Development and Validation of a 23-Gene Signature for Prognosis Prediction in Lung Adenocarcinoma

Jichang Liu
Institute of Oncology, Shandong Provincial Hospital, Cheeloo College of Medicine, Shandong university

Yadong Wang
Institute of Oncology, Shandong Provincial Hospital, Cheeloo College of Medicine, Shandong University.

Weiqing Zhong
Department of Radiology, The Third Affiliated Hospital of Shandong First Medical University (The Fourth People's Hospital of Jinan)

Yong Liu
Institute of Oncology, Shandong Provincial Hospital, Cheeloo College of Medicine, Shandong University.

Kai Wang
Institute of Oncology, Shandong Provincial Hospital, Cheeloo College of Medicine, Shandong University.

Lei Cong
Department of Oncology, Shandong Provincial Hospital Affiliated to Shandong First Medical University

Jiajun Du (✉ dujiajun@sdu.edu.cn)
Institute of Oncology, Shandong provincial Hospital, Cheeloo College of Medicine, Shandong University.
https://orcid.org/0000-0003-2406-9435

Primary research

Keywords: LUAD, EMT, immune, stromal, nomogram

DOI: https://doi.org/10.21203/rs.3.rs-272282/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background: Lung cancer remains the most fatal tumorous disease in the worldwide. Among that, lung adenocarcinoma (LUAD) was the most common histological type. A precise and concise prognostic model was urgently needed of LUAD. We developed a 23-gene signature for prognosis prediction based on EMT, immune and stromal datasets.

Methods: Univariate Cox regression analysis was performed to select genes which were significantly associated with overall survival (OS) of the TCGA LUAD cohorts. LASSO regression and multivariate Cox regression analysis was used to build the multi-gene signature. Enrichment analyses and a protein-protein interactions (PPI) network were performed to show the interaction and functions of the signature. A nomogram was developed based on risk score and other clinical features. Predictive performance of the signature was externally validated in two independent datasets from Gene Expression Omnibus (GSE37745 and GSE13213).

Results: A total of 1334 EMT, immune and stromal associated genes were obtained. After LASSO regression and multivariate Cox regression analysis, a 23-gene signature for risk stratification was built. K-M curves showed that the patients with high risk had a poorer outcome. Finally, a nomogram was built to predict prognosis. The predictive performance of the 23-gene signature was confirmed in internal and external validation.

Conclusion: We developed and verified a 23-gene signature based on EMT, immune and stromal gene sets. It provided a convenient and concise tool for risk stratification and individual medicine.

1. Introduction

Lung cancer remains the most commonly diagnosed cancer and the leading cause of oncological death, accounting for 18.4% of total cancer-related deaths worldwide, which makes it the most fatal tumorous disease(1). Non-small cell lung cancer (NSCLC) accounts of approximately 85% cases of lung cancer, among which, lung adenocarcinoma (LUAD) is the most common histological type, accounting for at least a half of all NSCLC cases(2, 3).

Despite continued development in treatment modalities, including targeted therapy and immunotherapy, the prognosis of patients has improved significantly but the 5-year survival rate of patients remains < 20% (4, 5). To date, histopathologic diagnosis and tumor staging system remain the main approaches of the prognostic prediction(6). However, these traditional approaches are not sufficient for assessing the prognosis of LUAD patients. Therefore, identification and validation of a series of specific prognostic biomarkers for screening and evaluating patients are imperative. In recent years, with the development of tumor biology, a series of biological behaviors and phenomena attracted our attention, including epithelial-mesenchymal transition (EMT) and tumor microenvironment (TME).
Epithelial–mesenchymal transition (EMT) is a complicated process involved in tumor metastasis\(^7\). During activation of the epithelial-mesenchymal transition (EMT) program, the cohesive and polarized tumor epithelial cells transform into a motile mesenchymal state with lacking the adhesion character of junction between epithelial cells\(^8\). EMT process results in the increase of tumor metastasis as well as multidrug resistance (MDR), which affect prognosis of patients\(^9,\,10\). However, EMT is a complex process involving genetic modification and immune alteration\(^8\). Recently, a series of studies indicate that the tumor microenvironment (TME) play an important role in EMT, metastasis and MDR\(^6\). Tumor microenvironment (TME) is a complex network, which comprises of a variety of cell type, including the epithelial cells, immune cells, stromal cells, cancer-associated fibroblasts (CAFs) and extracellular matrix (ECM)\(^11\). During TME, the immune cells and TME-induced factors are involved in EMT phenotype. TME-induced factors secreted by cancer cells and CAFs can directly induce the expression of EMT transcription factors or indirectly by activating other signaling pathways, such as WNT/\(\beta\) catenin signaling pathway, TNF-\(\alpha\) signaling pathway\(^6\). In lung cancer, a recent study demonstrates that the crosstalk between Th9/Th17 lymphocytes and cancer cells promote the EMT and metastasis\(^12\). The TME and EMT interact to promote MDR, metastasis and tumor progression together\(^6,\,12\). These findings indicate that EMT and TME-related genes may hold great promise as lung cancer prognostic markers.

