCGA-LUAD database

We researched the results as described in the flowchart (Supplementary Figure 1). This study analyzed three GEO datasets (GSE43458, GSE32863, and GSE27262), and 886, 1270, and 1921 DEGs were found, respectively. Then, we found 380 common DEGs by Venn diagram analysis. DEGs were verified in the TCGA-LUAD database (535 Lung adenocarcinoma tissues and 59 Lung cancer tissues), further confirming the differential expression of these 380 genes in lung adenocarcinoma and normal lung tissues (Figure 1A-D).

Figure 1 Screening of differential genes and establishing LASSO regression. A: Volcano map of the differential genes in the GSE43458; B: Volcano map of the differential genes in the GSE32863; C: Volcano map of the differential genes in the GSE27262; D: Venn diagram of the three Gene Expression Omnibus datasets; E: LASSO coefficients profiles of 380 common differential genes; F: LASSO regression with ten- fold cross-validation constructed the models.

Screening of prognostic differential genes and establishment of prognostic models

Univariate Cox regression analysis was performed on 380 DEGs in the training set. A total of 30 DEGs were related to the survival of patients with lung adenocarcinoma ($P < 0.05$) and further screened by LASSO regression (Figure 1E). Cross-validation was used to establish the model, as shown in Figure 1F. A total of 5 mRNAs (TCN1, CENPF, MAOB, CRTAC1, and PLEK2) were included in the model. Multivariate Cox regression analysis was performed for the above 5 mRNAs, and the risk scoring equation was established according to the corresponding regression coefficient. Risk score (RS) =

(0.00288∗ TCN1 EXP) + (0.0387∗ CENPF EXP) + (-0.0291∗ MAOB EXP) + (-0.0198 ∗CRTAC1 EXP) + (0.0214∗ PLEK2 EXP).

Verification of mRNA expression and genetic changes associated with 5 prognostic genes

Among the 566 patients included in the cBioPortal for Lung Adenocarcinoma (TCGA, PanCancer Atlas) database, 93 patients (16.4%) showed genetic changes in these 5 genes, among which missense mutations were the most common mutation type (Figure 2A). In the TCGA LUAD cohort, the mRNA expression levels of TCN1, CENPF, and PLEK2 were significantly increased in lung adenocarcinoma tissues compared with those in normal lung tissues, while MAOB and CRTAC1 were significantly decreased in lung adenocarcinoma tissues (Figure 2B). A human protein mapping database (http://www.proteinatlas.org) was used to explore the protein expression level. Immunohistochemical (IHC) results of four genes (TCN1 was not included in the database) in lung cancer and normal lung gland tissues are shown in Figure 2C. Consistent with the mRNA results, IHC results showed that CENPF and PLEK2 had significantly higher mean expression levels in lung adenocarcinoma tissue than in normal lung tissue. In contrast, the CRTAC1 expression level was higher in normal lung tissue than in lung adenocarcinoma tissue. MAOB showed no difference between normal and lung adenocarcinoma tissues (Figure 2C).

Evaluation of five-mRNA prognostic model

Each patient's risk score in the training group was calculated based on the above risk score function. The "SurvMiner" R software package was used to obtain the median critical point, and the patients were divided into a high-risk group and a low-risk group (Figure 3A). As the RS score increased, the patients' survival time was shortened, and the number of deaths increased significantly (Figure 3B). Figure 3C shows the heatmap of 5 prognostic genes in the high- and low-risk groups. The KM survival curve indicated a lower overall survival in the high-risk group than in the low-risk group ($P < 0.001$) (Figure 3D). To further verify the prognostic assessment model's accuracy, we used the R "survival ROC" package to draw the ROC curve (Figure 3E). The results showed that the AUC values of the risk score model for predicting the overall survival at 1, 3 and 5 years in patients with lung adenocarcinoma were 0.711, 0.668 and 0.728, respectively, indicating that the multigene model had a good predictive ability for the OS of patients with lung adenocarcinoma. Multiple Cox regression analysis showed that RS, along with patient age and stage, could be independent prognostic factors for lung adenocarcinoma patients (Figure 3F).

