Identification of EMT-Related lncRNAs as a Potential Prognostic Biomarker and Therapeutic Targets for Pancreatic Adenocarcinoma

Yanyao Deng
First Hospital of Changsha
Hai Hu
Central South University Third Xiangya Hospital
Le Xiao
First Hospital of Changsha
Ting Cai
Central South University Third Xiangya Hospital https://orcid.org/0000-0002-8910-3289
Wenzhe Gao
Central South University Third Xiangya Hospital
Hongwei Zhu
Central South University Third Xiangya Hospital
Shuai Wang (2201171015@csu.edu.cn)
Central South University Third Xiangya Hospital
Jixing Liu
Second Affiliated Hospital of Hainan medical University

Primary research

Keywords: pancreatic adenocarcinoma, epithelial-mesenchymal transition, long non-coding RNA, prognostic signature, ceRNA network

Posted Date: June 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-613873/v1

Version of Record: A version of this preprint was published at Journal of Oncology on April 11th, 2022.
See the published version at https://doi.org/10.1155/2022/8259951.

Abstract

Background: Epithelial-Mesenchymal Transition (EMT) can promote carcinoma progression by multiple mechanisms, many studies demonstrated the invasiveness of pancreatic adenocarcinoma (PAAD) associated with the EMT, but how it acts in a lncRNA dependent manner is unclear.

Methods: We investigated 146 PAAD samples from The Cancer Genome Atlas (TCGA) and 92 samples from the International Cancer Genome Consortium (ICGC). Gene set variation analysis (GSVA) and weighted correlation network analysis (WGCNA) were applied to explore the EMT related long non-coding RNAs (EMTlnc). Univariate Cox regression analysis was performed to screen their prognostic roles in PAAD patients. Least absolute shrinkage and selection operator (LASSO) Cox regression was used to establish an EMT-related lncRNA prognostic signature (EMT-LPS). We also established a competing endogenous RNA (ceRNA) network.

Results: 33 prognostic EMTlnc were identified as prognostic lncRNAs and an EMT-LPS were established. We divided the patients into low- and high-risk subgroups according to corresponding risk scores. The EMT-LPS showed a powerful prognostic predicting ability in stratification analysis. Principal component analysis (PCA) showed the low- and high-risk subgroups had distinct EMT status. Enrichment analysis indicated malignancy correlated biological processes, pathways and hallmarks were more common in the high-risk subgroup. Moreover, we constructed a nomogram that had a strong ability to forecast the overall survival (OS) of the PAAD patients in both datasets.

Conclusion: EMT-LPS are important factors in the carcinoma progression of PAAD and may help in decision making regarding the choice of prognosis assessment and provide us clues to design the new drugs for PAAD.

Introduction

Pancreatic adenocarcinoma (PAAD) is a neoplastic disease with extremely high malignancy and poor prognosis, the cure rate for PAAD is only 9%, moved to the third leading cause of cancer death, if untreated, the median survival of patients with metastatic disease is only 3 months[1]. Thus, searching for therapeutic targets for treating PAAD is urgent.

As a class of RNA molecules longer than 200 nucleotides in length, and without a protein-coding function, long non-coding RNAs (lncRNAs) are involved in a wide variety of cellular processes. Increasing evidence suggest that the dysregulation of lncRNA expression is implicated in multiple types of cancer[2], involved in cancer-related cellular processes such as proliferation, apoptosis, migration and invasion through regulation of gene expression[3–5]. In addition, lncRNAs can also serve as diagnostic or prognostic markers of various types of cancers, for instance, in hepatocellular carcinoma and prostate cancer[6–8].

The Epithelial-Mesenchymal Transition (EMT) plays prominent roles in the formation of the body plan and in the differentiation of multiple tissues and organs. It's a morphologic cellular program simply

defined as the phenotypic transition from an epithelial to a mesenchymal state. Study reveals EMT can adversely cause organ fibrosis and promote carcinoma progression by multiple mechanisms[9], and it's regulated by a complex network involving epigenetic modifications, transcriptional control, alternative splicing, protein stability, and subcellular localization[10–12]. Although EMT processes are documented in many in vitro cancer cell models, the significance of EMT during cancer progression and even its relevance in human cancer tissues are still a matter of controversy. A multitude of studies have demonstrated that the invasiveness of pancreatic adenocarcinoma associated with the EMT[13–16], but how it acts in a lncRNA-dependent way during PAAD progression is still unclear.

In this study, based on The Cancer Genome Atlas (TCGA) dataset (n = 146) and the International Cancer Genome Consortium (ICGC) dataset (n = 92), we identified the prognostic significance of EMTlnc by bioinformatic and statistical analysis of data from patients with PAAD. Our results showed that 33 EMTlnc had prognostic value in both TCGA and ICGC PAAD patients. Furthermore, we constructed an EMT-related lncRNA prognostic signature (EMT-LPS) based on the ability of 33 EMTlnc to predict the OS of PAAD patients. In the meanwhile, PC patients in low- and high-risk subgroups (categorized based on the EMT-LPS) had different prognosis and tumor hallmarks were more common in the high-risk subgroup. Furthermore, an accurate nomogram was constructed to predict OS in patients with PAAD and a ceRNA network was built to search the target miRNAs and mRNAs of these EMT-related prognostic lncRNAs.

Materials And Methods

Raw Data Acquisition

For training set, mRNA expression files [Fragments Per Kilobase of transcript per Million mapped reads (FPKM) normalized] and the corresponding clinicopathological data were acquired from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). For validation set, RNA-seq profile and related clinicopathological data were downloaded from UCSC Xena Database (http://xena.ucsc.edu/) and the International Cancer Genome Consortium Data Portal (https://dcc.icgc.org/). PAAD patients with missing OS values or OS < 30 days were excluded in order to reduce statistical bias in our analysis. Finally, 146 PAAD patients from TCGA were selected to construct the EMT-LPS and 92 PAAD patients from ICGC were included to test the EMT-LPS. Then, the FPKM data was underwent a log2 transformation. The gene annotation file "gencode.v22.annotation," which was downloaded from the TCGA database, was utilized to transform ENSEMBL ID to GENE SYMBOL. According to the GENCODE website (https://www.gencodegenes.org/human/), 9 types of transcripts (3prime_overlapping_ncRNA, antisense, bidirectional_promoter_lncRNA, lincRNA,macro_lncRNA, non_coding, processed_transcript, sense_intronic, sense_overlapping) were defined as lncRNAs. The ENSEMBL ID with the max average expression level to represent the expression level of this gene when there were multiple ENSEMBL ID annotated to the same gene symbol. Ultimately, 14805 lncRNAs and 19712 mRNAs were identified in the TCGA cohort, while 1440 lncRNAs and 15146 mRNAs were identified in the ICGC cohort.

EMTlnc Acquisition By GSVA and WGCNA

The HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION pathway, which contained 200 EMT-related gene, was obtained from package “msigdbr”[17, 18]. Then gene set variation analysis (GSVA), was used to calculated the values of EMT pathway of each PAAD patients in TCGA cohort. Weighted Gene Co-expression Network Analysis (WGCNA) was utilized to find significantly correlated lncRNAs to combine lncRNA modules and to search the relationship between each module and the EMT values estimated by GSVA of the 146 PAAD patients in TCGA cohort. Here, the power of β= 3 was chosen as the soft threshold to ensure a scale-free network. The dynamic tree cutting method was utilized to cluster the lncRNAs in layers, using 50 as a minimum size cutoff, and the cut height = 0.3 was applied to merge highly similar modules. Different lncRNA modules were labeled with different colors, and the gray module contained lncRNAs that cannot be merged. Then, Pearson correlation analysis was applied to evaluate the correlation between lncRNAs in each module and EMT values. Finally, the lncRNAs in model with abs(cor)>0.5&p.value<0.05 was defined as EMTlnc.