Due to the complexity of EMT and TME related genes, traditional single-gene markers are often not accurate enough for prognosis prediction. We have carried out systematic discovery and verification of biomarkers, and constructed a 23-gene signature using The Cancer Genome of Atlas (TCGA) datasets. The signature was based on EMT, immune and stromal (EIS) gene datasets, and external validation was performed in two independent LUAD datasets from Gene Expression Omnibus (GEO). The EIS signature may offer an accurate prediction and risk-stratification in LUAD patients’ prognosis and play an important role in the clinical management.

2. Materials And Methods

2.1 Study population

In the study, the RNA-sequencing (RNA-seq) datasets and corresponding clinical characteristics of 59 normal and 535 LUAD specimens were both downloaded from the TCGA website (https://portal.gdc.cancer.gov/). And the data was normalized and processed using R version 4.0.0 software. 500 patients with intact clinical characteristics were categorized as the training set, including gender, age, clinical stage, follow-up time and survival status.

The GSE37745 and GSE13213 datasets were obtained from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), which included 100 and 130 clinical and microarray-based gene expression data of LUAD specimens, respectively. They were categorized as the validation sets.

2.2. Enrichment analysis and PPI (protein-protein interactions) network construction
The top 300 differentially expressed genes between low risk and high risk groups were selected. After combining 23 genes of signature, Gene ontology (GO) analysis were performed in the 323 genes using R version 4.0.0 software.

We generated a PPI network of the 323 genes using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING, https://string-db.org, version 11.0). The minimum required interaction score was set to 0.9.

Next, we performed Gene Set Enrichment Analysis (GSEA) in differentially expressed genes between low risk and high risk groups (http://software.broadinstitute.org/gsea/).

2.3 ESTIMATE algorithm

As an algorithmic tool, ESTIMATE algorithm (Estimation of STromal and Immune cells in MAalignant Tumour tissues using Expression data) was performed to evaluated scores and infiltrating of stromal and immune cells between low and high risk groups. Besides, the tumor purity was calculated based on these. The R package “estimate” was applied in R version 4.0.0 software.

2.4. CIBERSORT estimation

Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts (CIBERSORT) was applied to assess the abundance ratio matrix of 22 kinds of immune cells. The differential abundance of immune cells infiltrates was obtained in cohorts with low and high risks. A violin illustration was plotted to visualize the results.

2.5. Statistical analysis

1184 EMT associated genes were identified from dbEMT 2.0 website(http://dbemt.bioinfo-minzhao.org/index.html), 102 immune(13) and 48 stromal(14) associated genes were selected from published datasets. R package “Edger” was utilized to identify differentially expressed genes of the 1334 genes with |the log(foldchange)| > 1.0 and FDR (adjusted p value) < 0.05. By univariate Cox proportional hazard regression analysis of selected genes, we attained the genes significantly related to overall survival (OS) of TCGA-LUAD patients. After performing the least absolute shrinkage and selection operator (LASSO) regression analysis of these OS associated genes, an EMT, immune and stromal associated multi-gene signature was developed. Kaplan-Meier curves were plotted to determine survival outcomes and the log-rank test was used to evaluate the statistical significance. We then calculated the risk score for each patient to predict OS. The cutoff value was taken as the median of the risk score, and the TCGA-LUAD cohort was dichotomized based on it. The Cox proportional hazards regression model was used to perform univariate and multivariate analysis to access the association between risk score and OS. A nomogram was constructed with age, stage and risk score, using the R library “survival” and “rms” package. Following that, the validation of the nomogram-based prediction model was accessed by calibration curves and Receiver Operating Characteristic (ROC) curves. Calibration was used to access the concordance between predicted and actual survival. ROC curves were used to evaluate the sensitivity and
specificity of survival prediction based on the nomogram. All statistics were performed by R software (Version 4.0.0; https://www.R-project.org). In all analyses, a P value of <0.05 was considered statistically significant.

