Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A novel epithelial-mesenchymal transition-related gene signature for prognosis prediction in patients with lung adenocarcinoma

Shengyu Feng [a,1], Ce Huang [a,1], Liuling Guo [a], Hao Wang [a], Hailiang Liu [a,b,*]

[a] Institute for Regenerative Medicine, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, 200123, China
[b] Key Laboratory of Xinjiang Phytomedicine Resource and Utilization of Ministry of Education, College of Life Sciences, Shihezi University, Shihezi, 832003, China

ARTICLE INFO

Keywords:
Epithelial-mesenchymal transition
Lung adenocarcinoma
Gene signature
Risk model
Immunity

ABSTRACT

Traditional pathological diagnoses and clinical methods are insufficient to accurately predict the prognosis of lung adenocarcinoma (LUAD). Epithelial-mesenchymal transition (EMT) process is closely related to tumor cell migration. However, the prognostic value of EMT-related genes in LUAD is still unclear. In this study, we collected bulk RNA-sequencing (RNA-seq) and microarray data of LUAD patients from public databases and identified different expressed EMT-related genes in tumor and normal tissues. Then, we used the least absolute shrinkage and selection operator Cox regression model to develop a multigene signature in the cancer genome atlas (TCGA) cohort and validated the model in the OncoSG (Singapore Oncology Data Portal) cohort as well as other datasets. Finally, we constructed a 12-gene signature to divide LUAD patients into high-risk and low-risk groups of overall survival (OS), which has a better stability and accuracy in predicating the OS of patients compared with some other published signatures of LUAD. In addition, evaluation of the risk model using the time-related receiver operating characteristic (ROC) curve confirmed the predictive ability of the model. Functional analysis showed that these genes are related to immunity. CD8 T cell and CD4 T cell types were significantly negatively correlated with the risk score in the analysis of immune infiltration. In general, our model provides useful information that may help clinicians better predict the prognosis of LUAD patients and provides potential targets for immunotherapy of LUAD.

1. Introduction

Lung cancer is the deadliest cancer with the highest morbidity and mortality worldwide and more than 1.3 million cases reported annually [1]. Lung cancer has two main types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC accounts for 85% of all cases, and lung adenocarcinoma represents the most common subtype of NSCLC, comprising approximately 40% of all lung cancer cases [2]. The standard therapeutic approach for LUAD is surgical resection, but tumor recurrence and metastasis limit the curative effect of resection, leading to a poor prognosis of LUAD patients. In the United States, the five-year OS rate of patients with metastatic NSCLC is less than 5% [3, 4]. Furthermore, despite great improvements in surgery, chemotherapy and radiation therapy, the mortality and metastasis rates remain high. Extensive evidence has shown that LUAD tends to metastasis at early stages; however, specific and sensitive biomarkers for the early detection of LUAD remain limited. Thus, identifying metastasis-related biomarkers that can accurately detect early-stage LUAD and even predict prognosis is crucial.

EMT is a biological process by which cells lose their epithelial characteristics and acquire mesenchymal characteristics. EMT has long been thought to be an one-way process involving a complete switch from epithelial state to mesenchymal state. However, it is becoming increasingly clear that EMT comprises a variety of hybrid forms, a phenotype known as "partial EMT" (P-EMT) [5, 6, 7]. Recent studies have demonstrated that EMT is closely related to tumor initiation and tumor cell migration [8]. EMT enhances cell migration, invasiveness and resistance to therapies and imparts polymer transfer properties [9]. EMT is affected by many factors, such as SNAIL, CXCL13, TWIST1 and ZEB1 [10, 11]. The SNAIL transcription factor superfamily plays an important role in the regulation of EMT. EMT can be induced by inhibiting the expression of the epithelial marker gene *CDH1* which encodes E-cadherin or increasing the expression of EMT drivers, such as ZEB-1 and ZEB-2. The expression of the repressor of cadherin can weaken the metastasis inhibitory

[*] Corresponding author.
E-mail address: hailiang_1111@tongji.edu.cn (H. Liu).
[1] These authors made equal contributions to this manuscript.

https://doi.org/10.1016/j.heliyon.2022.e08713
Received 9 July 2021; Received in revised form 16 November 2021; Accepted 31 December 2021

function of E-cadherin [12, 13]. In hepatic cell carcinoma and invasive lobular carcinoma, SNAIL1 of the SNAIL superfamily plays a role in tumor invasion, metastasis and poor differentiation, and its expression can be used as a predictor of poor patient prognosis [14, 15]. TWIST1 is a transcriptional repressor of E-cadherin, which simultaneously induces the transcription of N-cadherin and fibronectin and is associated with the development of several tumors [16, 17]. The ZEB family is similar to the SNAIL family as both consist of transcription factors with zinc finger domains. To regulate EMT, these SNAIL, CXCL13, TWIST1 and ZEB1 factors prevent the interaction between E-cadherin and desmosomes and upregulate N-cadherin and matrix metalloprotinases, causing epithelial cells to lose their polarity and adhesion abilities. In melanoma and thyroid cancer, researchers have found that the overexpression of ZEB can promote tumor cell invasion and metastasis [18, 19].

In this study, we downloaded the mRNA expression profiles, corresponding clinical data of LUAD patients and EMT-related genes from TCGA and dbEMT2.0 databases respectively. We comprehensively used more than 1,000 EMT-related genes in the EMT database for the analysis of this article, and screened more than 200 genes involved in the pathogenesis of LUAD. Then we developed a prognostic multi-gene signature with a good predictability using TCGA cohort, which was verified in the OncoSG cohort, compared with some other prognostic gene signatures of LUAD, the signature constructed in this paper have higher accuracy. Finally, enrichment analysis showed a strong correlation between EMT process and immune infiltration in LUAD, which provide a new target for immunotherapy. Therefore, the EMT-related gene risk signature can be used as an important indicator for predicting the prognosis of LUAD patients.

2. Materials and methods

Data collection and analysis: The RNA-seq data of LUAD patients were downloaded from TCGA. Data were obtained from 525 LUAD patients, and 493 samples were selected for subsequent analysis after removing 32 samples with missing prognostic statistics. The form of the downloaded gene expression data was log2 (count+1). The original data were converted into raw read count values, and normalized read values were used for the analysis. In addition, the RNA-seq data of 305 patients were downloaded from the OncoSG portal (https://src.gisapps.org/OncoSG_public/). Similarly, normalized read count values were used for subsequent data analyses. The TCGA cohort was used as the training set, and the OncoSG cohort was used as the validation set. The data from TCGA and OncoSG are both available on their official websites. The microarray data were downloaded from the NCBI GEO (Gene Expression Omnibus) database and analyzed by GEO2R tool.

Identification of DEGs related to EMT: The R package "DESeq2" [20] was used to identify DEGs (differentially expressed genes) between LUAD patients' tumor tissues and adjacent normal tissues. The DEG screening criteria were set with a false discovery rate (FDR) < 0.01 and $|\log2$ fold change $| \geq 1$. We also used the online analysis website GEPIA [21] (http://gepia.cancer-pku.cn/) to identify the DEGs in LUAD with the same threshold as above. The intersection between DEGs calculated by the two methods was used as the DEGs for further analysis. The interaction prediction of these genes was performed using the STRING database [22]. We downloaded 1012 genes related to EMT from the online database dbEMT2 [23] (http://dbemt.bioinfo-minzhao.org/), then the R package "VennDiagram" [24] was used to calculate the intersection of DEGs and EMT-related genes for subsequent analysis.

Construction and verification of the EMT-based prognostic risk model: Univariate Cox analysis of OS was applied to identify EMT-related genes with significant prognostic values. LASSO Cox regression analysis was performed to identify independent EMT-related genes with prognostic significance and calculate the risk regression coefficient of each gene. We then used the following formula to construct a prognostic risk model: risk score = sum (expression of each gene × corresponding coefficient). The above analysis was conducted using R. All patients were divided into a low-risk group and a high-risk group on the basis of the median of risk scores. The overall survival between the low-risk and high-risk groups was compared by Kaplan–Meier analysis with the log-rank test. The "survivalROC" R package [25] was applied to perform time-dependent ROC curve analysis to assess the predictive accuracy of the gene signature. Finally, the LUAD gene expression data in the OncoSG database were used to verify the model constructed above.