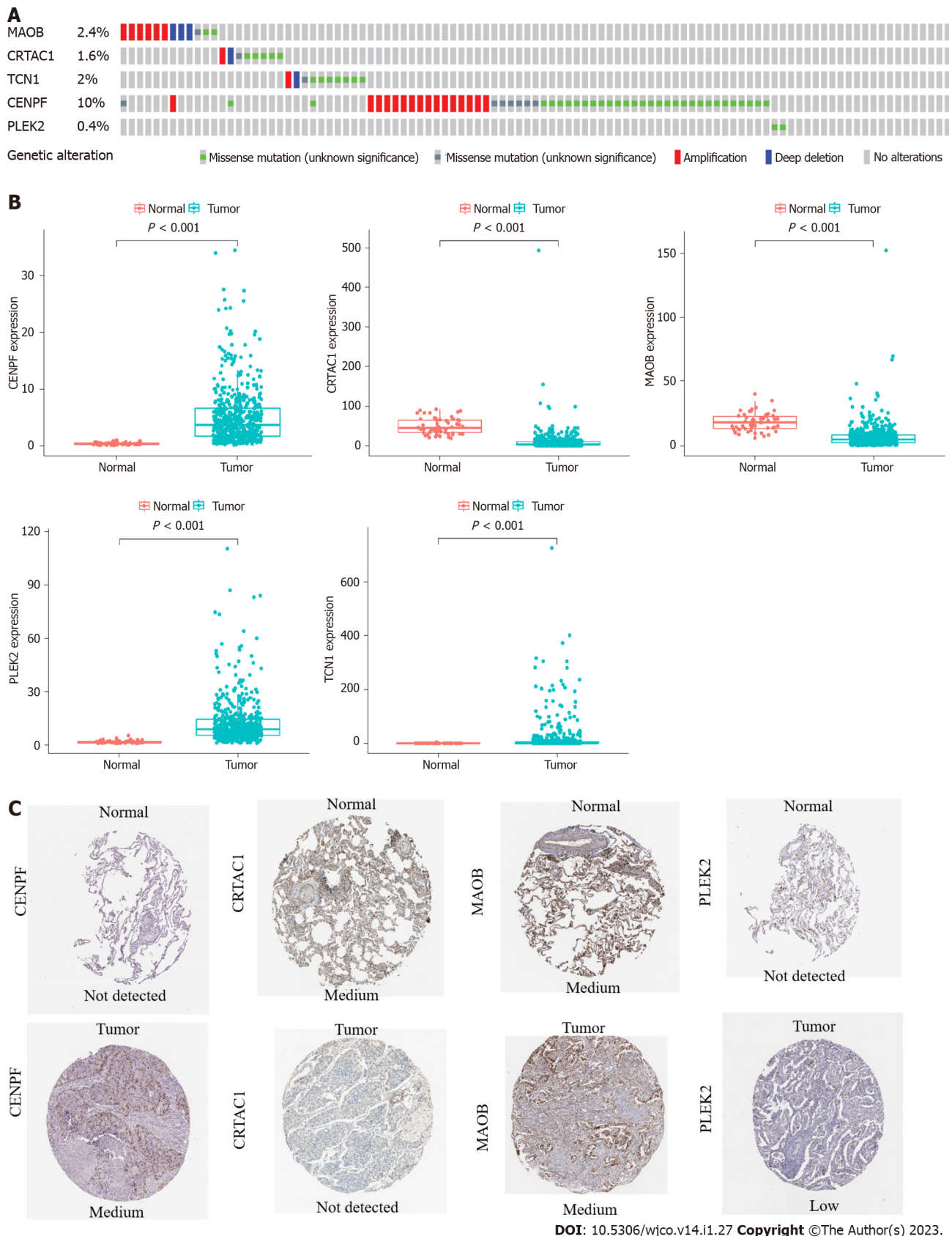### Internal and external validation of the five-mRNA prognostic signature

To verify the predictive value of the 5-mRNA prognostic signature, we used the same formula to calculate risk scores for patients with the internal validation set (n = 160), entire validation set (n = 535), and external validation set (GSE11969, n = 90). Consistent with the training group results, the OS of LUAD patients in the high-risk group was significantly lower than that in the low-risk group (Figure 4A -C). The KM survival analysis of the prognostic signature showed that the AUC values of the 1-year, 3-year, and 5-year OS of the internal validation set, the overall validation set and external validation set were 0.754, 0.630, 0.684, and 0.737, 0.701, 0.680, and 0.779, 0.752, 0.715, respectively (Figure 4D-F). Taken together, our results suggest that this 5-gene signature performs well in predicting overall survival in patients with lung adenocarcinoma.

Establish and verify the nomogram

To establish clinically applicable methods for predicting survival in patients with lung adenocarcinoma, we established a nomogram using three independent prognostic factors (including age, stage, and risk score) to predict 1-year, 3-year, and 5-year OS in patients with lung adenocarcinoma (Supplementary Figure 2A). The calibration diagram showed that the nomogram performs well (Supplementary Figure 2B). The AUC values of the 1-year, 2-year, and 3-year overall survival predictions of the nomograph were 0.760, 0.712, and 0.709, respectively (Supplementary Figure 3A). The KM chart effectively distinguishes the various risks of these categories, with people with higher scores having significantly poorer overall survival ($P < 0.001$) (Supplementary Figure 3B). The C-index (95%CI) of the age, stage, and risk score and combination models was 0.501 (0.480-0.522), 0.684 (0.662-0.076), 0.625 (0.604-0.646), and 0.726 (0.702-0.750), respectively (Supplementary Table 1). Thus, the nomogram performs well in predicting overall survival in patients with lung adenocarcinoma, which may be useful for patient counseling and clinical decision-making.