3. Results

3.1. A 23-genes signature was identified for LUAD patients’ risk stratification

We selected 1184 EMT associated genes from dbEMT 2.0 website(http://dbemt.bioinfo-minzhao.org/index.html) and a total of 150 immune and stromal related genes from published datasets to generate multigene signature. (Supplementary Table 1) A total of 1334 genes were subjected to differential expressed genes selection and univariate Cox regression analysis. After that, a total of 121 genes were identified which were significantly related to the OS of TCGA-LUAD patients. Performing LASSO regression analysis in the above 121 genes. (Figure 1 A and B). 46 genes were concluded in multivariate Cox regression analysis. After Akaike Information Criterion (AIC) based gene screening in multivariate Cox regression analysis, a 23-genes signature significantly associated to LUAD patients’ OS was identified, including KL, ECT2, FBLN5, TIMELESS, TERT, TYMS, FHL2, MNDA, TIMP1, IL11, CDH19, AOC4P, ETV1, KRT18, SATB2, EFNB2, TNFRSF11A, MEOX2, MEG3, ANGPTL4, WNT1, L1CAM, FURIN. Through the visualization of the results, a forest plot was developed (Figure 1 C).

The TCGA-LUAD patients were divided into two groups based on the expression level of the 23 genes, respectively. And the K-M curves were plotted to show the survival difference (Supplementary Figure 1). A nomogram based on the expression of these 23 genes was plotted (Supplementary Figure 2).

A heatmap was plotted to show the mRNA expression. The expression values were significantly different between normal and tumor specimens (Figure 2A and 2B). Following that we examined the genetic alteration of these 23 genes in the cBioPortal database (http://www.cbioportal.org). A total of 507 LUAD patients/samples with mutation and CNA data were concluded, 23 genes were altered in 272 (54%) of queried patients/samples. And the main alteration was amplification (Figure 2C). Frequent genetic alterations mean that these genes play an important role in the development of LUAD.

The risk score of each patient was calculated based on the Cox regression coefficients and the mRNA expression value. Set the median of the risk score as cutoff value and dichotomize the cohorts into high and low risk groups. A heatmap was plotted to show the gene expression difference between high and low risk groups (Figure 3A). The distribution of risk score and survival time from top to bottom was performed in Figure 3B and 3C, which showed the relationship between risk score and survival time in our LUAD cohort. The K-M survival curve showed that the patients with high risk had a poorer outcome (Figure 3D). ROC curve analysis was performed to evaluate the forecasting property of the 23-gene signature. The AUC values of 1, 3 and 5 years were 0.8, 0.783 and 0.7538, respectively (Figure 3E).

3.2. Establishment of a risk associated nomogram for prognostic prediction
We performed univariate and multivariate Cox regression analysis based on the risk score and various clinicopathological characteristics. After univariate analysis, statistically significant features were identified which included age (P=0.049), risk score (P<0.001), stage (P<0.001). Multivariate regression analysis was performed on the above features. A forest plot was built, revealing that the risk score served as an independent predictor of prognosis (Figure 4A).

Nomogram is considered a reliable tool for individualized risk assessment, by integrating multiple risk factors (15-17). In order to improve the accuracy of our gene signature prediction, by integrating the significant factors in univariate analysis (age, stage, risk score), we established a risk-nomogram to predict the 1-, 3- and 5-year OS of LUAD patients (Figure 4B). Each factor assigned points proportionally according to the impact of its risk on survival. A higher total score, based on the sum of the assigned score of each factor, is associated with a poorer prognosis. The internal validation of the model was performed by ROC curves and calibration plots in 1-, 3- and 5-years (Figure 4C and 4D). The AUC value of 1-, 3- and 5 years was 0.8, 0.784 and 0.73988 respectively, which showed the high predictive accuracy of the model.

3.3. Enrichment analysis and PPI (protein-protein interactions) network

GO analysis indicated that 23-gene signature and the top 300 differentially expressed genes was mainly related to DNA molecular structure and expression regulation. (Figure 5A). The PPI network showed the interaction between the 323 proteins. (Figure 5B) The minimum required interaction score was set 0.9 and the free molecules in the network were removed. GSEA analysis indicated that differentially expressed genes were significantly enriched in several essential pathways, including cell cycle, DNA replication, adherens junction, T cell receptor and B cell receptor signaling pathways. (Figure 5C)

3.4. Correlation between immune microenvironment, tumor staging and risk score.

CIBERSORT algorithm was applied to analyze the differential composition of tumor-infiltrating immune cells in cohorts with low and high risk. (Figure 6A) Several immune cells were significantly associated to risk score, including T cells CD4 memory resting, monocytes, macrophages M0, dendritic cells resting, mast cells resting and mast cells activated. ESTIMATE algorithm was performed to calculated immune score, stromal score and tumor purity between cohorts with low and high risk, which showed significant correlation. (Figure 6B and 6C). Besides, risk score was also significantly related to tumor staging. (Figure 6D)

3.5. Validation of the 23-gene signature in GEO LUAD cohorts.

We next evaluated the predictive property of the 23-gene signature in several independent LUAD cohorts (GSE13213 and GSE 37745) from GEO database. In each validation, we calculated risk score of each patients and divided patients into high and low risk groups based on the median of risk score. K-M curve was plotted to demonstrate that patients with a high risk had a worse outcome (Figure 7A). The patients
were ranked according to the risk and survival time (Figure 7B and 7C). The accuracy of prediction was evaluated by ROC curves and calibration plot of 1-, 3- and 5-years (Figure 7D and 7E).