Functional enrichment analysis of the 12-gene signature: The R package "DESeq2" was used to identify the DEGs in the low-risk and high-risk group with an FDR <0.01 and $|\log2$ fold change $| \geq 1$. GO and KEGG pathway enrichment analyses were performed on DEGs. Functional enrichment analysis and visualization were conducted using the R packages "clusterProfiler" and "GOplot" [26], and P-values were adjusted using the BH method. GSEA (Gene Set Enrichment Analysis) enrichment analysis was performed using the GSEA-P tool [27].

Evaluation of immune cell infiltration: The CIBERSORT method was used to assess the proportion of immune cells in different risk groups [28]. In total, we analyzed 22 human immune cell phenotypes in this study, including naive and memory B cells, gamma delta T cells, monocytes, plasma cells, and others.

Statistical analysis: All statistical tests were performed with R (version 3.6.1). The t-test was applied to compare gene expression between LUAD tissues and adjacent normal tissues and compare age, gender and tumor stage status in LUAD. The OS between low and high-risk groups was compared by Kaplan–Meier analysis with the log-rank test. All tests were two-tailed, and a p value <0.05 was considered significant.

3. Results

Clinical characteristics of patients: We collected the RNA-seq data and clinical information of 493 LUAD patients from the TCGA database to construct a prognosis risk model based on EMT-related genes and obtained the RNA-seq data and clinical information of 305 LUAD patients from OncoSG, which were used to verify the risk model. The data in OncoSG were collected from Lung Cancer Consortium Singapore. The specific sample information is shown in Table 1.

Identification of EMT-related DEGs in the TCGA cohort: Through t-SNE analysis of the data obtained from TCGA, we observed that the tumor tissues and adjacent normal tissues from LUAD patients were distributed in two different clusters (Figure 1A). We then identified the genes that were differentially expressed between tumor tissues and adjacent normal tissues both in our local analysis using R and online analysis with GEPIA tool. As shown in Figure 1B, there were 2147 upregulated and 843 downregulated genes in our local analysis and in common. The 2990 selected genes were compared with the dbEMT2.0 dataset, resulting in the identification of 93 upregulated EMT-related genes and 158 downregulated EMT-related genes, totaling 251 genes (Figure 1C). These genes were differentially expressed between tumors and normal tissues (Figure 1D). In addition, we performed an enrichment analysis of the identified 251 genes using MsigDB:HALLMARK and KEGG databases. The results revealed that epithelial-mesenchymal pathways in cancer, immune-related signals, such as TNF-α signaling via NFKB, and inflammatory responses were significantly enriched (Figure 1E).

Screening of EMT-associated DEGs related to LUAD tumor stage: The prognosis of LUAD is closely related to the stage of the tumor, and the metastasis of the tumor is an important factor in determining the stage. Recent studies have shown that EMT is closely associated with tumor initiation and tumor cell migration [8]. Therefore, we performed a time series analysis of these 251 genes to determine their relationship with LUAD tumor stage. We found that there were six main expression trends and 59 EMT-related genes were closely related to tumor stage (Supplementary Table 1), which were shown in Figure 2. Therefore, we used these tumor stage-related DEGs for further analysis.

Construction of the risk model of prognosis in the TCGA cohort: We identified 26 DEGs that were strongly correlated with patients' OS (P

Table 1. Clinical characteristics of TCGA LUAD and OncoSG LUAD patients.

	TCGA-LUAD cohort	LCCS-LUAD cohort
Num of Patients	493	305
Gender (num, %)		
Female	263 (53.3%)	149 (48.9%)
Male	230 (46.7%)	158 (51.1%)
Age at Diagnosis (num, %)		
<=49	98 (18.3%)	29 (9.5%)
50–59	148 (27.7%)	77 (25.2%)
60–69	148 (27.7%)	110 (36.1%)
70–79	114 (21.3%)	83 (27.2%)
>80	26 (5%)	6 (2.0%)
Tumor Stage (num, %)		
Stage I	268 (54.4%)	136 (44.6%)
Stage II	118 (24.0%)	57 (18.7%)
Stage III	80 (16.3%)	91 (29.8%)
Stage IV	25 (5.1%)	19 (6.2%)
Not report	2 (0.4%)	2 (0.6%)
Radiation Therapy (num, %)		
Yes	57 (11.6%)	na
No	341 (69.2%)	na
Not report	95 (19.3%)	na
Histological grade		
Well-differentiated	na	17 (5.6%)
Moderately-differentiated	na	134 (43.9%)
Poorly differentiated	na	30 (9.8%)
Not report	na	124 (40.7%)
Smoking History		
Yes	412 (83.6%)	112 (36.7%)
No	67 (14.0%)	189 (62.0%)
Not report	14 (2.4%)	4 (1.3%)
Mutation Count		
High (>51)	na	149 (48.9%)
Low (<51)	na	153 (50.2%)
Not Report	na	3 (0.9%)
Survival Time		
OS months (median)	36.3	21.4

< 0.05) using univariate Cox proportional hazards regression analysis in the training set (Figure 3A). Twelve of the 26 DEGs were low-risk genes, and the other 14 DEGs were high-risk genes for OS in LUAD patients. The co-expression analysis of these genes was shown in Figure 3B. Then, LASSO Cox regression analysis was performed to establish the prognostic model. Here, we determined a signature containing 12 genes, including *TYMS, ADAM12, GJB2, KRT8, LYPD3, ECT2, PSTPIP1, NDRG2, MAP3K3, SMAD9, ADM* and *ID2*, based on the best λ value (Figure 3C and D). Further, we verified the expression of these 12 genes in GSE33532 [29], GSE30219 [30] and GSE19804 [31] databases from the GEO database, they are indeed differentially expressed in normal lung tissues and tumor tissues of LUAD patients (Supplementary Figure 1). Then we obtained the following risk score calculation model:

Risk Score = 0.08 × geneExp (*TYMS*) + 0.03 × geneExp (*ADAM12*) + 0.01 × geneExp (*GJB2*) + 0.08 × geneExp (*KRT8*) + 0.04 × geneExp (*LYPD3*) + 0.01 × geneExp (*ECT2*) + 0.09 × geneExp (*ADM*) – 0.14 × geneExp (*PSTPIP1*) – 0.06 × geneExp (*NDRG2*) – 0.05 × geneExp (*MAP3K3*) – 0.09 × geneExp (*SMAD9*) – 0.06 × geneExp (*I* D2)

The median risk score value was used to classify patients into a high-risk group and a low-risk group (Figure 4A). According to t-SNE analysis, patients in the high-risk group and low-risk group were distributed in two clusters (Figure 4B). In addition, as shown in Figure 4C, the OS decreased with the increase of risk value. Similarly, the OS in the high-risk group was significantly worse than the OS in the low-risk group (Figure 4D). Finally, we evaluated the predictive performance of the risk model using the time-related ROC curve, and the area under the curve (AUC) reached 0.798 at 1 year, 0.669 at 2 years and 0.695 at 3 years (Figure 4E). These results demonstrate that our model can effectively predict the prognosis of LUAD patients.

Verification of the 12-gene signature in the OncoSG cohort: To test the validity of the model constructed with the TCGA cohort, we used the gene expression data in the OncoSG LCCS cohort to calculate the risk score of patients based on the formula constructed above, then divided patients into high-risk group and low-risk group (Figure 5A). The results obtained were similar to those in the TCGA cohort. The OS decreased with the increase of risk value (Figure 5B). The t-SNE analysis revealed that patients in different risk groups were distributed in two clusters (Figure 5C). Similarly, patients in the low-risk group had a better overall survival rate (Figure 5D). In addition, the time-related ROC curve analysis indicated that the AUC with those 12-gene markers was 0.798 at 1 year and 0.803 at 2 years (Figure 5E). This finding shows that our model is not only applicable to the TCGA cohort, but it also has a high prognostic ability in the OncoSG cohort.