Biological pathways of the five prognostic model genes were identified
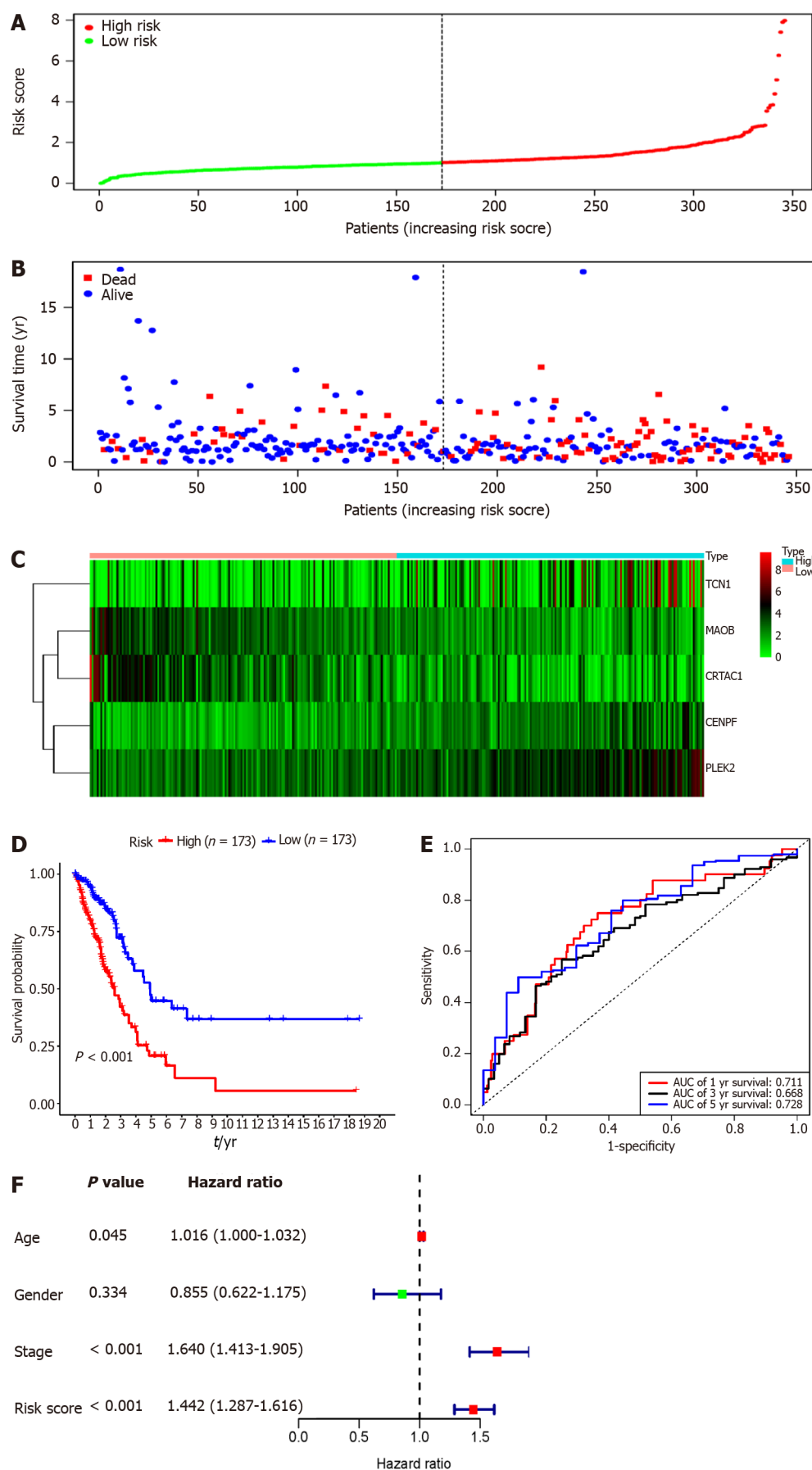
GSEA was performed to identify the potential biological processes of the 5 prognostic genes and showed that the samples with highly expressed TCN1, CENPF, and PLEK2 were enriched with focal adhesion, the p53 signaling pathway, and Toll-like sensors, respectively. MAOB and CRTAC1 samples were mediated in the cell cycle and ubiquitin-mediated proteolysis (Supplementary Figure 4), respectively.