Establishment and Verification of EMT-LPS

We used the Univariate Cox to choose prognostic EMTlnc in both TCGA cohort and ICGC cohort based on EMTlnc. Through taking intersection, 33 lncRNAs with p<0.05 were defined as shared prognostic EMTlnc. The LASSO Cox regression was used to selected the most useful prognostic lncRNAs and construct EMT-LPS involved 11 EMTlnc by using package “glmnet”[19] in R. Here, The “10-fold cross-validation” approach was used to facilitate parameter selection. The risk score of TCGA cohort and ICGC cohort patients were calculated as following formula :

$$\text{risk score} = \sum_{i=1}^{n} Coef\,(i) * x_i$$

(Coef(i) was the estimated regression coefficient dated from LASSO Cox regression analysis, and χi was the expression value of each selected EMTlnc). The median risk score in the all patients was used as the cutoff point that split PAAD patients into a high-risk group and a low-risk group both in TCGA cohort and ICGC cohort. Then Log-rank testing method was used to compare the differences of Overall survival outcomes between the high- and low-risk groups via Kaplan-Meier survival analysis.The receiver operating characteristic (ROC) curve analysis in the “survival ROC” package[20] was applied to examine the accuracy of the identified EMT-LPS.

Stratification Analysis

The whole TCGA cohort was stratified by age [≥60 years (n = 95) or <60 years (n = 43)], Gender[Female(n=60) or Male(n=78)], Grade[G1+G2(n=93), G3+G4(n=45)] and TNM stages [T1+T2(n=4), T3+T4(n=134), N0(n=38), N1(n=100)]. While the ICGC cohort was divided into different subgroup(age[≥60 years (n = 67) or <60 years (n = 21)], Gender[Female(n=47) or Male(n=41)], TNM stages[T1+T2(n=3), T3+T4(n=85), N0(n=30), N1(n=58)]). The Willcoxon rank sum test was used to

compare the risk score of different stratification cohort with package “ggpubr”(https://cran.r-project.org/web/packages/ggpubr). The formula of risk score acquired in the TCGA cohort was used to calculate the risk score of each PAAD patient in each stratification cohort, followed by grouping them into high- and low-risk groups. Log-rank testing method was used to compare the differences of Overall survival outcomes between the high- and low-risk groups via Kaplan-Meier survival analysis.

Principal Component Analysis(PCA) and Nomogram Construction

Based on the expression of 200 EMT-related genes of HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION pathway, the PCA was used to assess the differences between the low- and high-risk subgroups. The R package “rms”[21] was used to construct the nomogram to assess the 1- ,2- and 3-year survival possibility for PAAD patients both in TCGA and ICGC cohort. Calibration curve of the nomogram was generated to evaluate the consistency between its predicted values and the actual observed values by “rms” package.

Construction of the ceRNA Network

Using the TCGA cohort, the Differentially Expressed Genes (DEGs) between high-risk subgroup and low-risk subgroup were identified based on the threshold value of $|\log2(\text{Fold change})| > 1$ and $p < 0.05$ using Willcoxon rank sum test. Perl programming language was used to perform the prediction analysis of the target miRNAs of the 11 EMTlnc in the miRcode database(http://www.mircode.org/) and then the shared target mRNAs of these miRNAs found in the miRTarBase(http://mirtarbase.mbc.nctu.edu.tw/php/index.php), miRDB(http://mirdb.org/), and TargetScan database(http://www.targetscan.org/) were intersected with DEGs to obtain differential expressed target mRNA. Finally, The ceRNA network was plotted using the software of “Cytoscape”[20].

Enrichment Analysis

Respectively, the differential expression genes (DEGs) between low and high-subgroups and the differential expressed target mRNAs in the ceRNA network were then inputted into the “Metascape” website (http://metascape.org/)for functional and pathway enrichment analysis, which involved Canonical Pathways, Reactome Gene Sets, Gene Ontology (GO) Biological Processes and Kyoto Encyclopedia of Genes and Genomes Pathway (KEGG pathway). Additionally, we used GSEA software(http://software.broadinstitute.org/gsea/index.jsp) to investigate the tumor hallmarks that were more common in the high-risk subgroup compared with the low-risk subgroup.

Statistics Analysis

All dataset statistical analysis performed in this study were carried in R programming language(4.0.0). Kaplan-Meier, log rank tests and Univariate Cox regression were used to perform survival analysis based the expression of EMT related lncRNAs included in EMT-LPS. Univariate and multivariate Cox regression analyses were utilized to evaluate the independent prognostic value of the EMT-LPS regarding Overall survival outcomes.

Results

Identification of EMTlnc in PAAD Patients

Firstly, through the gene annotation file from the TCGA database, we identified 14805 lncRNAs in the TCGA dataset and 1440 lncRNAs in the ICGC dataset for the next analysis. Then we conducted GSVA analysis on EMT path-related genes in TCGA PAAD samples and obtained GSVA variation values. After applying WGCNA on the database of PAAD patients in TCGA, coexpression network by WGCNA analysis shown that the EMT related lncRNA model in PAAD samples were grouped into 48 models, contained MElightcyan and MEturquoise model with abs(cor)>0.5 and p.value <0.05, which were both included into further analysis in this research. The lncRNAs in those two models were defined as EMTlnc.

Combined with the prognostic information, univariate Cox regression was then implemented to screen EMT-related prognostic lncRNAs from the EMTlnc in both the TCGA and ICGC datasets ($p < 0.05$), respectively. Finally, we found that 33 EMTlnc were significantly correlated with the OS of PAAD patients in both two datasets through taking intersection. The work flow was shown in Figure 1A and the WGCNA in PAAD samples was shown in Figure 1B, the correlation between membership in the lightcyan module and membership in the EMT pathway of the eigengenes in lightcyan module by Pearson correlation analysis, and correlation between membership in the turquoise module and membership in the EMT pathway of the eigengenes in turquoise module by Pearson correlation analysis was shown in Figure 1C, Figure 1D respectively. The results of univariate Cox analysis of the 33 EMTlnc were shown in Table 1.

Table 1 | The thirty-three EMT-related prognostic lncRNAs

EMT related lncRNAs	TCGA				ICGC			
	HR	HR.95L	HR.95H	p-value	HR	HR.95L	HR.95H	p-value
AC017002.1	2.9985	1.0379	8.6629	4.25E-02	1.4835	1.0744	2.0483	1.66E-02
AC093850.2	1.2967	1.0605	1.5855	1.13E-02	1.1653	1.0152	1.3376	2.96E-02
LINC00152	1.9583	1.3243	2.8959	7.60E-04	1.9065	1.2703	2.8615	1.84E-03
LINC01116	1.7473	1.1670	2.6161	6.73E-03	1.5378	1.1805	2.0033	1.42E-03
MIR4435-1HG	2.3464	1.4175	3.8839	9.10E-04	2.3326	1.4546	3.7404	4.39E-04
RP11-274H2.3	5.8172	1.4789	22.8821	1.17E-02	1.3529	1.0352	1.7681	2.69E-02
RP11-400N13.3	1.6652	1.3155	2.1079	2.24E-05	1.2304	1.0762	1.4068	2.41E-03
RP11-417E7.1	1.5019	1.0033	2.2482	4.82E-02	1.3728	1.1670	1.6149	1.31E-04
RP11-554I8.2	1.6514	1.2114	2.2512	1.51E-03	1.1728	1.0457	1.3153	6.46E-03
UCA1	1.4171	1.2289	1.6341	1.62E-06	1.3475	1.1798	1.5391	1.10E-05
AC009506.1	0.2309	0.0902	0.5911	2.24E-03	0.6498	0.4526	0.9330	1.95E-02
AC096772.6	0.4648	0.2784	0.7762	3.40E-03	0.5803	0.3373	0.9985	4.93E-02
DANCR	0.3418	0.2100	0.5562	1.56E-05	0.6598	0.4665	0.9331	1.87E-02
FLJ37035	0.0148	0.0007	0.3109	6.69E-03	0.7340	0.5878	0.9165	6.35E-03
GS1-358P8.4	0.3759	0.2246	0.6289	1.95E-04	0.3935	0.2300	0.6734	6.67E-04
HNF1A-AS1	0.5779	0.4157	0.8034	1.11E-03	0.7402	0.5912	0.9267	8.70E-03
LINC00261	0.6925	0.5461	0.8780	2.41E-	0.8859	0.8096	0.9693	8.33E-