**Discussion**

With the advancement of diagnosis and treatment methods, the death rate of lung cancer continues to decrease significantly, lung cancer remains the most fatal neoplastic disease all over the world, nowadays(18). Among which, LUAD is the most common histological type and treatment of LUAD has made important progress in chemotherapy, targeted therapy and other treatment methods. There are still a series of problems, such as drug resistance and identification of novel biomarkers(19, 20). Therefore, the development of biomarkers is particularly important in the diagnosis and treatment of LUAD.

EMT is a complex cellular process, during which the epithelial features decrease and epithelial cells obtain mesenchymal phenotype(21). At the cellular level, activation of EMT leads to the loss of cell polarity, the interruption of cell-cell junctions, the degradation of basement membrane and the reorganization of extracellular matrix (ECM)(22). At the molecular level, activation of EMT repress the expression of epithelial cadherin (E-cadherin) and certain cytokeratins and active the expression of mesenchymal markers, such as neural cadherin (N-cadherin), vimentin, fibronectin and β1 and β3 integrins(23). The activation of EMT and the above process are orchestrated by EMT-inducing transcriptional factors (EMT-TFs). EMT is closely related to tumor malignant progression, which includes tumor-initiating properties, motility and resistance to both chemotherapy and immunotherapy(24–26). These brought EMT to our attention.

However, the activation of EMT rarely as a spontaneous process in carcinoma cells. It is reported that signals from the tumor microenvironment (TME) act on cancer cells to induce the expression of EMT-TF, thereby coordinating the expression of various components in the EMT-related program(22). A large number of studies verified that the activation of EMT were closely associated to TME-associated cells, including immune cells, stromal cells, cancer-associated fibroblasts (CAFs)(11, 22). CAFs coordinately induce the activation of EMT in carcinoma cells by secreting a series of cytokines and growth factors, such as TGFβ, IL-6, EGF, VEGF and HGF(27–30). The activated effector T cells facilitate the EMT activation by releasing cytokines such as IL-6, TNF, TGFβ and other unknown mechanisms(31). Besides, TAMs secrete an array of factors to induce EMT(32–35). However, the effect of TME on EMT is not unilateral. Acting reciprocally, the tumor cells with EMT-related markers can regulate the activity of cellular components in the tumor microenvironment(36). Under the influence of EMT-TFs, the cellular components with antitumor functions, such as cytotoxic CD8+ T cells and etc. are overwhelmed by immunosuppressive cells (37–39). In addition to changes in cell composition, immunomodulatory markers on cell surface were also altered which induce immune evasion (40, 41). In summary, through interaction, EMT and TME influence each other and play an important role in the malignant progression of tumors.
A series of studies indicate the potential of EMT markers as tumoral high-grade malignancy markers in NSCLC and other malignant tumors (42, 43). However, due to the complexity, continuity and interaction with other phenotypes, a recent review indicated that one or a small number of markers maybe hard to assess EMT status. A combination of markers is more reliable (21). Therefore, the junction of EMT and TME markers is necessary for clinical oncology.

In our study, we selected 1184 EMT associated genes from dbEMT 2.0 website (http://dbemt.bioinfo-minzhao.org/), 102 immune associated genes and 48 stromal associated genes from published datasets to build our signature. We performed univariate Cox proportional hazard regression analysis on the 1334 genes and selected genes significantly associated to overall survival (OS) of TCGA-LUAD patients. After lasso regression analysis and multivariate Cox regression analysis, 23 genes were selected to build our EMT, immune and stromal associated genes, including KL, ECT2, FBLN5, TIMELESS, TERT, TYMS, FHL2, MNDA, TIMP1, IL11, CDH19, AO4P, ETV1, KRT18, SATB2, EFNB2, TNFRSF11A, MEOX2, MEG3, ANGPTL4, WNT1, L1CAM, FURIN. The risk score was calculated of each patient to predict OS of LUAD patients. The K-M curves showed that the patients with high risk had a poorer outcome. More importantly prognostic power of the 23-gene signature then was validated in GEO cohorts. Gene signature is a powerful classifier and has been widely used to predict prognosis of patients (44). In published research, different EMT-associated signatures were built in head and neck squamous cell carcinoma and pan-cancer, which were highly relevant to TME markers, respectively (45, 46). However, as far as we know, there is no signature based on EMT and TME markers.