Figure 2 The expression and genetic alterations of the 5 prognostic genes in Lung adenocarcinoma. A: Genetic alterations rate of 5 model genes; B: Differential expression of the mRNA levels in lung adenocarcinoma tissues; C: Differential expression at the protein levels of the five model genes.
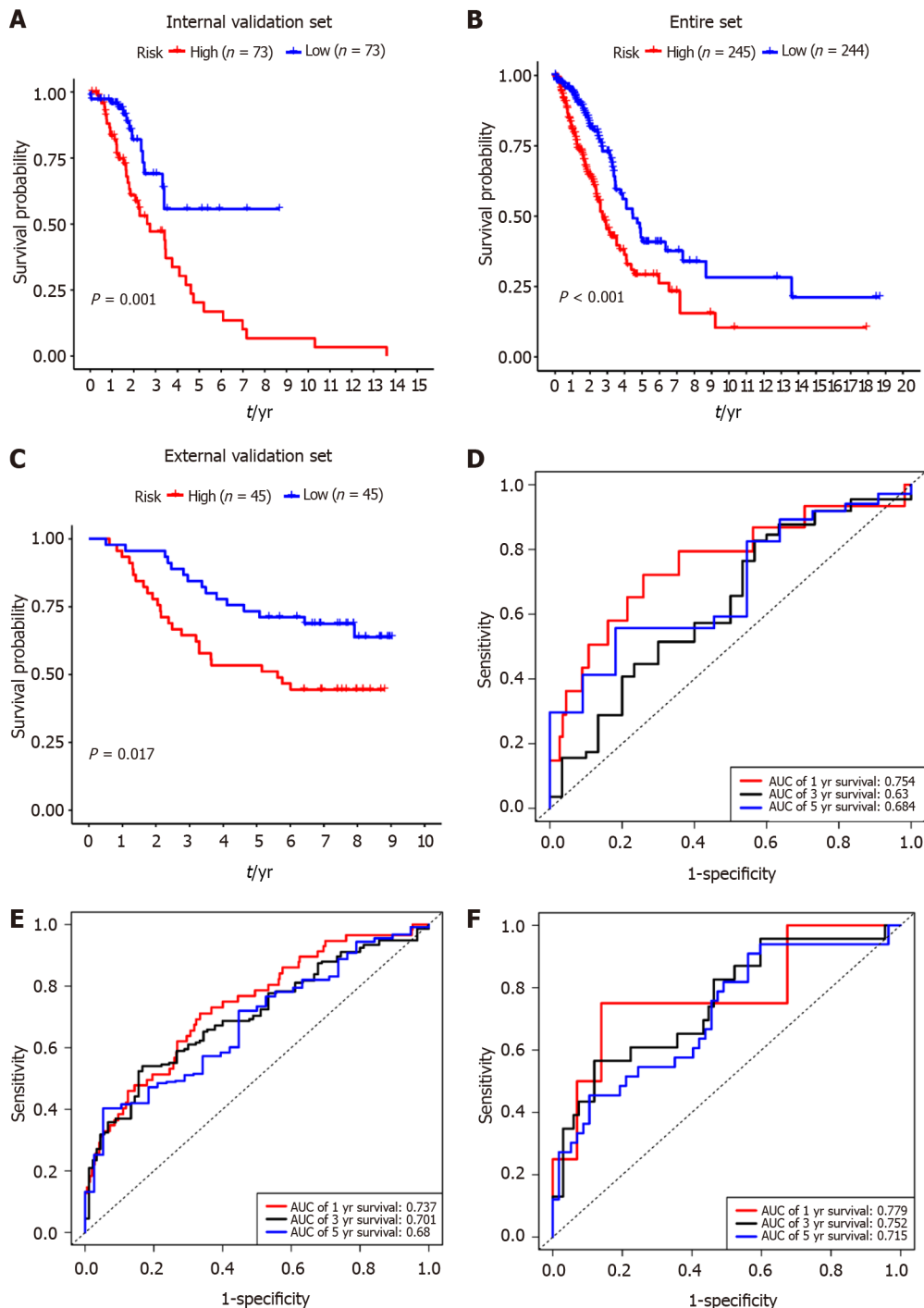
DISCUSSION

Recently, the tumor prognosis model based on the abnormal gene mRNAs has shown great potential due to its high prediction accuracy. Traditional clinicopathological parameters, such as tumor stage,

Figure 3 Risk score model, time-dependent receiver operating characteristic analysis, and survival analysis for the 5-gene signature in

Lung adenocarcinoma. A: The risk score curve divided the patients into the high-risk and low-risk groups; B: Distribution map of the patient's survival status; C: Heatmap of model genes in high and low risk groups; D: Kaplan-Meier survival analysis of the 5-mRNA-based prognostic signature; E: Receiver operating characteristic curves to evaluate the prognostic signature; F: Multiple cox regression analysis of the risk scores and clinical parameters.