				03				03
LINC01128	0.2883	0.1461	0.5691	3.37E-04	0.3458	0.1789	0.6681	1.58E-03
PART1	0.0495	0.0026	0.9468	4.59E-02	0.8745	0.7691	0.9944	4.09E-02
PP7080	0.6783	0.4953	0.9288	1.55E-02	0.7559	0.5729	0.9974	4.79E-02
PRKAG2-AS1	0.6626	0.4526	0.9702	3.44E-02	0.7983	0.6602	0.9652	2.00E-02
RP11-16P6.1	0.0868	0.0184	0.4096	2.02E-03	0.6257	0.3945	0.9926	4.64E-02
RP11-226L15.5	0.4160	0.2031	0.8521	1.65E-02	0.5727	0.3337	0.9830	4.31E-02
RP11-244O19.1	0.1963	0.0790	0.4874	4.50E-04	0.7268	0.5435	0.9720	3.14E-02
RP11-384L8.1	0.5399	0.3382	0.8617	9.78E-03	0.6399	0.5037	0.8129	2.55E-04
RP11-700H6.1	0.0668	0.0087	0.5136	9.31E-03	0.8289	0.7284	0.9433	4.45E-03
RP1-193H18.2	0.3373	0.1930	0.5895	1.36E-04	0.7516	0.6052	0.9334	9.77E-03
RP11-968O1.5	0.5038	0.2815	0.9017	2.10E-02	0.7650	0.5933	0.9863	3.88E-02
RP5-1033H22.2	0.6204	0.3941	0.9766	3.92E-02	0.8018	0.6884	0.9338	4.50E-03
RP5-1085F17.3	0.3766	0.2402	0.5904	2.07E-05	0.5024	0.2875	0.8779	1.56E-02
RP5-894A10.2	0.5887	0.3842	0.9020	1.49E-02	0.6040	0.4265	0.8554	4.52E-03
SLC25A25-AS1	0.6439	0.4690	0.8839	6.45E-03	0.7420	0.5867	0.9383	1.27E-02
XXbac-B135H6.15	0.4622	0.2895	0.7377	1.22E-03	0.6998	0.5027	0.9740	3.44E-02

Color shaded lncRNAs were risky lncRNAs and others were protective lncRNAs.

Construction of the EMT-LPS in the TCGA Dataset and Validation of the EMT-LPS in the ICGC Dataset

To build the EMT-LPS for forecasting the OS of PAAD patients, we performed a LASSO Cox analysis on the basis of the 33 EMT-related prognostic lncRNAs in the TCGA cohort and it generated the EMT-LPS which contains 11 EMTlnc and coefficient of each (Figures 2A,B). The EMT-LPS involved 11 lncRNAs and, for each patient in the TCGA dataset, a risk score was calculated based on the coefficient for each lncRNA (Figure 2C). Patients in the TCGA cohort were divided into low and high-risk subgroups based on the median value of risk scores. Kaplan-Meier survival curves depicted that PAAD patients with higher risk scores had worse clinical outcomes (lower OS rates and a shorter OS time) (Figure 2D). Risk score and survival status distributions are plotted in Figure 2F. And the ROC curves demonstrated that EMT-LPS harbored a promising ability to predict OS in the TCGA cohort (1-year AUC = 0.81, 2-year AUC = 0.86, 3-year AUC = 0.9; Figure 2H).

To validate the prognostic ability of EMT-LPS, we calculated risk scores for patients in the ICGC cohort using the same formula. PAAD patients in the ICGC dataset were assigned to low- and high-risk groups based on the median risk score. The results were consistent with the findings in the TCGA dataset: PAAD patients with higher risk scores had lower OS rates and a shorter OS time in the ICGC dataset (Figure 2E). Risk score and survival status distributions are shown in Figure 2G and it showed that patients with higher risk scores had shorter overall survival time and dead status. The ROC analysis also indicated that EMT-LPS had a strong prognostic value for PAAD patients in the ICGC dataset (1-year AUC = 0.8, 2-year AUC = 0.85, 3-year AUC = 0.89; Figure 2I). These results showed that the EMT-LPS had a robust and stable OS-predictive ability.

Prognostic Analysis of the Eleven EMTlnc

Eleven EMTlnc were included in the EMT-LPS and univariate Cox regression analysis was used to evaluate their prognostic roles. The forest plot shows that PP7080, xxbac-B135H6.15, RP5-1085F17.3, DANCR, AC096772.6, LINC01128 are protective factors with HR (Hazard ratio) < 1, while LINC01116, UCA1, RP11-400N13.3 are risk factors with HR < 1 in EMT patients (Figure 3A). The heatmap (Figure 3B) shows that RP11-55418.2, RP11-400N13.3, UCA1, LINC00152 and LINC01116 expression increased with increasing risk score, whereas the expression of the AC096772.6, xxbac-B135H6.15, LINC01128, RP5-1085F17.3, PP7080 and DANCR decreased with increasing risk score. Their expression levels were also related to the clinicopathological features of PAAD, including N_stage, T_stage, gender, age and WHO grade (Figure 3B). The Kaplan-Meier survival curves confirmed that higher expression of AC096772.6, xxbac-B135H6.15, LINC01128, RP5-1085F17.3, PP7080 and lower expression of RP11-55418.2, RP11-400N13.3, UCA1, LINC00152, DANCR and LINC01116 were associated with better OS in the TCGA dataset (Figures 3C-M). In the ICGC dataset, the heatmap (Supplementary Figure S2C) also showed that RP11-55418.2, RP11-400N13.3, UCA1, LINC00152 and LINC01116 expression increased with increasing risk score, whereas the expression of the AC096772.6, xxbac-B135H6.15, LINC01128, RP5-1085F17.3, PP7080 and DANCR decreased with increasing risk score. Their expression levels were related to the clinicopathological features of PAAD, such as N_stage, T_stage, gender, and age.

Stratification Analysis of the EMT-LPS

We attempted to identify whether clinicopathological features were associated with the risk score. In TCGA dataset, the results revealed that PAAD patients with WHO grade III + □ had higher risk scores, while the risk score was not associated with age, gender, N_stage and T_stage (Figures 4A-E). To better assess the prognostic ability of the EMT-LPS, we performed a stratification analysis to confirm whether it retains its ability to predict OS in various subgroups. In contrast with patients with lower risk, higher risk PAAD patients had worse OS in age ≥ 60 (Figures 4F, G). Likewise, we confirmed that EMT-LPS retained its ability to predict OS for patients by female or male (Figures 4H,I) , patients with grade □+□ or grade □+□, and patients with TNM stage N0, N1 or T3+T4(Figures 4J-N). In ICGC dataset, the results showed that PAAD patients with N1_stage, T3_stage and T4_stage had higher risk scores (Supplementary Figure S2D-G) . These data indicated that it could be a potential predictor for PAAD patients.

Principal Component Analysis

Based on the expression value of the 200 EMT-related genes, principal component analysis (PCA) was performed to assess the differences between the low- and high-risk subgroups (Supplementary Figure S2A,B). The results showed that the low and high-risk patients in both the TCGA and ICGC datasets were distributed in distinct directions. These results may suggest that differential EMT statuses exist in different risk subgroups.

Pathway and Process Enrichment Analysis and Gene Set Enrichment Analysis (GSEA)

For investigating the potential biological process and pathway involving in the molecular heterogeneity between the low- and high-risk subgroups, we identified 710 differential expression genes (DEGs) [|log2 (fold change)| > 1 and $p < 0.05$] between the low- and high-risk subgroups in the TCGA cohort. These DEGs were primarily enriched in these terms: chemical synaptic transmission, Neuronal System, regulation of membrane potential, behavior, plasma membrane bounded cell projection morphogenesis, GABAergic synapse(Figures 5A). Gene set enrichment analysis revealed that two tumor hallmarks were enriched in the high-risk subgroup, the interferon alpha response and the interferon gamma response (Figure 5B,C) . These results may give us some insights into the cellular biological effects related to the EMT-LPS.