Comparison with other signatures of LUAD: There are some already published prognostic signatures for lung adenocarcinoma, but the genes they used to build the model were not so strongly associated with cancer development, and they didn't compare it with other signature. Therefore, We calculate the risk scores of patients in the two additional databases, GSE30219 and GSE31210 [32], using the method applied in this paper and the method mentioned in Liu et al [33] and Huang et al [34], respectively. It can be clearly seen from the results that the signature constructed in this paper has a better prediction effect, while the signature in Liu et al only has a good predict performance in GSE31210 data set, and the signature in Huang et al has a poor predict performance in both data sets (Supplementary Figure 2). This is further proof that the signature we built has more accurate predictive results and it's more suitable to predict the prognosis of lung adenocarcinoma.

Independent prognosis value of the 12-gene signature: Furthermore, we explored whether the risk value can independently affect the survival prognosis of patients. The results of the univariate Cox analysis are shown in Table 2. Tumor stage, age, smoking status and risk value were significantly related to prognosis in the TCGA cohort (Table 2). The multivariate Cox analysis result showed that the risk score model was significantly correlated with OS [hazard ratio (HR) = 2.01, 95% confidence interval (CI) = 1.41–2.93, $P < 0.01$) in the TCGA training cohort. We obtained similar results in OncoSG cohort, and the univariate Cox analysis indicated that tumor stage and risk value were significantly related to the prognosis of LUAD patients. The nomograms of this model in the training set and the testing set are shown in Supplementary Figure 3.

Functional Annotation of the Signature of 12 EMT-Related Genes: To explore the biological functions and pathways related to the risk model in LUAD patients, we conducted KEGG and GO analysis of the DEGs between the high-risk group and the low-risk group. The results showed that the DEGs were enriched in several EMT-related biological processes, such as focal adhesion and cell-substrate adhesion ($P < 0.001$) (Figure 6A, B and C). In addition, these genes were highly enriched in immune-related pathways, including the IL-17 signaling pathway, regulation of T-helper 1 cell cytokine production and cytokine secretion ($P < 0.001$). Through GESA analysis, we obtained results similar to those above. EMT pathways, angiogenesis and TNFα signaling via NFKB were significantly enriched (Figure 6D and E). Similar results were observed in the OncoSG cohort (Supplementary Figure 4A).

Associations of EMT-related gene risk scores with immune cell infiltration: The enrichment analysis showed in the Figure 6 revealed that EMT-related gene risk scores were highly relevant to the immune status. We calculate the enrichment scores of multiple immune cell subpopulations using CIBERSORT.

Interestingly, CD8 T cells and naive B cells were significantly enriched in the low-risk group. However, some immune cells, such as macrophages

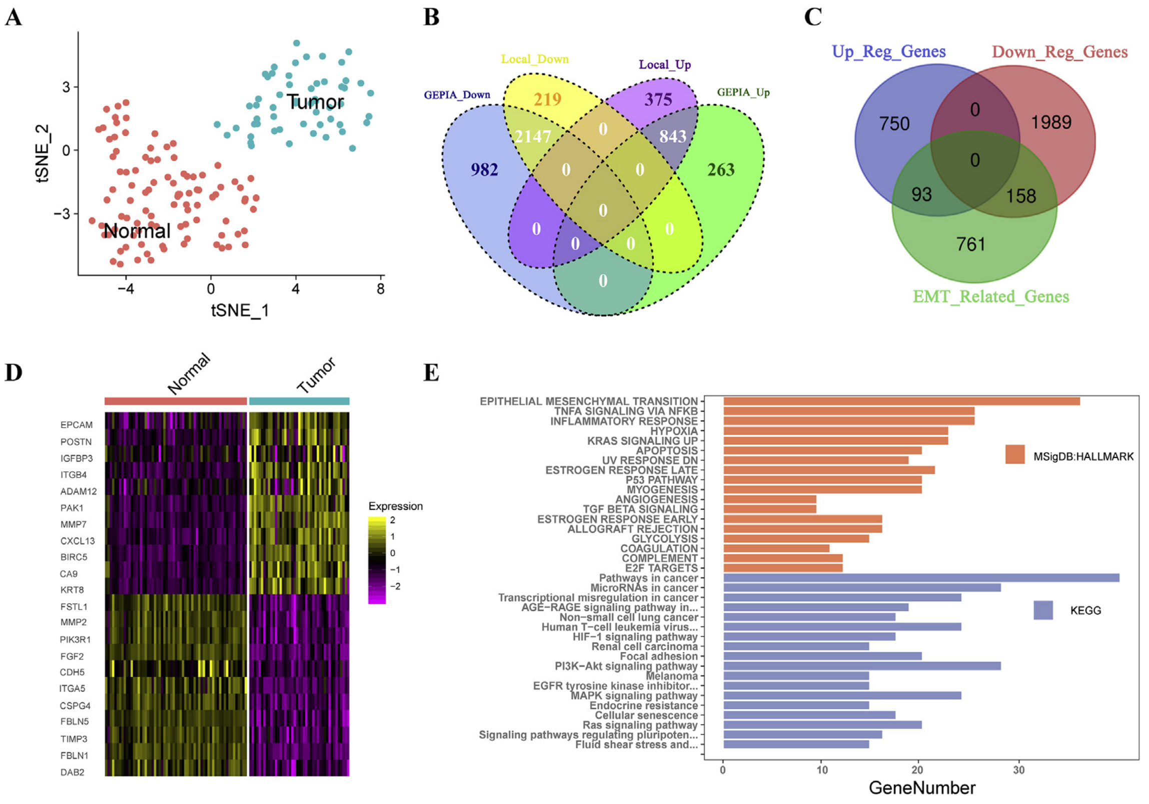

Figure 1. Identification of candidate genes related to EMT in the TCGA cohort. (A) t-SNE analysis of TCGA cohort. (B) The Venn diagram analysis of DEGs in the local and online tool analysis. (C) Venn diagrams used to identify DEGs related to EMT. (D) The expression heatmap of some DEGs between normal and cancer tissues. (E) Enrichment analysis of 251 EMT-related DEGs using MsigDB: HALLMARK and KEGG databases.