DOI: 10.5306/wjco.v14.i1.27 **Copyright** ©The Author(s) 2023.

Figure 4 Internal and external validation of the 5-gene prognostic signature. A: Internal validation by survival analysis; B: The entire dataset validation by survival analysis; C: GSE11969-based external validation by survival analysis; D: Internal validation by receiver operating characteristic (ROC) curves; E: The entire dataset validation by ROC curves; F: GSE11969-based external validation by ROC curves.

have been used to reflect and predict disease progression. However, a single clinical parameter has poor predictive ability for prognosis[3,6,7,10]. In this study, we identified 380 reliable lung adenocarcinoma differential genes by comprehensive analysis of three GEO datasets combined with data from the TCGA-LUAD. Univariate, LASSO and multivariate Cox analyses of DEGs were performed to establish a

WJCO | https://www.wjgnet.com   35   January 24, 2023 | Volume 14 | Issue 1 |

prognostic risk model for lung adenocarcinoma based on 5 mRNAs (TCN1, CENPF, MAOB, CRTAC1, and PLEK2). These five new genes were significantly correlated with the prognosis of LUAD patients. MAOB and CRTAC1 were negative prognostic genes, while TCN1, CENPF and PLEK2 were positive prognostic genes. Recently, several studies have revealed the important role of these five genes in cancer progression. Monoamine oxidase B (MAOB) is an enzyme located on the outer mitochondrial membrane. It is responsible for catalyzing monoamine oxidation to produce hydrogen peroxide and is mainly involved in the metabolism of neurotransmitters[19]. The relationship between MAOB and tumors is less reported. It has been reported that MAOB mRNA is significantly lower in the saliva of oral squamous cell carcinoma patients than in that of healthy controls[20]. Xu *et al*[21] found that MAOB is a key DNA methylation driver gene for prostate cancer and plays an important role in the DNA methylation of prostate cancer patients through a comprehensive analysis of the TCGA methylation data. There is no previous report about MAOB in lung adenocarcinoma. CRTAC1 encodes human chondrogenic acid protein 1, which can be used as a marker of chondrocytes to distinguish human chondrocytes from osteoblasts and mesenchymal stem cells in cultures[22]. Currently, this gene is rarely reported in tumors. TCN1 is a member of the vitamin B12-binding protein family and is a 60-70 kDa molecular weight protein. High levels of TCN1 are primarily related to abnormal granulocyte proliferation. TCN1 is overexpressed in a variety of malignancies, such as pancreas, breast, and colon cancer, and is associated with tumor progression and metastasis[23-25]. TCN1 was significantly associated with advanced colorectal cancer[24] and laryngeal cancer[26]. Centromere protein F (CENPF), as an important member of the centromere protein family, is a component of the centromere complex and plays an important regulatory role in mitosis[27]. CENPF expression is abnormally increased in a variety of malignant tumors and is associated with the prognosis of patients[28,29]. Using bioinformatics and immunohistochemical analysis, CENPF overexpression was associated with poor prognosis of breast cancer and tumor bone metastasis[30]. Through comprehensive analysis of three GEO databases, CENPF was identified as a key gene with prognostic value in lung adenocarcinoma, which was consistent with our research results[31]. Pleckstrin-2 (PLEK2) is a 353 amino acid protein encoded by the PLEK2 gene in the human genome and is widely expressed in various tissues. Its overexpression contributes to the formation of large apolipoproteins, thereby promoting cell proliferation[32]. PLEK2 has been found to be related to the invasion and metastasis of multiple tumors. In gallbladder cancer (GBC), PLEK2 overexpression enhances the epithelial-mesenchymal transformation (EMT) process in GBC cells, leading to subsequent higher rates of cell migration, invasion, and liver metastasis[33]. The overexpression of PLEK2 also significantly promoted the EMT and migration of non-small cell lung cancer and destroyed the vascular endothelial barrier[34]. After identifying the five prognostic gene markers, we also conducted internal and external validation to confirm their predictive value and revealed that the prognostic signatures had good prognostic diagnostic value. To improve the prognostic predictive power of the five prognostic gene markers, a predictive nomogram combining risk scores and conventional clinical prognostic parameters (including age and tumor stage) was constructed to enable clinicians to determine the prognosis of each patient. Its graphical scoring system is easy to understand and helps customize treatment and medical decisions. The prognostic models and nomograms associated with five-gene characteristics have not been reported. Hence, our study may be useful prognostic and diagnostic classification tools for lung adenocarcinoma. Our study still has some limitations. First, the study only focuses on transcriptome sequencing data. If other omics techniques, such as DNA methylation and single nucleotide polymorphisms, can be analyzed together, more favorable results may be obtained. Second, our research is limited to the bioinformatics analysis of the TCGA and GEO databases. Although we have verified the accuracy of the models internally and externally, the verification of large samples in the clinical diagnosis and treatment process will further enhance their diagnostic accuracy and clinical value.