EMT-LPS Was an Independent Prognostic Factor for PAAD Patients

We used univariate and multivariate Cox analyses to assess whether the EMT-LPS was an independent prognostic factor for patients with PAAD. Based on the data of PAAD patients in the TCGA dataset, univariate Cox analysis indicated that EMT-LPS was remarkably associated with OS [Hazard Ratio (HR): 3.832, 95% CI: 2.435-6.030, p < 0.001; Figure 6A] and multivariate Cox analysis further showed that EMT-LPS was an independent predictor of OS (HR: 3.573, 95% CI: 2.248-5.681, p < 0.001; Figure 6A). The conclusion was validated in the ICGC dataset, which confirmed that EMT-LPS was an independent predictor of OS for PAAD patients in the ICGC validation dataset (univariate: HR: 2.667, 95% CI: 1.923-3.697, p < 0.001; multivariate: HR: 2.787 95% CI: 1.945-3.94, p < 0.001; Figure 6B). These results indicated that our EMT-LPS, as an independent prognostic indicator, might be useful for clinical prognosis evaluation.

Construction and Validation of the EMT-LPS Based Nomogram

To create a clinically applicable quantitative tool to predict the OS of PAAD patients, we established a nomogram using the risk status (based on EMT-LPS), gender, age, T_stage, N_stage and WHO grade in the TCGA dataset and it was also tested in the ICGC dataset (Figure 6C). Calibration plots showed that the observed vs. predicted rates of 1-, 2- and 3-year OS showed perfect concordance in the TCGA (Figures 7A-C) and ICGC cohorts (Supplementary Figure S1B-D). Then time-dependent ROC curves were used to assess the prognostic predictive ability of the nomogram and other predictors (risk score, gender, age, T_stage, N_stage and WHO grade) in the TCGA (Figures 7D-F) and (Supplementary Figure S1E-G) and the results revealed that, compared with the other predictors, the nomogram had excellent accuracy regarding 1-, 2- and 3- year OS (AUC = 0.79, 0.83, and 0.86 in the TCGA, and 0.8, 0.84, and 0.89 in the ICGC , respectively). These data indicated that the nomogram has a robust and stable ability to predictive the OS for PAAD patients.

Construction of the ceRNA Network and Functional Enrichment Analysis

To further understand how the EMTlnc regulate mRNA expression by sponging miRNAs in PAAD, we constructed a ceRNA network based on the EMTlnc. Two of twelve lncRNAs were extracted from the miRcode database and thirteen pairs of interaction between the two lncRNAs and thirty-four miRNAs were identified. Then we used three databases (miRTarBase, miRDB, and TargetScan) to search target mRNAs based on the thirty-four miRNAs and totally 1539 mRNAs were identified in all the three databases. What's more, these target mRNA were intersected with DEGs to botain differential expressed target mRNA. Ultimately, two lncRNAs, twelve miRNAs and thirteen mRNAs were included in our ceRNA network (Figure 8A). Furthermore, the 1539 target mRNAs were used to implemented functional enrichment analysis in the Metascape online tool and we found that these genes were enriched in vasculature development, pathway in cancer, regulation of cellular response to stress, Wnt signaling pathway, tissue

morphogenesis, insulin signaling, regulation of cellular protein localization, response to growth factor, nagative regulation of cell differentiation (Figures 8B-D). These data may provide us some clues for finding the potential functions of these EMTlnc in PAADs.

Disscussion

A total of 268 PAAD patients from the TCGA and ICGC datasets were included in our study to exploit the prognostic significance of EMTlnc. Thirty-three EMTlnc were proven to have prognostic value in both the TCGA and ICGC datasets, and eleven of them were used to establish an EMT-LPS for predicting the OS of PAAD patients. Based on the median risk score, PAAD patients were divided into the low- and high-risk subgroups, and the high-risk group had worse clinical outcomes and enrichment of tumor hallmarks and certain malignant related pathways. Multivariate Cox regression analysis showed that EMT-LPS was an independent risk factor for OS.

Furthermore, combining EMT-LPS with gender, age and World Health Organization Grade, T_stage and N_stage, we set up a nomogram, and it had a strong ability to predict OS in PAAD patients in the TCGA and ICGC datasets. Finally, A ceRNA network include two EMTlnc, twelve miRNAs and thirteen mRNAs were established for observing the latent functions of these EMTlnc.

Accumulating evidence shows that lncRNAs orchestrate multiple cellular processes by modulating EMT in various cell types. MALAT1 and lnc-ATB can motivate EMT during the period of silica-induced pulmonary fibrosis by competitively binding miR-503 and miR-200c, respectively[22, 23]. The lncRNA ROR regulates multiple signaling pathways include breast, bladder, and nasopharyngeal EMT[24-26]. Furthermore , hypoxia reinforces exosome-mediated reciprocal movement of the lncRNA UCA1 into bladder cancer cells, which accelerates cancer growth and progression by inducing EMT[27].

Study had revealed that EMT had impact on cancer invasion and progression, and lncRNAs may act as ceRNAs, targeting EMT regulators so as to influence tumor aggressive progression. Liu et al[28] found TGFBI and FN1 regulates EMT as a ceRNA for miR-21 and miR-200c, respectively. And the abundance of ceRNA can determines the reversibility of EMT. We implemented functional enrichment analysis based on the ceRNAs and found that genes were enriched in vasculature development, pathway in cancer, regulation of cellular response to stress, Wnt signaling pathway and some other pathways. Combining all the evidence together, we believe that EMT is targeted at lncRNAs, and we had better with more attention to the interactions and functions of lncRNAs and EMT thus identifing potential prognostic markers or therapeutic targets of cancers.

We identified 33 EMT-related prognostic lncRNAs from 268 PAAD patients, and eleven of them were included in the EMT-LPS. Yang et al uncovered that RP11-400N13.3 reacts as an oncogenic lncRNA in colorectal cancer, and can accelerate colorectal cancer progression via modulating the miR-4722-3p/P2RY8 axis[29]. LINC00152 was firstly found over expressed in gastric cancer and served as an oncogene in gliomas, liver, lung, and colorectal cancer[30-32]. It may promote the cell proliferation and invasion capability of these cancer cells by regulating GFR, EZH2, miR-16, and miR-139-5p[33-37] . Wang

et al confirmed that LINC01116 was expressed at markedly higher level in glioma and relevant to the clinicopathological characteristics and survival of glioma patients, LINC01116 promoted tumor proliferation and neutrophil recruitment by regulating IL-1β in glioma[38]. Li et al discovered that LINC01128 resisted acute myeloid leukemia through regulating miR-4260/NR3C2[39]. Tang et al reported that DANCR was downregulated by miR-135a through regulating of downstream protein NLRP3 in pancreatic cancer[40], which is consistent with our results.

Several of the eleven lncRNAs were reported to be associated with cancer progression, but there have been few reports regarding PAAD, and reports on how the lncRNAs interact with EMT-related genes have been even rarer. Hence, we hope that our results contribute to identify the prognostic lncRNAs that EMT regulators might target, then providing insights into their potential roles in PAAD tumorigenesis and progression.

In summary this study included two PAAD datasets, the TCGA and ICGC datasets, and our results were derived and validated using them, but there were several limitations in our study. More independent PAAD cohorts should be used to validate the identified prognostic EMTlnc. Additionally, researches aimed at revealing the specific mechanisms and genes that regulate EMT in PAAD are advocated, which will enable the design of new drugs or strategies for managing PAAD.

Conclusions

EMT-LPS are important factors in the carcinoma progression of PAAD and may help in decision making regarding the choice of prognosis assessment and provide us clues to design the new drugs for PAAD.

Declarations

Acknowledgements

None.

Authors' contributions

Conceptualization, Hongwei Zhu and Jixing Liu; methodology, Shuai Wang and Hai Hu; software, Shuai Wang and Hai Hu; validation, Yanyao Deng, Hai Hu. and Le Xiao; formal analysis, Shuai Wang, Hai Hu and Ting Cai; resources, Jixing Liu; data curation, Ting Cai and Wenzhe Gao; writing—original draft preparation, Yanyao Deng and Hai Hu; writing—review and editing, Le Xiao and Ting Cai; visualization, Shuai Wang; supervision, Wenzhe Gao; project administration, Hongwei Zhu and Jixing Liu; funding acquisition, Yanyao Deng and Hongwei Zhu. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Hunan Provincial Science and Technology Plan Project, grant number 2019JJ80066, Scientific Research Project of Health and Family Planning Commission of Hunan Province of China, grant number B20-17202, National Natural Science Foundation for Young Scholars of China, , grant number 82000614, Natural Science Foundation of Hunan Province, China, grant number 2020JJ5876 and Science and Technology Project of Changsha, Hunan, China, grant number kq2004146.