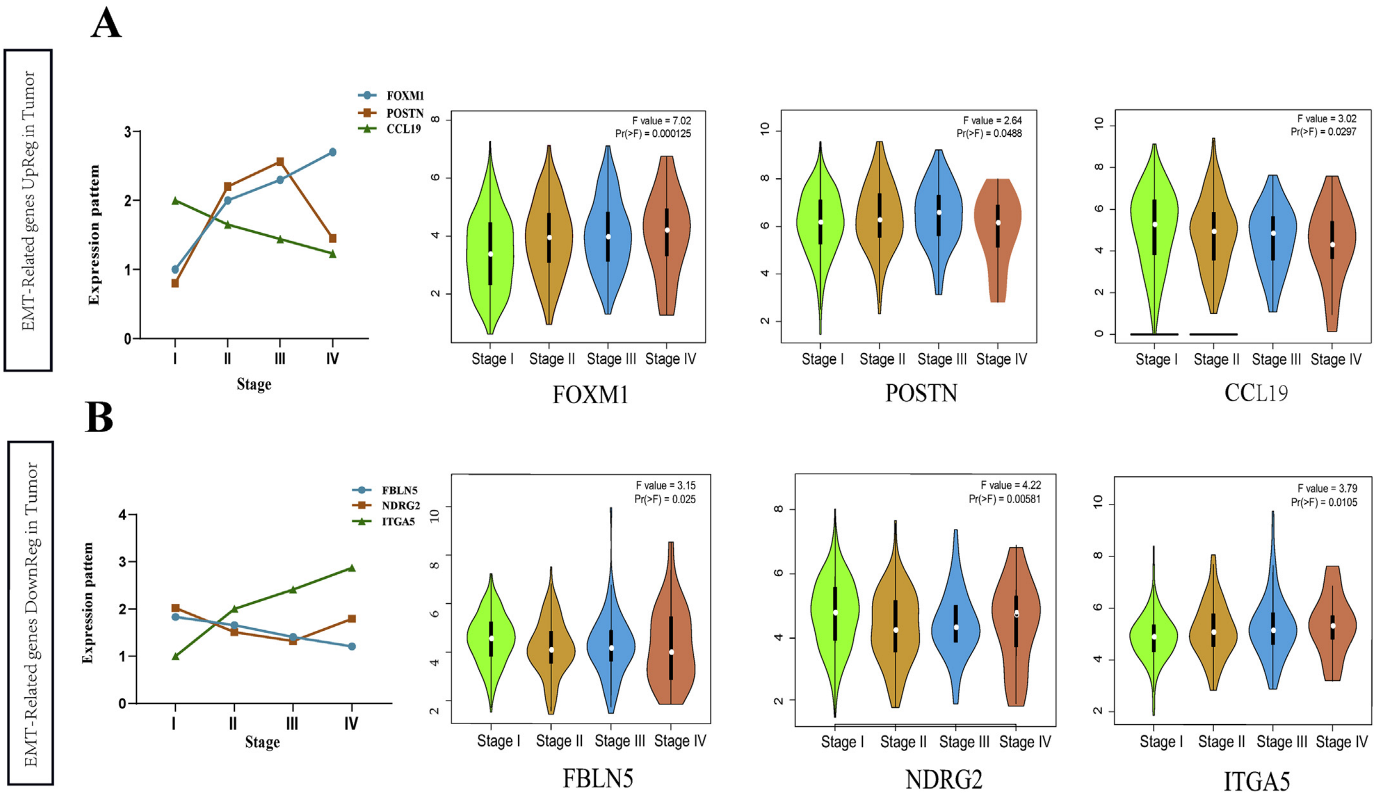

Figure 2. Screening of EMT-associated DEGs related to LUAD tumor stage. (A) The expression trends of three example EMT-related genes that were upregulated in tumor. (B) The expression trends of three example EMT-related genes that were downregulated in tumor. The correlation analysis was performed with the online tool GEPIA. The method for differential gene expression analysis was a one-way ANOVA using the pathological stage as a variable for calculating differential expression.

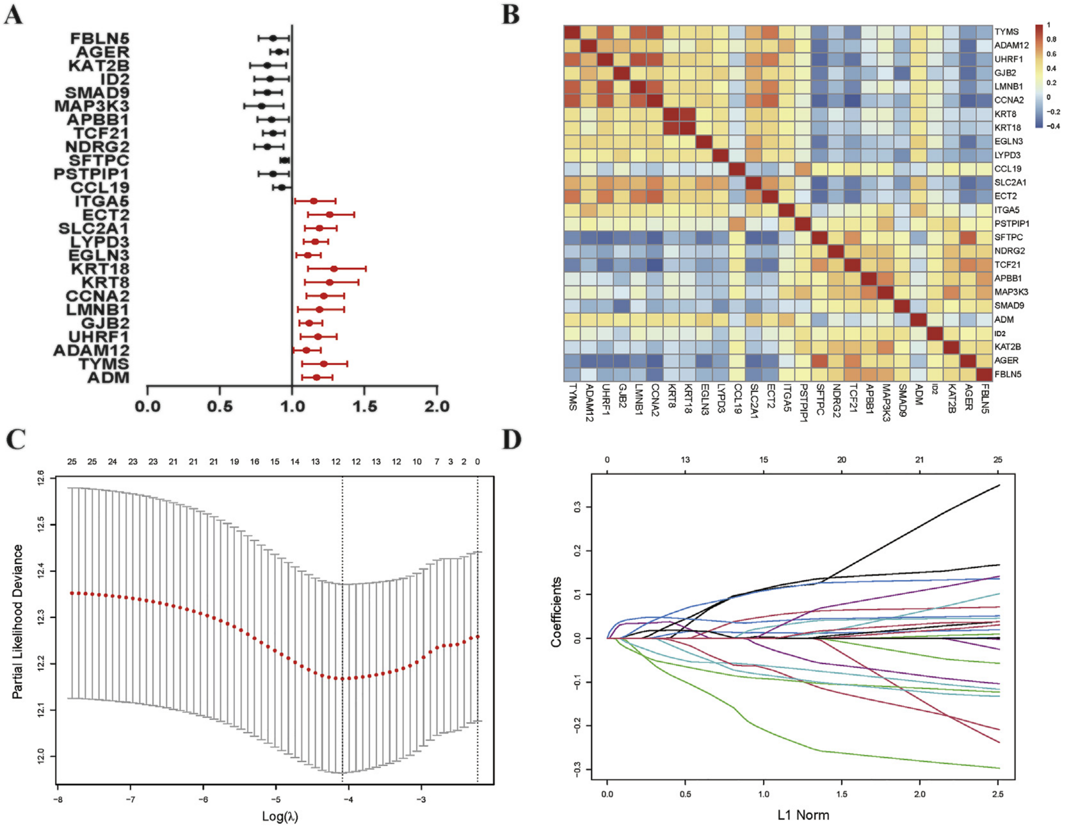

Figure 3. Establishment of the prognostic risk model in the TCGA cohort. (A) Univariate Cox analysis identified 26 EMT-related DEGs that were correlated with the overall survival of LUAD patients. Red lines indicate high-risk genes, and black lines indicate low-risk genes. (B) Co-expression analysis of genes in the final signature. (C) 1,000-fold cross-validation for tuning parameter selection in the least absolute shrinkage and selection operator model. (D) LASSO coefficient profiles of the most useful prognostic genes. Each line indicates an individual gene in the LASSO model.

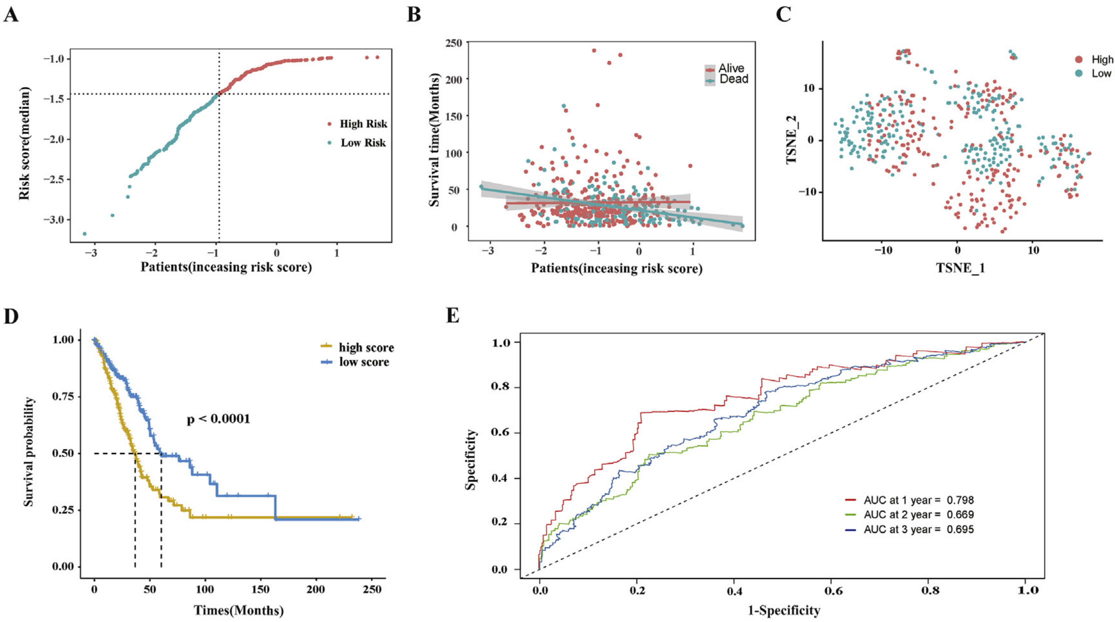

Figure 4. Prognostic analysis of 12-gene signature model in the TCGA cohort. (A) Distribution of risk scores in TCGA cohort. (B) t-SNE analysis of TCGA cohort. (C) Distribution of OS status, OS and risk score in TCGA cohort. (D) Kaplan–Meier curve of OS for patients in the high-risk group and low-risk group. (E) The AUC of the ROC curve over time in the TCGA cohort.