CONCLUSION

In summary, our study identified a 5-gene model and prognostic nomogram that combined gene models and clinical prognostic factors to predict the overall survival rate of lung adenocarcinoma patients, and this nomogram may be of great significance for the selection of personalized treatment options and clinical medical decisions in patients with lung adenocarcinoma.

ARTICLE HIGHLIGHTS

Research background
Lung adenocarcinoma patients with localized or locally advanced disease have a high risk of death, and their 5-year overall survival rate is less than 15%.

Research motivation

To evaluate the prognosis of Lung adenocarcinoma (LUAD) patients and optimize treatment, effective clinical research prediction models.

Research objectives

To identify reliable prognostic biomarkers to predict clinical outcomes and to help clinicians to make accurate clinical treatment decisions.

Research methods

The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) were used to screen for differential genes for lung adenocarcinoma. Univariate regression analysis combined with LASSO regression analysis was used to screen for prognostic-related genes. Multivariate Cox regression analysis was applied to establish the risk score equation and construct the survival prognosis model.

Research results

We establish a prognostic risk model for lung adenocarcinoma based on 5 mRNAs (TCN1, CENPF, MAOB, CRTAC1, and PLEK2). These five new genes were significantly correlated with the prognosis of LUAD patients. To improve the prognostic predictive power of the five prognostic gene markers, a predictive nomogram combining risk scores and conventional clinical prognostic parameters (including age and tumor stage) was constructed to enable clinicians to determine the prognosis of each patient.

Research conclusions

A 5-mRNA-based model was constructed to predict the prognosis of lung adenocarcinoma, which may provide clinicians with reliable prognostic assessment tools and help clinical treatment decisions.

Research perspectives

Our study identified a 5-gene model and constructed a nomogram which may have important implications for clinical medical decision and personalized treatment of patients with lung adenocarcinoma.

FOOTNOTES

REFERENCES

1   **Meng W**, Ye Z, Cui R, Perry J, Dedousi-Huebner V, Huebner A, Wang Y, Li B, Volinia S, Nakanishi H, Kim T, Suh SS, Ayers LW, Ross P, Croce CM, Chakravarti A, Jin VX, Lautenschlaeger T. MicroRNA-31 predicts the presence of lymph node metastases and survival in patients with lung adenocarcinoma. *Clin Cancer Res* 2013; **19**: 5423-5433 [PMID: 23946296 DOI: 10.1158/1078-0432.CCR-13-0320]

2   **Drosten M**, Barbacid M. Modeling K-Ras-driven lung adenocarcinoma in mice: preclinical validation of therapeutic targets. *J Mol Med (Berl)* 2016; **94**: 121-135 [PMID: 26526121 DOI: 10.1007/s00109-015-1360-5]

3   **Li S**, Xuan Y, Gao B, Sun X, Miao S, Lu T, Wang Y, Jiao W. Identification of an eight-gene prognostic signature for lung adenocarcinoma. *Cancer Manag Res* 2018; **10**: 3383-3392 [PMID: 30237740 DOI: 10.2147/CMAR.S173941]

4   **Zuo S**, Wei M, Zhang H, Chen A, Wu J, Wei J, Dong J. A robust six-gene prognostic signature for prediction of both disease-free and overall survival in non-small cell lung cancer. *J Transl Med* 2019; **17**: 152 [PMID: 31088477 DOI: 10.1186/s12967-019-1899-y]

5   **He R**, Zuo S. A Robust 8-Gene Prognostic Signature for Early-Stage Non-small Cell Lung Cancer. *Front Oncol* 2019; **9**: 693 [PMID: 31417870 DOI: 10.3389/fonc.2019.00693]

6   **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* 2017; **109** [PMID: 27707839 DOI: 10.1093/jnci/djw200]

7   **Yue C**, Ma H, Zhou Y. Identification of prognostic gene signature associated with microenvironment of lung adenocarcinoma. *PeerJ* 2019; **7**: e8128 [PMID: 31803536 DOI: 10.7717/peerj.8128]

8   **Huang P**, Cheng CL, Chang YH, Liu CH, Hsu YC, Chen JS, Chang GC, Ho BC, Su KY, Chen HY, Yu SL. Molecular gene signature and prognosis of non-small cell lung cancer. *Oncotarget* 2016; **7**: 51898-51907 [PMID: 27437769 DOI: 10.18632/oncotarget.10622]