Availability of data and materials

All datasets generated for this study are included in the article.

Ethics approval and consent to participate

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication

All authors read the final manuscript and agreed to publish it.

Competing interests

The authors declare no conflict of interest.

References

1. Tempero MA. NCCN Guidelines Updates: Pancreatic Cancer. J Natl Compr Canc Netw. 2019;17(5.5):603–5.
2. Shi X, Sun M, Liu H, Yao Y, Song Y. Long non-coding RNAs: a new frontier in the study of human diseases. CANCER LETT. 2013;339(2):159–66.
3. Liu Q, Huang J, Zhou N, Zhang Z, Zhang A, Lu Z, Wu F, Mo YY. LncRNA loc285194 is a p53-regulated tumor suppressor. NUCLEIC ACIDS RES. 2013;41(9):4976–87.
4. Yao N, Fu Y, Chen L, Liu Z, He J, Zhu Y, Xia T, Wang S: **Long non-coding RNA NONHSAT101069 promotes epirubicin resistance, migration, and invasion of breast cancer cells through NONHSAT101069/miR-129-5p/Twist1 axis**. *ONCOGENE* 2019, **38**(47):7216–7233.
5. Zheng J, Zhang H, Ma R, Liu H, Gao P. Long non-coding RNA KRT19P3 suppresses proliferation and metastasis through COPS7A-mediated NF-kappaB pathway in gastric cancer. ONCOGENE. 2019;38(45):7073–88.
6. Yuan W, Sun Y, Liu L, Zhou B, Wang S, Gu D. Circulating LncRNAs Serve as Diagnostic Markers for Hepatocellular Carcinoma. CELL PHYSIOL BIOCHEM. 2017;44(1):125–32.
7. Arriaga-Canon C, De La Rosa-Velazquez IA, Gonzalez-Barrios R, Montiel-Manriquez R, Oliva-Rico D, Jimenez-Trejo F, Cortes-Gonzalez C, Herrera LA. The use of long non-coding RNAs as prognostic biomarkers and therapeutic targets in prostate cancer. Oncotarget. 2018;9(29):20872–90.

8. Wang P, Ning S, Zhang Y, Li R, Ye J, Zhao Z, Zhi H, Wang T, Guo Z, Li X. Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer. NUCLEIC ACIDS RES. 2015;43(7):3478–89.
9. Ye X, Weinberg RA. Epithelial-Mesenchymal Plasticity: A Central Regulator of Cancer Progression. TRENDS CELL BIOL. 2015;25(11):675–86.
10. Diaz-Lopez A, Moreno-Bueno G, Cano A. Role of microRNA in epithelial to mesenchymal transition and metastasis and clinical perspectives. CANCER MANAG RES. 2014;6:205–16.
11. Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial-mesenchymal transition. Nat Rev Mol Cell Biol. 2014;15(3):178–96.
12. Warzecha CC, Carstens RP. Complex changes in alternative pre-mRNA splicing play a central role in the epithelial-to-mesenchymal transition (EMT). SEMIN CANCER BIOL. 2012;22(5–6):417–27.
13. Rhim AD, Mirek ET, Aiello NM, Maitra A, Bailey JM, McAllister F, Reichert M, Beatty GL, Rustgi AK, Vonderheide RH, et al. EMT and dissemination precede pancreatic tumor formation. CELL. 2012;148(1–2):349–61.
14. Jolly MK, Tripathi SC, Jia D, Mooney SM, Celiktas M, Hanash SM, Mani SA, Pienta KJ, Ben-Jacob E, Levine H. Stability of the hybrid epithelial/mesenchymal phenotype. Oncotarget. 2016;7(19):27067–84.
15. Alderton GK. Metastasis: Epithelial to mesenchymal and back again. NAT REV CANCER. 2013;13(1):3.
16. Li W, Kang Y. Probing the Fifty Shades of EMT in Metastasis. Trends Cancer. 2016;2(2):65–7.
17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
18. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. CELL SYST. 2015;1(6):417–25.
19. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J STAT SOFTW. 2010;33(1):1–22.
20. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. BIOMETRICS. 2005;61(1):92–105.
21. Chen S, Ma W, Cao F, Shen L, Qi H, Xie L, Wu Y, Fan W. Hepatocellular Carcinoma Within the Milan Criteria: A Novel Inflammation-Based Nomogram System to Assess the Outcomes of Ablation. FRONT ONCOL. 2020;10:1764.
22. Liu Y, Li Y, Xu Q, Yao W, Wu Q, Yuan J, Yan W, Xu T, Ji X, Ni C. Long non-coding RNA-ATB promotes EMT during silica-induced pulmonary fibrosis by competitively binding miR-200c. Biochim Biophys Acta Mol Basis Dis. 2018;1864(2):420–31.
23. Yan W, Wu Q, Yao W, Li Y, Liu Y, Yuan J, Han R, Yang J, Ji X, Ni C. MiR-503 modulates epithelial-mesenchymal transition in silica-induced pulmonary fibrosis by targeting PI3K p85 and is sponged

by lncRNA MALAT1. Sci Rep. 2017;7(1):11313.

24. Hou P, Zhao Y, Li Z, Yao R, Ma M, Gao Y, Zhao L, Zhang Y, Huang B, Lu J. LincRNA-ROR induces epithelial-to-mesenchymal transition and contributes to breast cancer tumorigenesis and metastasis. CELL DEATH DIS. 2014;5:e1287.
25. Chen Y, Peng Y, Xu Z, Ge B, Xiang X, Zhang T, Gao L, Shi H, Wang C, Huang J. LncROR Promotes Bladder Cancer Cell Proliferation, Migration, and Epithelial-Mesenchymal Transition. CELL PHYSIOL BIOCHEM. 2017;41(6):2399–410.
26. Li L, Gu M, You B, Shi S, Shan Y, Bao L, You Y. Long non-coding RNA ROR promotes proliferation, migration and chemoresistance of nasopharyngeal carcinoma. CANCER SCI. 2016;107(9):1215–22.
27. Xue M, Chen W, Xiang A, Wang R, Chen H, Pan J, Pang H, An H, Wang X, Hou H, et al. Hypoxic exosomes facilitate bladder tumor growth and development through transferring long non-coding RNA-UCA1. MOL CANCER. 2017;16(1):143.
28. Liu Y, Xue M, Du S, Feng W, Zhang K, Zhang L, Liu H, Jia G, Wu L, Hu X, et al. Competitive endogenous RNA is an intrinsic component of EMT regulatory circuits and modulates EMT. NAT COMMUN. 2019;10(1):1637.
29. Yang H, Li Q, Wu Y, Dong J, Lao Y, Ding Z, Xiao C, Fu J, Bai S: **Long noncoding RNA RP11400N13.3 promotes the progression of colorectal cancer by regulating the miR47223p/P2RY8 axis**. *ONCOL REP* 2020, **44**(5):2045–2055.
30. Bian Z, Zhang J, Li M, Feng Y, Yao S, Song M, Qi X, Fei B, Yin Y, Hua D, et al: **Long non-coding RNA LINC00152 promotes cell proliferation, metastasis, and confers 5-FU resistance in colorectal cancer by inhibiting miR-139-5p**. *ONCOGENESIS* 2017, **6**(11):395.
31. Chen X, Li D, Gao Y, Tang W, Iw L, Cao Y, Hao B. Long Intergenic Noncoding RNA 00152 Promotes Glioma Cell Proliferation and Invasion by Interacting with MiR-16. CELL PHYSIOL BIOCHEM. 2018;46(3):1055–64.
32. Pang Q, Ge J, Shao Y, Sun W, Song H, Xia T, Xiao B, Guo J. Increased expression of long intergenic non-coding RNA LINC00152 in gastric cancer and its clinical significance. Tumour Biol. 2014;35(6):5441–7.
33. Chen WM, Huang MD, Sun DP, Kong R, Xu TP, Xia R, Zhang EB, Shu YQ. Long intergenic non-coding RNA 00152 promotes tumor cell cycle progression by binding to EZH2 and repressing p15 and p21 in gastric cancer. Oncotarget. 2016;7(9):9773–87.
34. Li Q, Shao Y, Zhang X, Zheng T, Miao M, Qin L, Wang B, Ye G, Xiao B, Guo J. Plasma long noncoding RNA protected by exosomes as a potential stable biomarker for gastric cancer. Tumour Biol. 2015;36(3):2007–12.
35. Ji J, Tang J, Deng L, Xie Y, Jiang R, Li G, Sun B. LINC00152 promotes proliferation in hepatocellular carcinoma by targeting EpCAM via the mTOR signaling pathway. Oncotarget. 2015;6(40):42813–24.
36. Cai Q, Wang ZQ, Wang SH, Li C, Zhu ZG, Quan ZW, Zhang WJ. Upregulation of long non-coding RNA LINC00152 by SP1 contributes to gallbladder cancer cell growth and tumor metastasis via PI3K/AKT pathway. AM J TRANSL RES. 2016;8(10):4068–81.