Bioinformatic analysis demonstrated that the 23-gene signature mainly related to cell cycle and cell division, ubiquitin-mediated proteolysis and cell adhesion molecules. Interestingly, a large number of genes from 23-EIS signature are related to cell cycle, division and DNA replication. For example, Epithelial cell transforming 2 (ECT2), an activator of Rho family GTPases, was involved in a variety of cell process, including adhesion, migration, cell division, growth, and centromere maintenance (47). As a oncoprotein, ECT2 is reported to play an important role in the development of multiple cancers including LUAD (48). Moreover, a series of genes, such as TIMELESS (49), TERT (50) and FHL2 (51) were also involved in cell cycle process, more than EMT, immune and stromal.

In addition, several key molecules of the EIS signature were also involved in classic signal pathways. EFNB2, cell surface transmembrane ligand for Eph receptors, is involved in Eph receptors and ephrins signaling, which affects tumor growth, invasion, metastasis and angiogenesis (52). Moreover, TIMP1 (metalloproteinase inhibitor 1) binds to matrix metalloproteinases (MMPs) and inactivates them, which is closely associated to cell motility, invasion and angiogenesis. TIMP1 also acts as a growth factor to regulate cell differentiation, migration and cell death directly (53).

Importantly, we established a nomogram to predict the 1-, 3- and 5-year OS of LUAD patients. By integrating multiple risk factors, nomogram is considered as a reliable tool for risk prediction (54). As a predictive model, each factor associated with OS of patients was included into the nomogram, such as age, gender, and TNM stage. Besides, risk score based on gene signature can also incorporated as a
predictive factor. According to the impact of its risk on survival, each factor was assigned points proportionally. In our study, after univariate analysis, the significant clinical factors (age, stage) and risk score was included into the nomogram. Due to the combination of the multi-gene signature, the predictive performance of the nomogram maybe more accurate than traditional nomogram with only clinical features. Since EMT is related to tumor metastasis and drug resistance, in addition to OS, this signature may have a potential predictive effect on metastasis possibility and medication nonadherence. Moreover, the detection method is simple and feasible, which is suitable to validate our gene-signature in biopsies.

However, certain limitations existed in our study. For example, the clinical features of the public cohort were not complete enough, such as neoadjuvant treatment information and smoking status of patients. Although the 23-gene signature can predict OS of patients accurately based on gene expression level, a deeper research of gene, such as gene mutation, was not included in our research. And also, the possibility of the genes being potential therapeutic targets was not clarified.

In conclusion, we developed and verified a 23-gene signature of EMT, immune and stromal based on TCGA and GEO datasets. These genes are all of great importance and expected to be new targets. Besides, a nomogram based on these genes and clinical features was developed, which could accurately predict a 1-, 3- and 5-year survival time of LUAD patients. More importantly, these findings provide new ideas for systematically evaluation of prognosis of LUAD patients and effective personalized treatment.

**Abbreviations**

NSCLC, non-small cell lung cancer; LUAD, lung adenocarcinoma; EMT, epithelial–mesenchymal transition; TME, tumor microenvironment; OS, overall Survival; MDR, multidrug resistance; CAFs, cancer-associated fibroblasts; ECM, extracellular matrix; TAMs, tumor associated macrophages; EIS, EMT, immune and stromal; EMT, immune and stromal; TCGA, The Cancer Genome of Atlas; GEO, Gene Expression Omnibus; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene ontology; ROC, Receiver Operating Characteristic; AIC, Akaike Information Criterion; AUC: area under the curve; LASSO, least absolute shrinkage and selection operator; EMT-TFs, EMT-inducing transcriptions factors; MMPs, matrix metalloproteinases.

**Declarations**

6. ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was based on publicly available data from the TCGA database, and did not involve interaction with human subjects or the use of personally identifiable information. The study did not require informed consent for TCGA registration cases, and the author obtained a "limited use data agreement" from TCGA. No trial registration was required.

7. CONSENT TO PUBLISH
8. AVAILABILITY OF DATA AND MATERIALS

The dataset supporting the conclusions of this article is available in the TCGA website (https://portal.gdc.cancer.gov/) and additional file.

9. COMPETING INTERESTS

The authors declare that they have no competing interests.

10. FUNDING

This work is supported by National Natural Science Foundation of China (Grant No. 81672288) and Key Research and Development Plan of Shandong Province (Grant No. 2018GSF118028).