9   **Shahid M**, Choi TG, Nguyen MN, Matondo A, Jo YH, Yoo JY, Nguyen NN, Yun HR, Kim J, Akter S, Kang I, Ha J, Maeng CH, Kim SY, Lee JS, Kim SS. An 8-gene signature for prediction of prognosis and chemoresponse in non-small cell lung cancer. *Oncotarget* 2016; **7**: 86561-86572 [PMID: 27863408 DOI: 10.18632/oncotarget.13357]

10   **Krzystanek M**, Moldvay J, Szüts D, Szallasi Z, Eklund AC. A robust prognostic gene expression signature for early stage lung adenocarcinoma. *Biomark Res* 2016; **4**: 4 [PMID: 26900477 DOI: 10.1186/s40364-016-0058-3]

11   **Zengin T**, Önal-Süzek T.   Analysis of genomic and transcriptomic variations as prognostic signature for lung adenocarcinoma. BMC bioinformatics 2020, 21(Suppl 14):368[DOI: 10.1186/s12859-020-03691-3].

12   **Li YY**, Yang C, Zhou P, Zhang S, Yao Y, Li D. Genome-scale analysis to identify prognostic markers and predict the survival of lung adenocarcinoma. *J Cell Biochem* 2018; **119**: 8909-8921 [PMID: 30105759 DOI: 10.1002/jcb.27144]

13   **Xie H**, Xie C. A Six-Gene Signature Predicts Survival of Adenocarcinoma Type of Non-Small-Cell Lung Cancer Patients: A Comprehensive Study Based on Integrated Analysis and Weighted Gene Coexpression Network. *Biomed Res Int* 2019; **2019**: 4250613 [PMID: 31886214 DOI: 10.1155/2019/4250613]

14   **Zhang MY**, Liu XX, Li H, Li R, Liu X, Qu YQ. Elevated mRNA Levels of AURKA, CDC20 and TPX2 are associated with poor prognosis of smoking related lung adenocarcinoma using bioinformatics analysis. *Int J Med Sci* 2018; **15**: 1676-1685 [PMID: 30588191 DOI: 10.7150/ijms.28728]

15   **Xia Q**, Li Z, Zheng J, Zhang X, Di Y, Ding J, Yu D, Yan L, Shen L, Yan D, Jia N, Chen W, Feng Y, Wang J. Identification of novel biomarkers for hepatocellular carcinoma using transcriptome analysis. *J Cell Physiol* 2019; **234**: 4851-4863 [PMID: 30272824 DOI: 10.1002/jcp.27283]

16   **Thul PJ**, Lindskog C. The human protein atlas: A spatial map of the human proteome. *Protein Sci* 2018; **27**: 233-244 [PMID: 28940711 DOI: 10.1002/pro.3307]

17   **Huang JT**, Wang J, Srivastava V, Sen S, Liu SM. MicroRNA Machinery Genes as Novel Biomarkers for Cancer. *Front Oncol* 2014; **4**: 113 [PMID: 24904827 DOI: 10.3389/fonc.2014.00113]

18   **Li Z**, Chen Z, Hu G, Zhang Y, Feng Y, Jiang Y, Wang J. Profiling and integrated analysis of differentially expressed circRNAs as novel biomarkers for breast cancer. *J Cell Physiol* 2020; **235**: 7945-7959 [PMID: 31943203 DOI: 10.1002/jcp.29449]

19   **Edmondson DE**. Hydrogen peroxide produced by mitochondrial monoamine oxidase catalysis: biological implications. *Curr Pharm Des* 2014; **20**: 155-160 [PMID: 23701542 DOI: 10.2174/13816128113190990406]

20   **Oh SY**, Kang SM, Kang SH, Lee HJ, Kwon TG, Kim JW, Lee ST, Choi SY, Hong SH. Potential Salivary mRNA Biomarkers for Early Detection of Oral Cancer. *J Clin Med* 2020; **9** [PMID: 31963366 DOI: 10.3390/jcm9010243]

21   **Xu N**, Wu YP, Ke ZB, Liang YC, Cai H, Su WT, Tao X, Chen SH, Zheng QS, Wei Y, Xue XY. Identification of key DNA methylation-driven genes in prostate adenocarcinoma: an integrative analysis of TCGA methylation data. *J Transl Med* 2019; **17**: 311 [PMID: 31533842 DOI: 10.1186/s12967-019-2065-2]

22   **Steck E**, Bräun J, Pelttari K, Kadel S, Kalbacher H, Richter W. Chondrocyte secreted CRTAC1: a glycosylated extracellular matrix molecule of human articular cartilage. *Matrix Biol* 2007; **26**: 30-41 [PMID: 17074475 DOI: 10.1016/j.matbio.2006.09.006]