37. Chen QN, Chen X, Chen ZY, Nie FQ, Wei CC, Ma HW, Wan L, Yan S, Ren SN, Wang ZX. Long intergenic non-coding RNA 00152 promotes lung adenocarcinoma proliferation via interacting with EZH2 and repressing IL24 expression. MOL CANCER. 2017;16(1):17.
38. Wang T, Cao L, Dong X, Wu F, De W, Huang L, Wan Q. LINC01116 promotes tumor proliferation and neutrophil recruitment via DDX5-mediated regulation of IL-1beta in glioma cell. CELL DEATH DIS. 2020;11(5):302.
39. Li H, Tian X, Wang P, Hu J, Qin R, Xu R, Liu K, Hao J, Tian N. LINC01128 resisted acute myeloid leukemia through regulating miR-4260/NR3C2. CANCER BIOL THER. 2020;21(7):615–22.
40. Tang Y, Cao G, Zhao G, Wang C, Qin Q. LncRNA differentiation antagonizing non-protein coding RNA promotes proliferation and invasion through regulating miR-135a/NLRP37 axis in pancreatic cancer. Invest New Drugs. 2020;38(3):714–21.

Figures

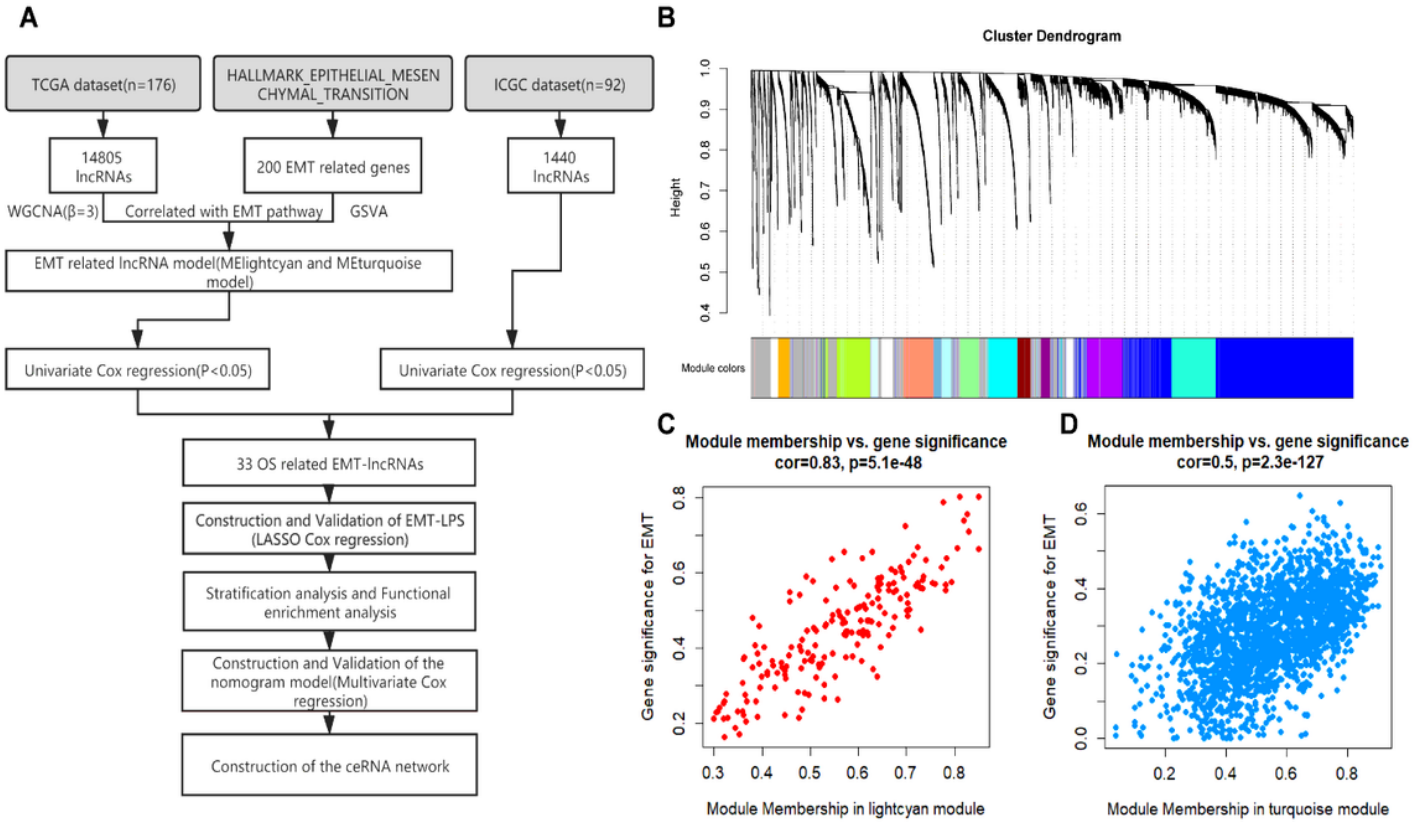


Figure 1

(A) Study flow chart. (B)Weighted correlation network analysis (WGCNA) of lncRNAs in Pancreatic adenocarcinoma (PAAD) samples. Gene clustering tree (dendrogram) obtained by hierarchical clustering of adjacency-based dissimilarity. (C) Correlation between membership in the lightcyan module and membership in the EMT pathway of the eigengenes in lightcyan module by Pearson correlation analysis. Cor, correlation coefficient. (D) Correlation between membership in the turquoise module and membership

in the autophagy pathway of the eigengenes in turquoise module by Pearson correlation analysis. Cor, correlation coefficient.