11. AUTHOR'S CONTRIBUTIONS

Jiajun Du and Lei Cong contributed to the designing and supervising the study, and correspondence. Co-authors Jichang Liu and Yadong Wang collected and analyzed the data and completed the manuscript. Weiqing Zhong, Kai Wang, Yong Liu helped to search for references and offered guidelines of statistical methods. All authors have read and approved the final version to be published.

12. ACKNOWLEDGEMENT

Not applicable.

13. AUTHOR'S INFORMATION

1Institute of Oncology, Shandong Provincial Hospital·Cheeloo College of Medicine, Shandong University; Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, PR China.

2Department of Thoracic Surgery, Shandong Provincial Hospital, Cheeloo College of Medicine, Shandong University; Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, PR China.

4Department of Oncology, Shandong Provincial Hospital affiliated to Shandong First Medical University, Jinan, PR China.

Highlights:

EMT, immune and stromal associated genes significantly related to OS of TCGA-LUAD patients were selected.

A 23-gene signature was developed and served as an independent risk factor.

A nomogram was developed for prognosis prediction based on the signature and clinical features.
References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.

2. Kumarakulasinghe NB, van Zanwijk N, Soo RA. Molecular targeted therapy in the treatment of advanced stage non-small cell lung cancer (NSCLC). Respirology. 2015;20(3):370–8.

3. Huang C, Qu X, Du J. Proportion of lung adenocarcinoma in female never-smokers has increased dramatically over the past 28 years. J Thorac Dis. 2019;11(7):2685–8.

4. Bittner N, Ostoros G, Geczi L. New treatment options for lung adenocarcinoma—in view of molecular background. Pathol Oncol Res. 2014;20(1):11–25.

5. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69(1):7–34.

6. Erin N, Grahovac J, Brozovic A, Efferth T. Tumor microenvironment and epithelial mesenchymal transition as targets to overcome tumor multidrug resistance. Drug Resist Updat. 2020;53:100715.

7. Datar I, Schalper KA. Epithelial-Mesenchymal Transition and Immune Evasion during Lung Cancer Progression: The Chicken or the Egg? Clin Cancer Res. 2016;22(14):3422–4.

8. Pietila M, Ivaska J, Mani SA. Whom to blame for metastasis, the epithelial-mesenchymal transition or the tumor microenvironment? Cancer Lett. 2016;380(1):359–68.

9. Zheng H, Kang Y. Multilayer control of the EMT master regulators. Oncogene. 2014;33(14):1755–63.

10. Assaraf YG. The role of multidrug resistance efflux transporters in antifolate resistance and folate homeostasis. Drug Resist Updat. 2006;9(4–5):227–46.

11. Yue C, Ma H, Zhou Y. Identification of prognostic gene signature associated with microenvironment of lung adenocarcinoma. PeerJ. 2019;7:e8128.

12. Salazar Y, Zheng X, Brunn D, Raifer H, Picard F, Zhang Y, et al. Microenvironmental Th9 and Th17 lymphocytes induce metastatic spreading in lung cancer. J Clin Invest. 2020;130(7):3560–75.

13. Rooney M, Shukla S, Wu C, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160:48–61.

14. Moffitt R, Marayati R, Flate E, Volmar K, Loeza S, Hoadley K, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet. 2015;47(10):1168–78.

15. Valentini V, van Stiphout RG, Lammering G, Gambacorta MA, Barba MC, Bebenek M, et al. Nomograms for predicting local recurrence, distant metastases, and overall survival for patients with locally advanced rectal cancer on the basis of European randomized clinical trials. J Clin Oncol. 2011;29(23):3163–72.

16. Han DS, Suh YS, Kong SH, Lee HJ, Choi Y, Aikou S, et al. Nomogram predicting long-term survival after d2 gastrectomy for gastric cancer. J Clin Oncol. 2012;30(31):3834–40.
17. Karakiewicz PI, Briganti A, Chun FK, Trinh QD, Perrotte P, Ficarra V, et al. Multi-institutional validation of a new renal cancer-specific survival nomogram. J Clin Oncol. 2007;25(11):1316–22.

18. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70(1):7–30.

19. Boumahdi S, de Sauvage FJ. The great escape: tumour cell plasticity in resistance to targeted therapy. Nat Rev Drug Discov. 2020;19(1):39–56.

20. Rosenzweig K, Gomez J. Concurrent Chemotherapy and Radiation Therapy for Inoperable Locally Advanced Non-Small-Cell Lung Cancer. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2017;35(1):6–10.

21. Yang J, Antin P, Berx G, Blanpain C, Brabletz T, Bronner M, et al. Guidelines and definitions for research on epithelial-mesenchymal transition. Nat Rev Mol Cell Biol. 2020;21(6):341–52.