and activated dendritic cells, were significantly enriched in the high-risk group in the TCGA cohort (Figure 7, all adjusted $P < 0.05$). These results were consistent with the previous functional enrichment analysis. Immune infiltration analysis of the OncoSG cohort showed that in addition to CD8 T cells, activated CD4 T cells and natural killer cells were also enriched in the low-risk group (Supplementary Figure 4B).

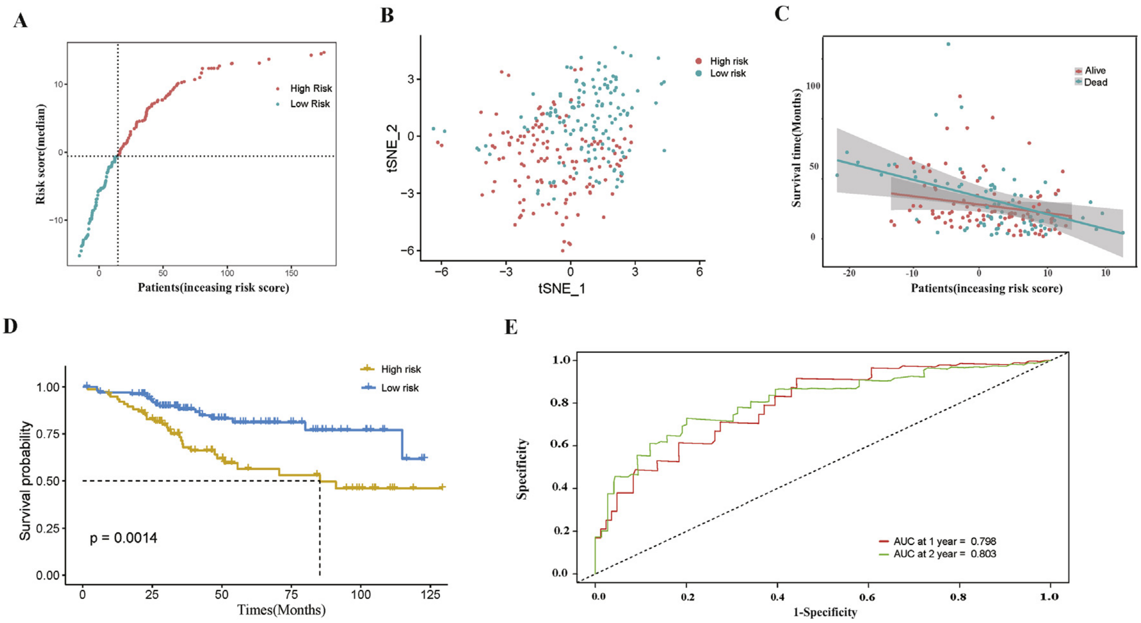

Figure 5. Verification of the 12-gene signature in the OncoSG cohort. (A) Distribution of risk scores in the OncoSG cohort. (B) t-SNE analysis of OncoSG cohort. (C) Distribution of OS status, OS and risk score in the OncoSG cohort. (D) Kaplan–Meier curve of OS for patients in the high-risk group and low-risk group. (E) The AUC of the ROC curve over time in the OncoSG cohort.

Table 2. Univariate and multivariate Cox analysis of the 12-gene signature in the two LUAD cohorts.

Factors	TCGA-cohort (Training)				LCCS-cohort (Validation)			
	Univariate Cox analysis		Multivariate Cox analysis		Univariate Cox analysis		Multivariate Cox analysis	
	HR (95% CI)	p	HR (95% CI)	P	HR (95% CI)	p	HR (95% CI)	p
Gender								
Male	Reference				Reference			
Female	1.32 (0.74–1.87)	0.42			1.49 (0.84–2.81)	0.21		
Age								
Young (>60)	Reference				Reference			
Old (<60)	1.36 (0.98–2.01)	0.05	1.21 (0.62–1.54)	0.23	1.53 (0.88–1.98)	0.14		
Stage								
Low (Stage i, stage ii)					Reference			
High (stage iii, stage iv)	1.98 (1.21–2.78)	<0.01	1.67 (0.89–2.48)	0.01	2.05 (1.45–2.84)	<0.01	1.85 (1.21–2.36)	0.02
Histological Grade								
Well differentiated					Reference			
Poorly differentiated					1.35 (0.75–1.47)	0.122		
Radiation Therapy								
Yes	Reference							
No	1.12 (0.78–1.23)	0.51						
Smoking History								
No	Reference				Reference			
Yes	1.32 (0.87–1.65)	0.05	1.54 (0.98–2.87)	0.12	1.52 (0.8–2.9)	0.21		
Risk Score								
Low risk	Reference				Reference			
High risk	2.12 (1.14–2.95)	<0.01	2.01 (1.41–2.93)	<0.01	1.51 (1.21–2.08)	<0.01	1.23 (0.98–1.52)	0.04

4. Discussion

In this study, we used public LUAD gene expression datasets to identify 12 EMT-related genes that can be used as prognostic predictors of LUAD. For the first time, we established a 12 EMT-related gene signatures and validated it in the OncoSG cohort, providing a novel prognostic model of LUAD related to EMT. Significantly, this model showed good prognostic value in TCGA and several other datasets. Besides, compared with other published signatures in LUAD, the signature constructed in this paper has better stability and accuracy.

It is well known that LUAD is a malignant disease with high heterogeneity. Even for patients with similar clinical characteristics and medical histories, their prognoses vary substantially. Previously, the prediction of prognosis in LUAD patients was largely dependent on clinical and pathological analysis, but these methods are insufficient. It is difficult to detect tumor metastasis and recurrence in the early stages.