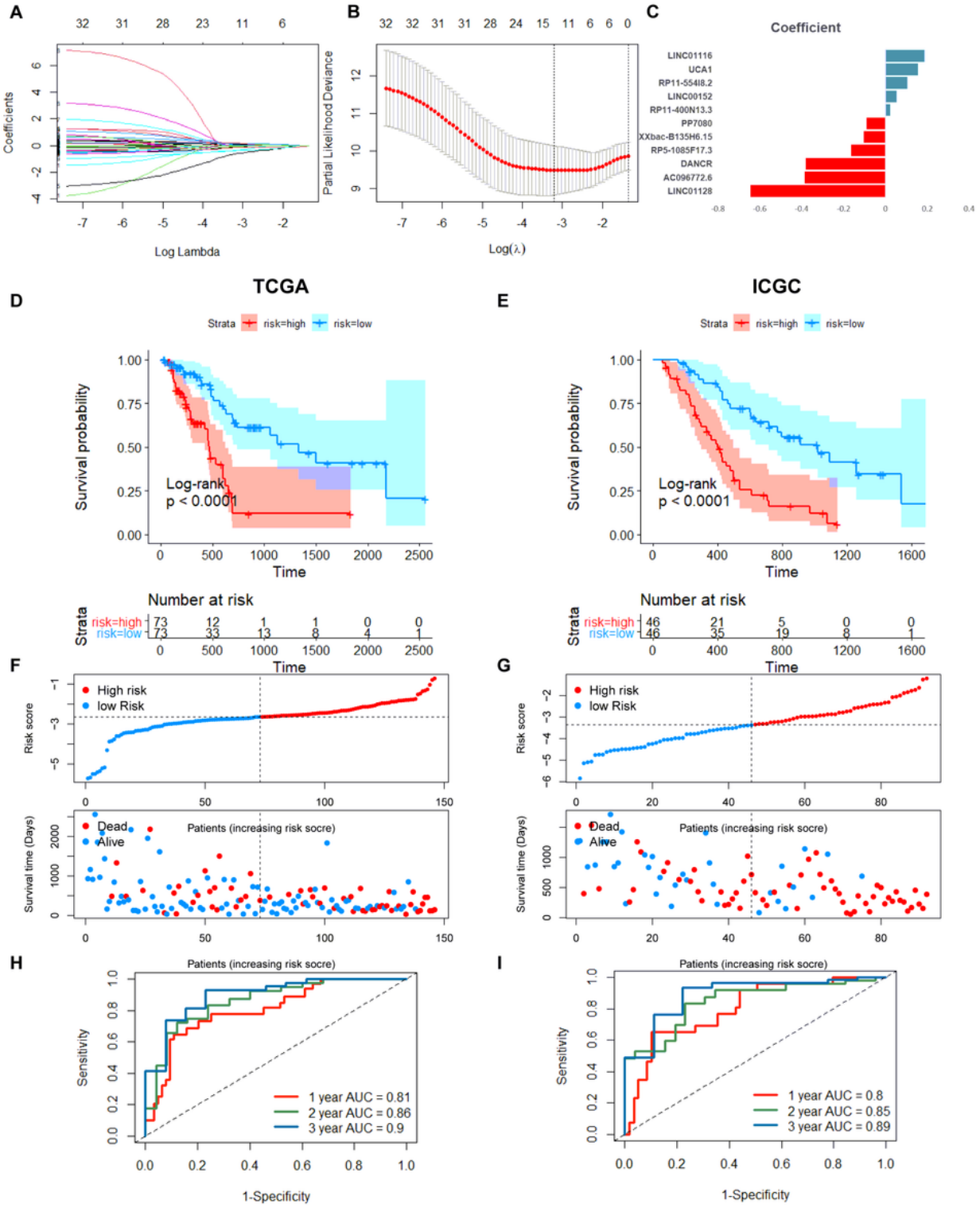


Figure 2

Construction of the EMT-LPS in the TCGA dataset and validation in ICGC dataset. (A-C) Least absolute1 shrinkage and selection operator (LASSO) regression was performed, calculating the minimum criteria (A-B) and coefficients (C). (D-E) Kaplan–Meier survival curve for the high- and low-risk groups divided by the

cutoff value in the TCGA dataset(D) and ICGC dataset (E), respectively. p-values were obtained via log-rank test. The distribution of risk score and survival state of the selected eleven EMT lncRNAs in the TCGA dataset(F) and ICGC dataset(G), respectively. (H-I) The receiver operating characteristic curve (ROC) for the prognosis prediction of the signature at 1/2/3 years of overall survival (OS) in the TCGA dataset (H) and ICGC dataset (I), respectively.

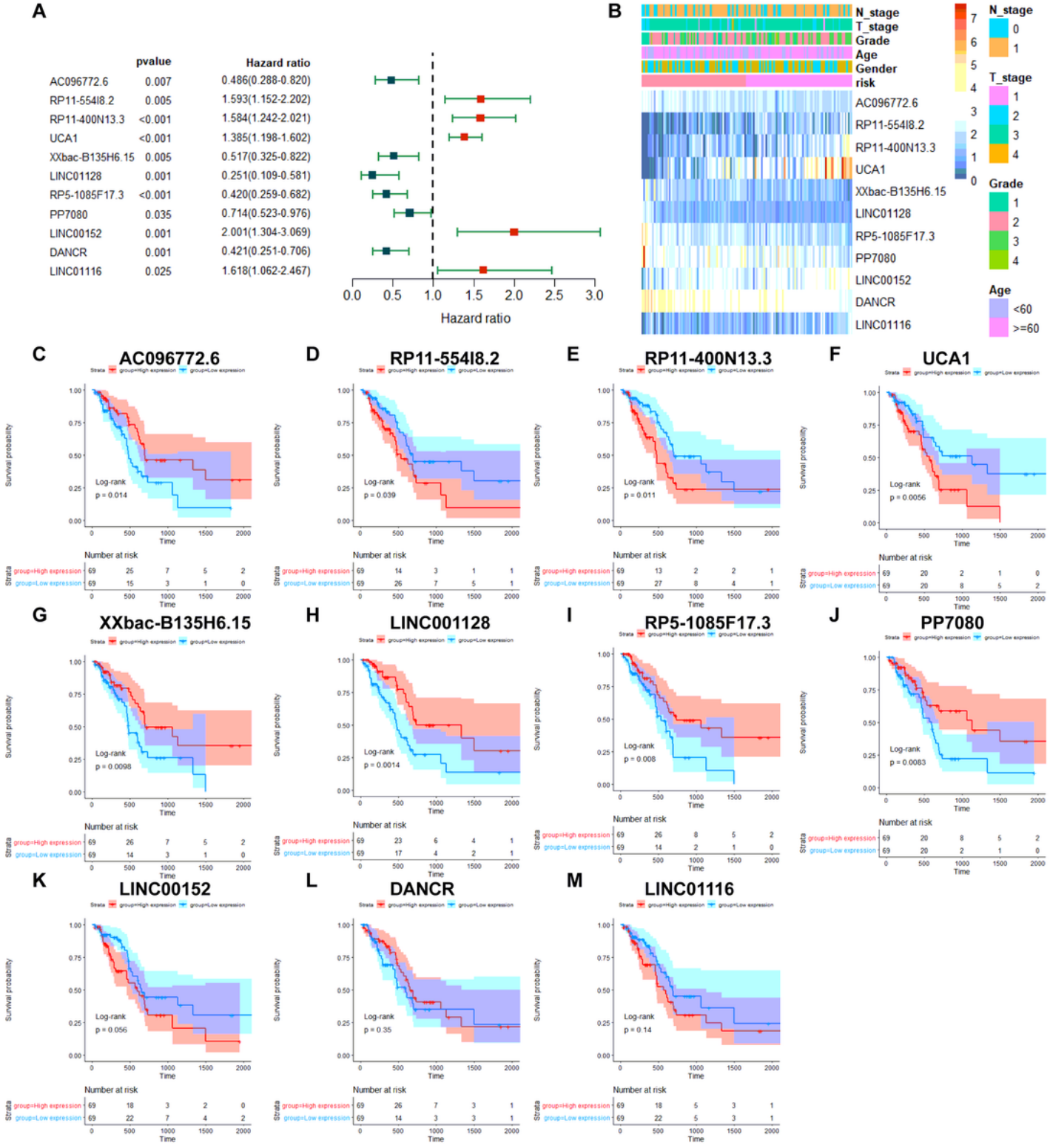


Figure 3

Prognostic Analysis of the elven EMT-Related lncRNAs in TCGA. (A) Forest plot of the prognostic ability of the eleven EMT-related lncRNAs included in the prognostic signature. (B) Heatmap of the associations between the expression levels of the eleven EMT-related lncRNAs and clinicopathological features in the TCGA dataset. (C–M) Kaplan–Meier curves showing that patients with different expression levels of the eleven EMT-related lncRNAs had different overall survival.