22. Dongre A, Weinberg RA. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. Nat Rev Mol Cell Biol. 2019;20(2):69–84.

23. Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial-mesenchymal transition. Nat Rev Mol Cell Biol. 2014;15(3):178–96.

24. Mani S, Guo W, Liao M, Eaton E, Ayyanan A, Zhou A, et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell. 2008;133(4):704–15.

25. Singh A, Settleman J. EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. Oncogene. 2010;29(34):4741–51.

26. Shibue T, Weinberg R. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. Nature reviews Clinical oncology. 2017;14(10):611–29.

27. Pistore C, Giannoni E, Colangelo T, Rizzo F, Magnani E, Muccillo L, et al. DNA methylation variations are required for epithelial-to-mesenchymal transition induced by cancer-associated fibroblasts in prostate cancer cells. Oncogene. 2017;36(40):5551–66.

28. Shintani Y, Fujiwara A, Kimura T, Kawamura T, Funaki S, Minami M, et al. IL-6 Secreted from Cancer-Associated Fibroblasts Mediates Chemoresistance in NSCLC by Increasing Epithelial-Mesenchymal Transition Signaling. Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer. 2016;11(9):1482–92.

29. Yu Y, Xiao C, Tan L, Wang Q, Li X, Feng Y. Cancer-associated fibroblasts induce epithelial-mesenchymal transition of breast cancer cells through paracrine TGF-β signalling. British journal of cancer. 2014;110(3):724–32.

30. Zhao L, Ji G, Le X, Luo Z, Wang C, Feng M, et al. An integrated analysis identifies STAT4 as a key regulator of ovarian cancer metastasis. Oncogene. 2017;36(24):3384–96.

31. Chen Q, Yang D, Zong H, Zhu L, Wang L, Wang X, et al. Growth-induced stress enhances epithelial-mesenchymal transition induced by IL-6 in clear cell renal cell carcinoma via the Akt/GSK-3β/β-catenin signaling pathway. Oncogenesis. 2017;6(8):e375.

32. Bonde A, Tischler V, Kumar S, Soltermann A, Schwendener R. Intratumoral macrophages contribute to epithelial-mesenchymal transition in solid tumors. BMC Cancer. 2012;12:35.
33. Fan Q, Jing Y, Yu G, Kou X, Ye F, Gao L, et al. Tumor-associated macrophages promote cancer stem cell-like properties via transforming growth factor-beta1-induced epithelial-mesenchymal transition in hepatocellular carcinoma. Cancer letters. 2014;352(2):160–8.

34. Bates R, Mercurio A. Tumor necrosis factor-alpha stimulates the epithelial-to-mesenchymal transition of human colonic organoids. Molecular biology of the cell. 2003;14(5):1790–800.

35. Che D, Zhang S, Jing Z, Shang L, Jin S, Liu F, et al. Macrophages induce EMT to promote invasion of lung cancer cells through the IL-6-mediated COX-2/PGE/β-catenin signalling pathway. Molecular immunology. 2017;90:197–210.

36. Terry S, Savagner P, Ortiz-Cuaran S, Mahjoubi L, Saintigny P, Thiery J, et al. New insights into the role of EMT in tumor immune escape. Molecular oncology. 2017;11(7):824–46.

37. Gabrilovich D, Nagaraj S. Myeloid-derived suppressor cells as regulators of the immune system. Nature reviews Immunology. 2009;9(3):162–74.

38. Kerkar S, Restifo N. Cellular constituents of immune escape within the tumor microenvironment. Cancer research. 2012;72(13):3125–30.

39. Gajewski T, Schreiber H, Fu Y. Innate and adaptive immune cells in the tumor microenvironment. Nature immunology. 2013;14(10):1014–22.

40. Noman M, Janji B, Abdou A, Hasmim M, Terry S, Tan T, et al. The immune checkpoint ligand PD-L1 is upregulated in EMT-activated human breast cancer cells by a mechanism involving ZEB-1 and miR-200. Oncoimmunology. 2017;6(1):e1263412.

41. Chen L, Gibbons D, Goswami S, Cortez M, Ahn Y, Byers L, et al. Metastasis is regulated via microRNA-200/ZEB1 axis control of tumour cell PD-L1 expression and intratumoral immunosuppression. Nature communications. 2014;5:5241.

42. Mahmood M, Ward C, Muller H, Sohal S, Walters E. Epithelial mesenchymal transition (EMT) and non-small cell lung cancer (NSCLC): a mutual association with airway disease. Medical oncology (Northwood, London, England). 2017;34(3):45.