23   **Chong LY**, Cheok PY, Tan WJ, Thike AA, Allen G, Ang MK, Ooi AS, Tan P, Teh BT, Tan PH. Keratin 15, transcobalamin I and homeobox gene Hox-B13 expression in breast phyllodes tumors: novel markers in biological classification. *Breast Cancer Res Treat* 2012; **132**: 143-151 [PMID: 21574054 DOI: 10.1007/s10549-011-1555-6]

24   **Chu CM**, Yao CT, Chang YT, Chou HL, Chou YC, Chen KH, Terng HJ, Huang CS, Lee CC, Su SL, Liu YC, Lin FG, Wetter T, Chang CW. Gene expression profiling of colorectal tumors and normal mucosa by microarrays meta-analysis using prediction analysis of microarray, artificial neural network, classification, and regression trees. *Dis Markers* 2014; **2014**: 634123 [PMID: 24959000 DOI: 10.1155/2014/634123]

25   **Wu Y**, Wei J, Ming Y, Chen Z, Yu J, Mao R, Chen H, Zhou G, Fan Y. Orchestrating a biomarker panel with lncRNAs and mRNAs for predicting survival in pancreatic ductal adenocarcinoma. *J Cell Biochem* 2018; **119**: 7696-7706 [PMID: 29923223 DOI: 10.1002/jcb.27119]

26   **Wang Y**, Yue C, Fang J, Gong L, Lian M, Wang R, Feng L, Ma H, Ma Z, Liu H. Transcobalamin I: a novel prognostic biomarker of neoadjuvant chemotherapy in locally advanced hypopharyngeal squamous cell cancers. *Onco Targets Ther* 2018; **11**: 4253-4261 [PMID: 30100732 DOI: 10.2147/OTT.S166514]

27   **Varis A**, Salmela AL, Kallio MJ. Cenp-F (mitosin) is more than a mitotic marker. *Chromosoma* 2006; **115**: 288-295 [PMID: 16565862 DOI: 10.1007/s00412-005-0046-0]

28   **Brendle A**, Brandt A, Johansson R, Enquist K, Hallmans G, Hemminki K, Lenner P, Försti A. Single nucleotide polymorphisms in chromosomal instability genes and risk and clinical outcome of breast cancer: a Swedish prospective case-control study. *Eur J Cancer* 2009; **45**: 435-442 [PMID: 19008095 DOI: 10.1016/j.ejca.2008.10.001]

29   **Chen WB**, Cheng XB, Ding W, Wang YJ, Chen D, Wang JH, Fei RS. Centromere protein F and survivin are associated with high risk and a poor prognosis in colorectal gastrointestinal stromal tumours. *J Clin Pathol* 2011; **64**: 751-755 [PMID: 21613637 DOI: 10.1136/jcp.2011.089631]

30   **Sun J**, Huang J, Lan J, Zhou K, Gao Y, Song Z, Deng Y, Liu L, Dong Y, Liu X. Overexpression of CENPF correlates with poor prognosis and tumor bone metastasis in breast cancer. *Cancer Cell Int* 2019; **19**: 264 [PMID: 31632198 DOI: 10.1186/s12935-019-0986-8]

31   **Song YJ**, Tan J, Gao XH, Wang LX. Integrated analysis reveals key genes with prognostic value in lung adenocarcinoma. *Cancer Manag Res* 2018; **10**: 6097-6108 [PMID: 30538558 DOI: 10.2147/CMAR.S168636]

32   **Hu MH**, Bauman EM, Roll RL, Yeilding N, Abrams CS. Pleckstrin 2, a widely expressed paralog of pleckstrin involved in actin rearrangement. *J Biol Chem* 1999; **274**: 21515-21518 [PMID: 10419454 DOI: 10.1074/jbc.274.31.21515]

33   **Luo Y**, Robinson S, Fujita J, Siconolfi L, Magidson J, Edwards CK, Wassmann K, Storm K, Norris DA, Bankaitis-Davis D, Robinson WA, Fujita M. Transcriptome profiling of whole blood cells identifies PLEK2 and C1QB in human melanoma. *PLoS One* 2011; **6**: e20971 [PMID: 21698244 DOI: 10.1371/journal.pone.0020971]

34   **Wu DM**, Deng SH, Zhou J, Han R, Liu T, Zhang T, Li J, Chen JP, Xu Y. PLEK2 mediates metastasis and vascular invasion *via* the ubiquitin-dependent degradation of SHIP2 in non-small cell lung cancer. *Int J Cancer* 2020; **146**: 2563-2575 [PMID: 31498891 DOI: 10.1002/ijc.32675]