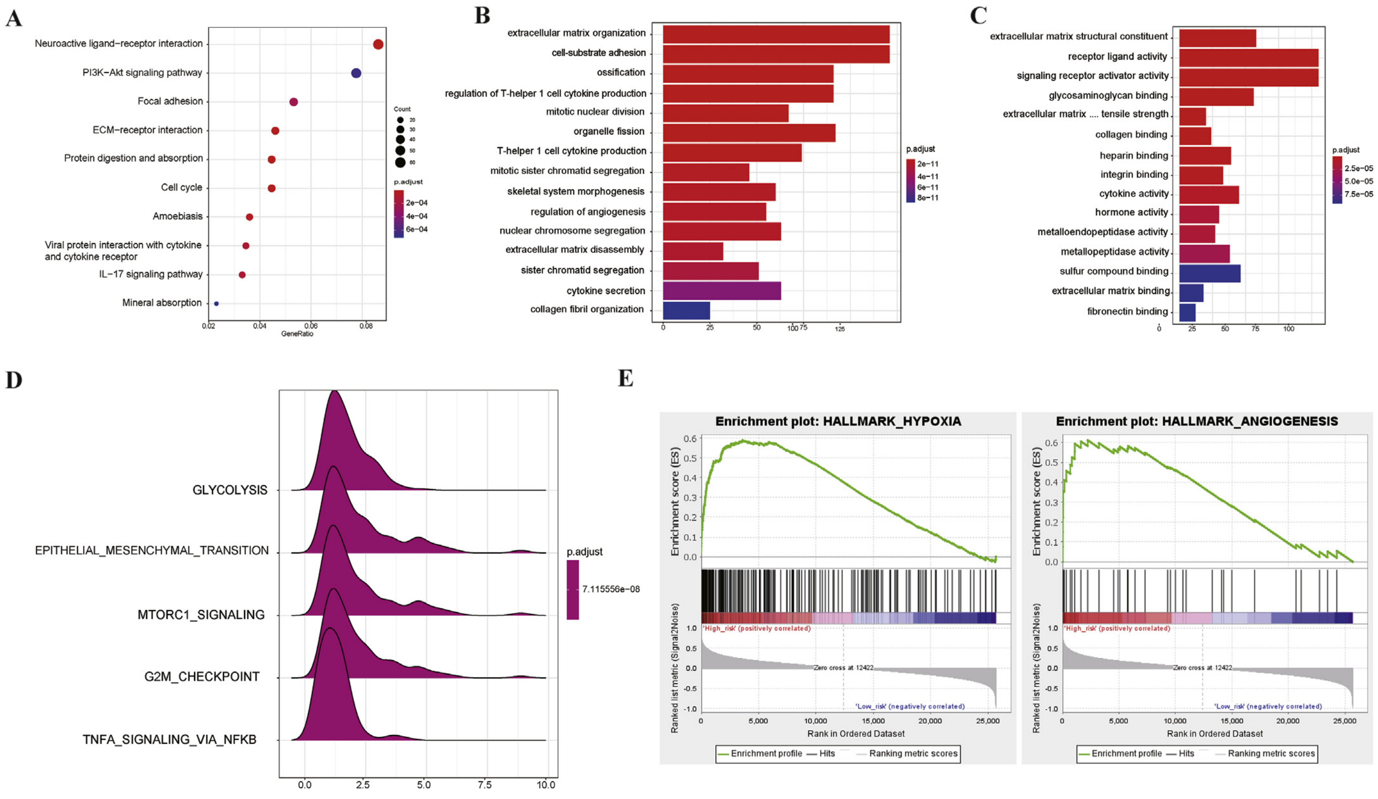

Figure 6. Functional analysis of risk model in TCGA cohort. (A–C) KEGG pathway, GO biological process and GO molecular function analysis of the DEGs between high-risk group and low-risk group. (D–E) GESA analysis of the high-risk group and low-risk group.

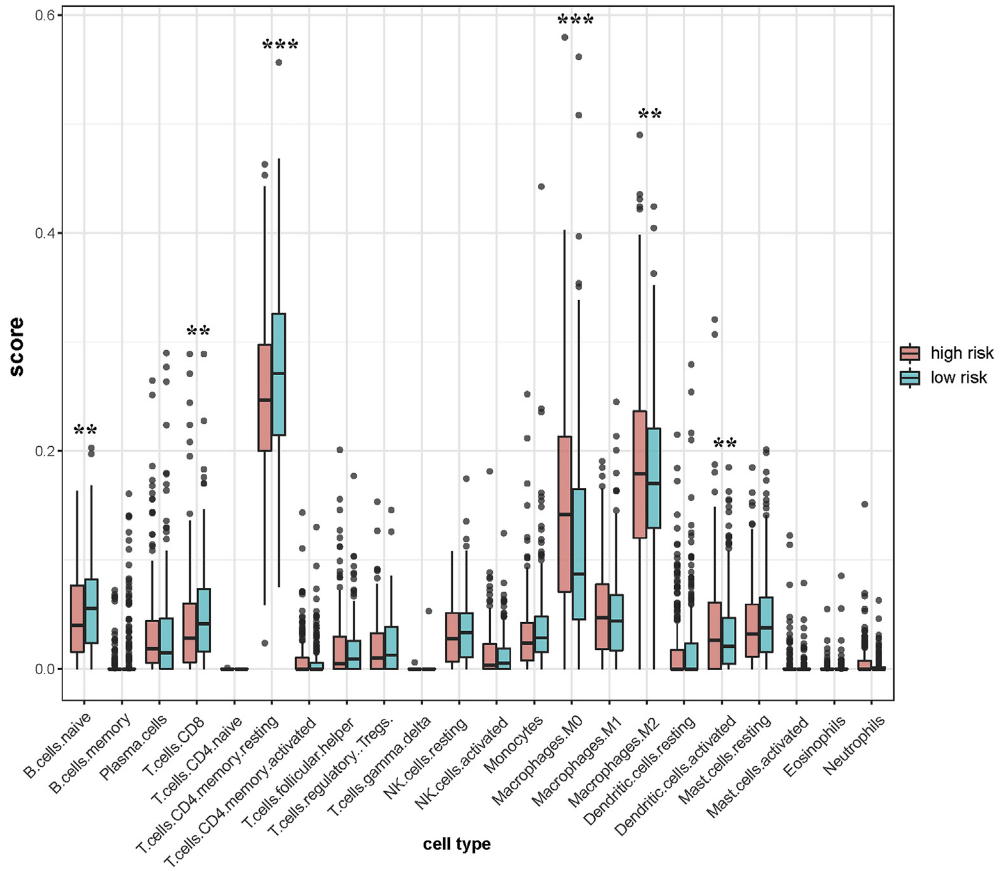

Figure 7. Comparison of the immune-related cells' scores between different risks groups in the TCGA cohort. The scores of 16 immune cells and their different cell states are displayed in boxplots.

With the advances in molecular biology and genomics, significant progress has been made in the prediction of tumor occurrence and development from a more microscopic and genetic perspective [35, 36, 37]. For example, 16 genes related to the survival of NSCLC patients were identified through the analysis of microarray data and risk scores, and multiple genes, including *DUSP6, MMD, STAT1, ERBB3* and *LCK,* were selected as a gene signature. The results showed that five gene markers were closely related to the OS rate (sensitivity = 98%, resolution = 93%, forward prediction conversion rate = 95%, reverse predictor variable = 98%, overall accuracy = 96%) [37]. In another study, researchers conducted a meta-analysis of datasets from seven different microarray studies of NSCLC to analyze DEGs related to the survival time (below 2 years and more than 5 years). Kaplan–Meier analysis of the OS rate in stage I of NSCLC patients with 64 gene expression characteristics showed that there was a significant difference between the OS rate in high-risk and low-risk groups [36]. However, most of those genes were shown to be highly expressed in the early stages of LUAD tumors. More importantly, these genes were not closely related to tumor metastasis or recurrence, which suggests that these gene signatures may fail in predicting the OS rate and recurrence-free survival of LUAD patients. So we compare our signature with two other signatures, which are constructed by glycolysis-related genes [33] and stem cell-related genes [34]. As expected, the signature constructed by EMT-related genes has a better stability and accuracy in predicting the OS of LUAD patients.

Immune infiltration of tumors is tightly related to clinical outcomes in LUAD [38]. Tumor-infiltrating immune cells (TIICs) affect the progression of cancer and are appealing therapeutic targets [39]. EMT has been an active research area in cancer biology in the past few years, however, the potential regulatory role between EMT and LUAD immunity is still elusive.

We performed KEGG and GO enrichment analysis based on the DEGs between high-risk group and low-risk group. Unexpectedly, our findings suggest that several biological pathways correlated with immunity have been enriched. The patients in high-risk groups in TCGA and OncoSG cohort have a higher fraction of macrophage M0, macrophage M2 and activated dendritic cells, on the contrary, CD8 T cell and memory CD4 T cell enriched in the low-risk group. Tumor killer cells were significantly reduced in the high-risk group, suggesting that EMT-related high-risk genes inhibit the immune response to tumor cells to some extent. Similarly, it has been reported that T cell differentiation in lung adenocarcinoma is shaped by tumor mutations [40] and EMT signature is inversely associated with T-cell infiltration in NSCLC [41]. So, the EMT-related high-risk genes may be a potential target for adenocarcinoma immunotherapy. Besides, Over the past decades, substantial literature has linked EMT to the pathophysiology of chronic obstructive pulmonary disease (COPD) [42, 43, 44], which is a recognized strong independent risk factor for the development of lung cancer. Our analysis results also showed that smoking status was significantly related to the risk grade. Abnormal phenotypes of M1/M2 macrophages in the small airway wall have been reported in smokers and those with COPD [45], and our analysis of immune infiltration showed that M0 and M2 macrophages were enriched in the high-risk groups, indicating that EMT-related genes may be involved in its progression of COPD by influencing the immune process.