43. Kalluri R. EMT: when epithelial cells decide to become mesenchymal-like cells. J Clin Investig. 2009;119(6):1417–9.

44. Wang Y, Zhang Q, Gao Z, Xin S, Zhao Y, Zhang K, et al. A novel 4-gene signature for overall survival prediction in lung adenocarcinoma patients with lymph node metastasis. Cancer cell international. 2019;19:100.

45. Mak MP, Tong P, Diao L, Cardnell RJ, Gibbons DL, William WN, et al. A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. Clin Cancer Res. 2016;22(3):609–20.

46. Cao R, Yuan L, Ma B, Wang G, Qiu W, Tian Y. An EMT-related gene signature for the prognosis of human bladder cancer. J Cell Mol Med. 2020;24(1):605–17.

47. Chen M, Pan H, Sun L, Shi P, Zhang Y, Li L, et al. Structure and regulation of human epithelial cell transforming 2 protein. Proc Natl Acad Sci U S A. 2020;117(2):1027–35.
48. Justilien V, Ali SA, Jamieson L, Yin N, Cox AD, Der CJ, et al. Ect2-Dependent rRNA Synthesis Is Required for KRAS-TRP53-Driven Lung Adenocarcinoma. Cancer Cell. 2017;31(2):256–69.

49. Zhang W, He W, Shi Y, Zhao J, Liu S, Zhang F, et al. Aberrant TIMELESS expression is associated with poor clinical survival and lymph node metastasis in early-stage cervical carcinoma. Int J Oncol. 2017;50(1):173–84.

50. Boscolo-Rizzo P, Giunco S, Rampazzo E, Bruttì M, Spinato G, Menegaldo A, et al. TERT promoter hotspot mutations and their relationship with TERT levels and telomere erosion in patients with head and neck squamous cell carcinoma. J Cancer Res Clin Oncol. 2020;146(2):381–9.

51. Sun L, Yu S, Xu H, Zheng Y, Lin J, Wu M, et al. FHL2 interacts with EGFR to promote glioblastoma growth. Oncogene. 2018;37(10):1386–98.

52. Pasquale E. Eph receptors and ephrins in cancer: bidirectional signalling and beyond. Nature reviews Cancer. 2010;10(3):165–80.

53. Jackson HW, Defamie V, Waterhouse P, Khokha R. TIMPs: versatile extracellular regulators in cancer. Nat Rev Cancer. 2017;17(1):38–53.

54. Liang W, Zhang L, Jiang G, Wang Q, Liu L, Liu D, et al. Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. J Clin Oncol. 2015;33(8):861–9.

Figures
Figure 1

The performance of LASSO regression analysis and establishment of prognostic gene signature. (A) LASSO coefficient profiles of the 46 genes. (B) The optimal parameter in the LASSO model. (C) After multivariate regression analysis, the 23-gene signature was established and visualized by forest plot.
Figure 2

The gene expression level and genetic alteration of the 23 genes. (A-B) The difference of gene expression level in normal and tumor TCGA-LUAD patients. (C) Genetic alteration of the 23 genes in 507 LUAD specimens.
Figure 3

Characteristics of the prognostic gene signature. (A) Heatmap of the 23 genes expression profiles in TCGA-LUAD cohorts. (B-C) The distribution of risk score and survival time for TCGA-LUAD cohorts. (D) The K-M curves showed patients with high risks were related to poor survival rate. (E) Time-dependent ROC curves to describe the specificity and sensitivity of the risk score.
Figure 4

Establishment and validation of a nomogram. (A) The risk score served as an independent factor for prognostic prediction. (B) The risk associated nomogram was established for prognostic prediction. (C) ROC curves of 1-, 3- and 5-year in TCGA-LUAD cohorts. (D) Calibration plots of 1-, 3- and 5-year in TCGA-LUAD cohorts.
Figure 5

Enrichment analysis and PPI network. (A) GO analysis of the 23 EMT, immune and stromal associated genes and 300 altered genes. (B) Proteins interaction of the 323 molecules (C) GSEA analysis between cohorts with low and high risk.
Figure 6

immune microenvironment and tumor staging are associated with risk score. (A) The differential composition of tumor-infiltrating immune cells in cohorts with low and high risk. (B) Heatmap of tumor purity, ESTIMATE score, immune score and stromal score in low and high group. (C) Immune score, stromal score and tumor purity were significantly associated to risk score. (D) Tumor staging and TNM staging were associated to risk score.
Figure 7

Validation of the 23-gene signature in GEO LUAD cohorts. (A) K-M curve showed that patients with a high risk had a worse outcome. (B) The distribution of risk score and survival time for GEO-LUAD cohorts. (C) ROC curves of 1-, 3- and 5-year in GEO validation cohorts. (D) Calibration plots of 1-, 3- and 5-year in GEO validation cohorts.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.docx