There are limitations to this study. First, this study mainly focused on the role of EMT in predicting the prognosis of LUAD and did not investigate many other oncogenic genes that are actively involved in LUAD tumor cell proliferation and invasion. Secondly, the risk prediction model designed based on different EMT status markers may be more accurate in predicting the prognosis of patients because EMT is a state composed of various intermediate states rather than just epithelial and mesenchymal two status. Thirdly, this is a retrospective study driven by a hypothesis, and all data were obtained from public databases, therefore, the results still require relevant experimental and more clinical verification.

5. Conclusion

In summary, our study established a novel LUAD prognosis model based on EMT-related genes. This model has been demonstrated to be independently related to OS in TCGA and several other cohorts, and the prediction of our model is more accurate than some of the other gene signatures of LUAD. More importantly, these genes can be detected by RT-PCR, which is straightforward to use in a clinical setting. Our findings provide important indicators for the prognosis prediction of LUAD patients and a potential target for adenocarcinoma immunotherapy.

Declarations

Author contribution statement

Hailiang Liu: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Shengyu Feng: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Ce Huang & Hao Wang: Analyzed and interpreted the data; Wrote the paper.

Liuling Guo: Performed the experiments; Wrote the paper.

Funding statement

This work was supported by grants from the National Key Research and Development Program of China (2020YFC2002800), National Natural Science Foundation of China (31671539) and Major Program of Development Fund for Shanghai Zhangjiang National Innovation Demonstration Zone (ZJ2018-ZD-004).

Data availability statement

The main datasets of LUAD were collected from TCGA (https://portal.gdc.cancer.gov/), OncoSG (https://src.gisapps.org/OncoSG_public/) and GEO database. The raw microarray data is available at GEO: GSE30219, GSE33532, GSE31210 and GSE19804. The datasets used and/or analyzed during the current study are available from the corresponding author on request.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e08713.

References

[1] S. McGuire, World cancer report 2014. Geneva, Switzerland: World Health Organization, international agency for research on cancer, WHO press, 2015, Adv. Nutr. 7 (2) (2016) 418–419.

[2] H. Lemjabbar-Alaoui, O.U. Hassan, Y.W. Yang, P. Buchanan, Lung cancer: biology and treatment options, Biochim. Biophys. Acta 1856 (2) (2015) 189–210.

[3] B.A. Williams, H. Sugimura, C. Endo, F.C. Nichols, S.D. Cassivi, M.S. Allen, P.C. Pairolero, C. Deschamps, P. Yang, Predicting postrecurrence survival among completely resected nonsmall-cell lung cancer patients, Ann. Thorac. Surg. 81 (3) (2006) 1021–1027.

[4] K.C. Arbour, G.J. Riely, Systemic therapy for locally advanced and metastatic non-small cell lung cancer: a review, JAMA 322 (8) (2019) 764–774.

[5] A.D. Grigore, M.K. Jolly, D. Jia, M.C. Farach-Carson, H. Levine, Tumor budding: the name is EMT. Partial EMT, J. Clin. Med. 5 (5) (2016).

[6] N.M. Aiello, R. Maddipati, R.J. Norgard, D. Balli, J. Li, S. Yuan, T. Yamazoe, T. Black, A. Sahmoud, E.E. Furth, D. Bar-Sagi, B.Z. Stanger, EMT subtype influences epithelial plasticity and mode of cell migration, Dev. Cell 45 (6) (2018) 681–695, e4.

[7] I. Pastushenko, A. Brisebarre, A. Sifrim, M. Fioramonti, T. Revenco, S. Boumahdi, A. Van Keymeulen, D. Brown, V. Moers, S. Lemaire, S. De Clercq, E. Minguijón, C. Balsat, Y. Sokolow, C. Dubois, F. De Cock, S. Scozzaro, F. Sopena, A. Lanas, N. D'Haene, I. Salmon, J.C. Marine, T. Voet, P.A. Sotiropoulou, C. Blanpain, Identification of the tumour transition states occurring during EMT, Nature 556 (7702) (2018) 463–468.

[8] M.A. Nieto, R.Y. Huang, R.A. Jackson, J.P. Thiery, EMT: 2016, Cell 166 (1) (2016) 21–45.

[9] V. Mittal, Epithelial mesenchymal transition in tumor metastasis, Annu. Rev. Pathol. 13 (2018) 395–412.

[10] Y. Wang, J. Shi, K. Chai, X. Ying, B.P. Zhou, The role of snail in EMT and tumorigenesis, Curr. Cancer Drug Targets 13 (9) (2013) 963–972.

[11] S. Biswas, S. Sengupta, S. Roy Chowdhury, S. Jana, G. Mandal, P.K. Mandal, N. Saha, V. Malhotra, A. Gupta, D.V. Kuprash, A. Bhattacharyya, CXCL13-CXCR5 co-expression regulates epithelial to mesenchymal transition of breast cancer cells during lymph node metastasis, Breast Cancer Res. Treat. 143 (2) (2014) 265–276.

[12] A. Barrallo-Gimeno, M.A. Nieto, The Snail genes as inducers of cell movement and survival: implications in development and cancer, Development 132 (14) (2005) 3151–3161.

[13] A.M. Krebs, J. Mitschke, M. Lasierra Losada, O. Schmalhofer, M. Boerries, H. Busch, M. Boettcher, D. Mougiakakos, W. Reichardt, P. Bronsert, V.G. Brunton, C. Pilarsky, T.H. Winkler, S. Brabletz, M.P. Stemmler, T. Brabletz, The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer, Nat. Cell Biol. 19 (5) (2017) 518–529.

[14] M.J. Blanco, G. Moreno-Bueno, D. Sarrio, A. Locascio, A. Cano, J. Palacios, M.A. Nieto, Correlation of Snail expression with histological grade and lymph node status in breast carcinomas, Oncogene 21 (20) (2002) 3241–3246.

[15] A. Miyoshi, Y. Kitajima, S. Kido, T. Shimonishi, S. Matsuyama, K. Kitahara, K. Miyazaki, Snail accelerates cancer invasion by upregulating MMP expression and is associated with poor prognosis of hepatocellular carcinoma, Br. J. Cancer 92 (2) (2005) 252–258.

[16] C.S. Scanlon, E.A. Van Tubergen, R.C. Inglehart, N.J. D'Silva, Biomarkers of epithelial-mesenchymal transition in squamous cell carcinoma, J. Dent. Res. 92 (2) (2013) 114–121.

[17] Z. Zhao, M.A. Rahman, Z.G. Chen, D.M. Shin, Multiple biological functions of Twist1 in various cancers, Oncotarget 8 (12) (2017) 20380–20393.

[18] S. Li, H.Y. Zhang, Z.X. Du, C. Li, M.X. An, Z.H. Zong, B.Q. Liu, H.Q. Wang, Induction of epithelial-mesenchymal transition (EMT) by Beclin 1 knockdown via posttranscriptional upregulation of ZEB1 in thyroid cancer cells, Oncotarget 7 (43) (2016) 70364–70377.

[19] Y. Chen, X. Lu, D.E. Montoya-Durango, Y.H. Liu, K.C. Dean, D.S. Darling, H.J. Kaplan, D.C. Dean, L. Gao, Y. Liu, ZEB1 regulates multiple oncogenic components involved in Uveal melanoma progression, Sci. Rep. 7 (1) (2017) 45.

[20] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (12) (2014) 550.

[21] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, Z. Zhang, GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses, Nucleic Acids Res. 45 (W1) (2017) W98–w102.

[22] D. Szklarczyk, J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N.T. Doncheva, A. Roth, P. Bork, L.J. Jensen, C. von Mering, The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, Nucleic Acids Res. 45 (D1) (2017) D362–d368.

[23] M. Zhao, Y. Liu, C. Zheng, H. Qu, dbEMT 2.0: an updated database for epithelial-mesenchymal transition genes with experimentally verified information and precalculated regulation information for cancer metastasis, J. Genet. Genomics 46 (12) (2019) 595–597.

[24] K. Colwill, S. Gräslund, A roadmap to generate renewable protein binders to the human proteome, Nat. Methods 8 (7) (2011) 551–558.

[25] P.J. Heagerty, T. Lumley, M.S. Pepe, Time-dependent ROC curves for censored survival data and a diagnostic marker, Biometrics 56 (2) (2000) 337–344.

[26] G. Yu, L.G. Wang, Y. Han, Q.Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, Omics 16 (5) (2012) 284–287.

[27] A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, J.P. Mesirov, GSEA-P: a desktop application for gene set enrichment analysis, Bioinformatics 23 (23) (2007) 3251–3253.

[28] A.M. Newman, C.L. Liu, M.R. Green, A.J. Gentles, W. Feng, Y. Xu, C.D. Hoang, M. Diehn, A.A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles, Nat. Methods 12 (5) (2015) 453–457.

[29] K. Quek, J. Li, M. Estecio, J. Zhang, J. Fujimoto, E. Roarty, L. Little, C.W. Chow, X. Song, C. Behrens, T. Chen, W.N. William, S. Swisher, J. Heymach, I. Wistuba, J. Zhang, A. Futreal, J. Zhang, DNA methylation intratumor heterogeneity in localized lung adenocarcinomas, Oncotarget 8 (13) (2017) 21994–22002.

[30] S. Rousseaux, A. Debernardi, B. Jacquiau, A.L. Vitte, A. Vesin, H. Nagy-Mignotte, D. Moro-Sibilot, P.Y. Brichon, S. Lantuejoul, P. Hainaut, J. Laffaire, A. de Reyniès, D.G. Beer, J.F. Timsit, C. Brambilla, E. Brambilla, S. Khochbin, Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers, Sci. Transl. Med. 5 (186) (2013), 186ra66.

[31] T.P. Lu, C.K. Hsiao, L.C. Lai, M.H. Tsai, C.P. Hsu, J.M. Lee, E.Y. Chuang, Identification of regulatory SNPs associated with genetic modifications in lung adenocarcinoma, BMC Res. Notes 8 (2015) 92.

[32] M. Yamauchi, R. Yamaguchi, A. Nakata, T. Kohno, M. Nagasaki, T. Shimamura, S. Imoto, A. Saito, K. Ueno, Y. Hatanaka, R. Yoshida, T. Higuchi, M. Nomura, D.G. Beer, J. Yokota, S. Miyano, N. Gotoh, Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma, PLoS One 7 (9) (2012), e43923.

[33] C. Liu, Y. Li, M. Wei, L. Zhao, Y. Yu, G. Li, Identification of a novel glycolysis-related gene signature that can predict the survival of patients with lung adenocarcinoma, Cell Cycle 18 (5) (2019) 568–579.

[34] Z. Huang, M. Shi, H. Zhou, J. Wang, H.J. Zhang, J. Shi, Prognostic signature of lung adenocarcinoma based on stem cell-related genes, Sci. Rep. 11 (1) (2021) 1687.

[35] D.G. Beer, S.L. Kardia, C.C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M. Taylor, M.D. Iannettoni, M.B. Orringer, S. Hanash, Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (8) (2002) 816–824.

[36] Y. Lu, W. Lemon, P.Y. Liu, Y. Yi, C. Morrison, P. Yang, Z. Sun, J. Szoke, W.L. Gerald, M. Watson, R. Govindan, M. You, A gene expression signature predicts survival of patients with stage I non-small cell lung cancer, PLoS Med. 3 (12) (2006) e467.

[37] H.Y. Chen, S.L. Yu, C.H. Chen, G.C. Chang, C.Y. Chen, A. Yuan, C.L. Cheng, C.H. Wang, H.J. Terng, S.F. Kao, W.K. Chan, H.N. Li, C.C. Liu, S. Singh, W.J. Chen, J.J. Chen, P.C. Yang, A five-gene signature and clinical outcome in non-small-cell lung cancer, N. Engl. J. Med. 356 (1) (2007) 11–20.

[38] Y. Qu, B. Cheng, N. Shao, Y. Jia, Q. Song, B. Tan, J. Wang, Prognostic value of immune-related genes in the tumor microenvironment of lung adenocarcinoma and lung squamous cell carcinoma, Aging (Albany NY) 12 (6) (2020) 4757–4777.

[39] C. Jochems, J. Schlom, Tumor-infiltrating immune cells and prognosis: the potential link between conventional cancer therapy and immunity, Exp. Biol. Med. 236 (5) (2011) 567–579.

[40] E. Ghorani, J.L. Reading, J.Y. Henry, M.R. de Massy, R. Rosenthal, V. Turati, K. Joshi, A.J.S. Furness, A.B. Aissa, S.K. Saini, S. Ramskov, A. Georgiou, M.W. Sunderland, Y.N.S. Wong, M.V. De Mucha, W. Day, F. Galvez-Cancino, P.D. Becker, I. Uddin, M. Ismail, T. Ronel, A. Woolston, M. Jamal-Hanjani, S. Veeriah, N.J. Birkbak, G.A. Wilson, K. Litchfield, L. Conde, J.A. Guerra-Assunção, K. Blighe, D. Biswas, R. Salgado, T. Lund, M. Al Bakir, D.A. Moore, C.T. Hiley, S. Loi, Y. Sun, Y. Yuan, K. AbdulJabbar, S. Turajilic, J. Herrero, T. Enver, S.R. Hadrup, A. Hackshaw, K.S. Peggs, N. McGranahan, B. Chain, C. Swanton, S.A. Quezada, The T cell differentiation landscape is shaped by tumour mutations in lung cancer, Nat. Can. 1 (5) (2020) 546–561.

[41] Y.K. Chae, S. Chang, T. Ko, J. Anker, S. Agte, W. Iams, W.M. Choi, K. Lee, M. Cruz, Epithelial-mesenchymal transition (EMT) signature is inversely associated with T-cell infiltration in non-small cell lung cancer (NSCLC), Sci. Rep. 8 (1) (2018) 2918.

[42] M.Q. Mahmood, E.H. Walters, S.D. Shukla, S. Weston, H.K. Muller, C. Ward, S.S. Sohal, β-catenin, Twist and Snail: transcriptional regulation of EMT in smokers and COPD, and relation to airflow obstruction, Sci. Rep. 7 (1) (2017) 10832.

[43] M.Q. Mahmood, C. Ward, H.K. Muller, S.S. Sohal, E.H. Walters, Epithelial mesenchymal transition (EMT) and non-small cell lung cancer (NSCLC): a mutual association with airway disease, Med. Oncol. 34 (3) (2017) 45.

[44] E.H. Walters, S.D. Shukla, M.Q. Mahmood, C. Ward, Fully integrating pathophysiological insights in COPD: an updated working disease model to broaden therapeutic vision, Eur. Respir. Rev. 30 (160) (2021).

[45] M.S. Eapen, P.M. Hansbro, K. McAlinden, R.Y. Kim, C. Ward, T.L. Hackett, E.H. Walters, S.S. Sohal, Abnormal M1/M2 macrophage phenotype profiles in the small airway wall and lumen in smokers and chronic obstructive pulmonary disease (COPD), Sci. Rep. 7 (1) (2017) 13392.