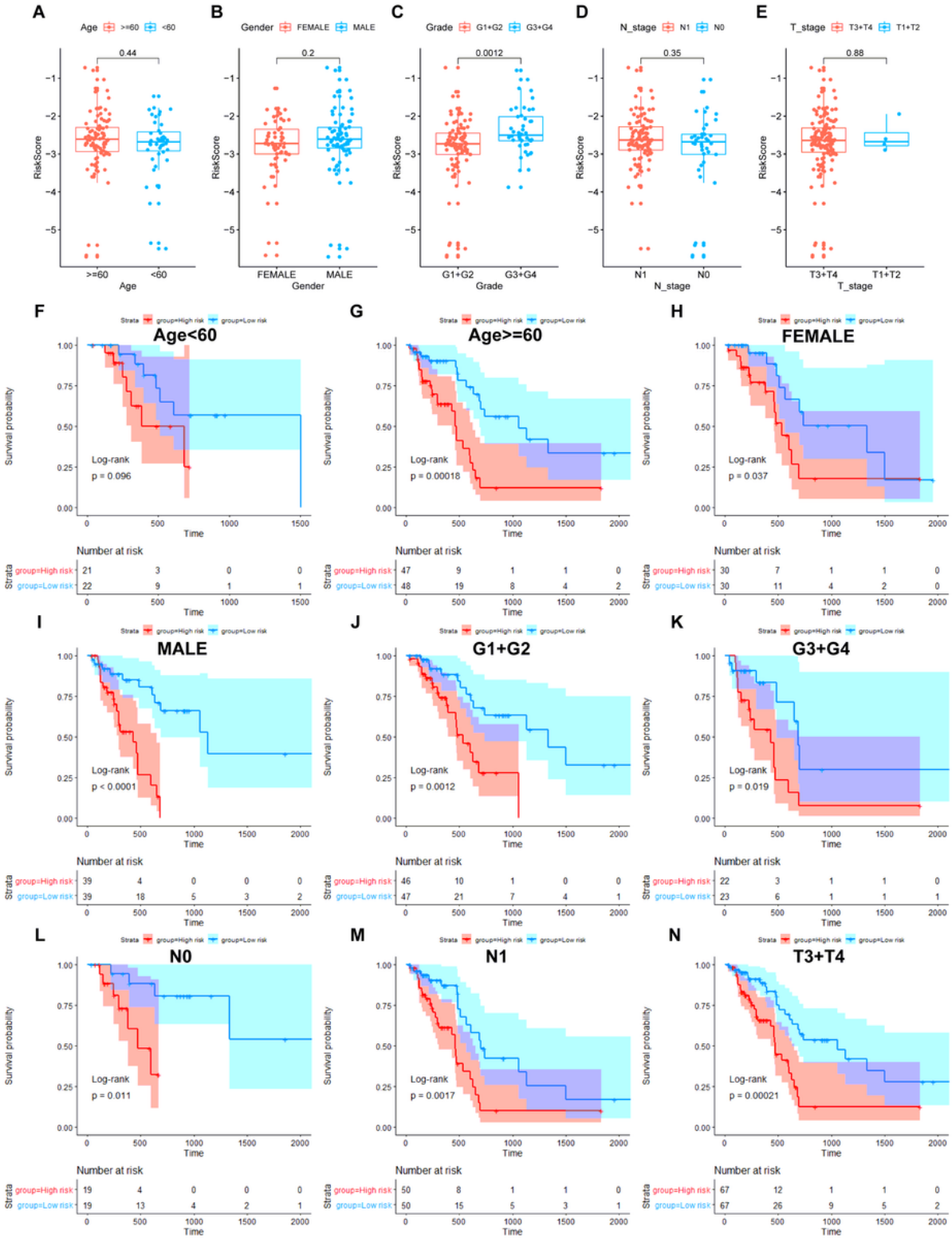
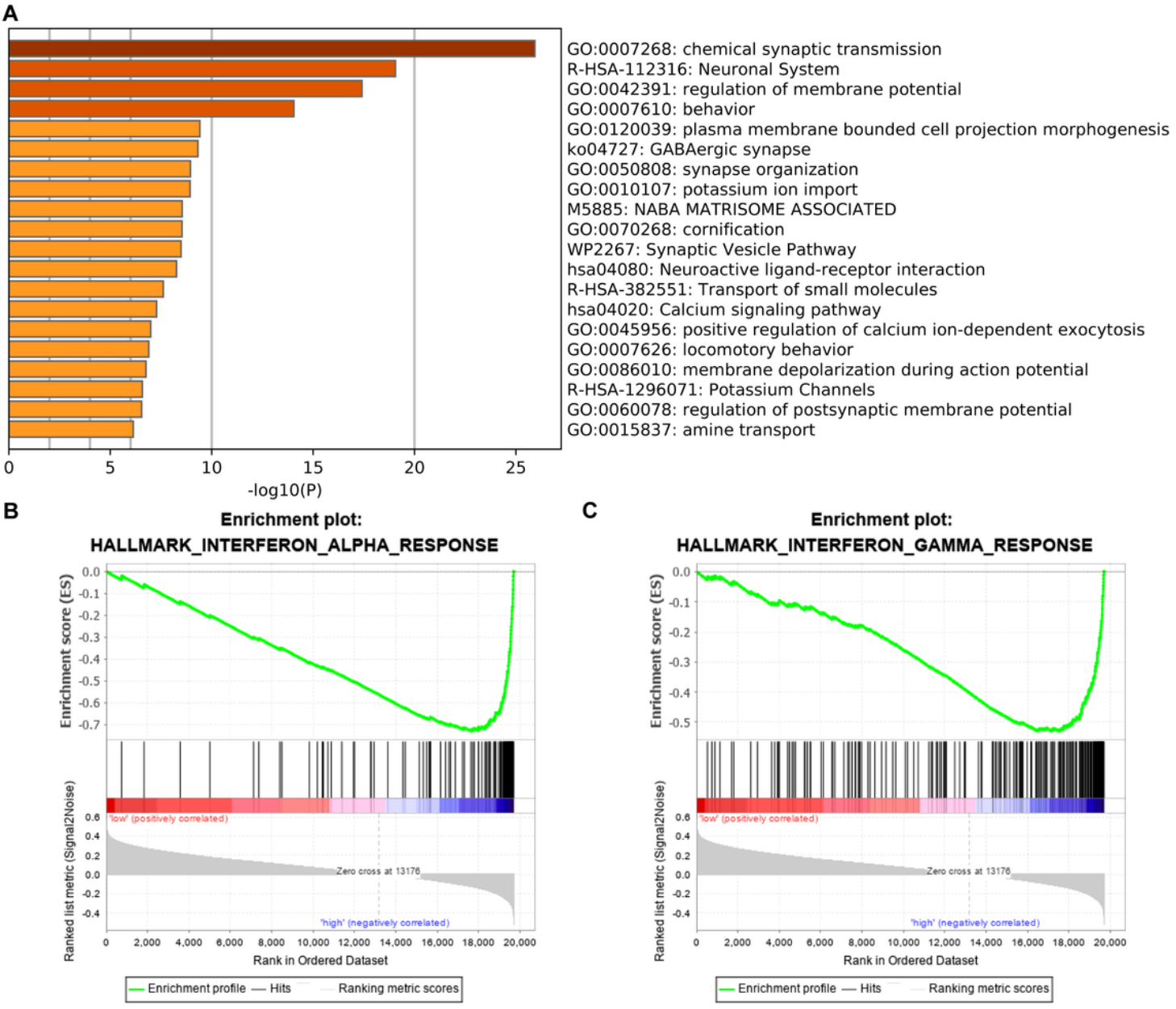


Figure 4

Stratification Analysis of the EMT-LPS in TCGA patients with different clincial characteristic. (A–E) Patients with different clinicopathological features (only including Grade) had different levels of risk scores, calculated based on the EMT-related lncRNA prognostic signature(EMT-LPS). (G–L) The EMT-LPS retained its prognostic value in multiple subgroups of PAAD patients (including patients aged ≤ 60 or > 60 years, female or male patients, patients with grade I +II or grade III+IV, and patients with TNM stage N0, N1 or T3+T4).



Upregulated in class	High risk	Upregulated in class	High risk
Enrichment Score (ES)	-0.7310267	Enrichment Score (ES)	-0.52983075
Normalized Enrichment Score (NES)	-1.8306823	Normalized Enrichment Score (NES)	-1.6644077
Nominal p-value	0.003968254	Nominal p-value	0.049822062
FDR q-value	0.16821192	FDR q-value	0.3240651
FWER p-Value	0.122	FWER p-Value	0.302

Figure 5

Functional analysis of 710 differentially expressed mRNA (DEms) between the low- and high-risk subgroups. (A) Heatmap of enriched terms across the inputted gene list, colored according to p-value. (B-C) Gene set enrichment analysis (GSEA) indicating that tumor hallmarks were enriched in the high-risk subgroup.

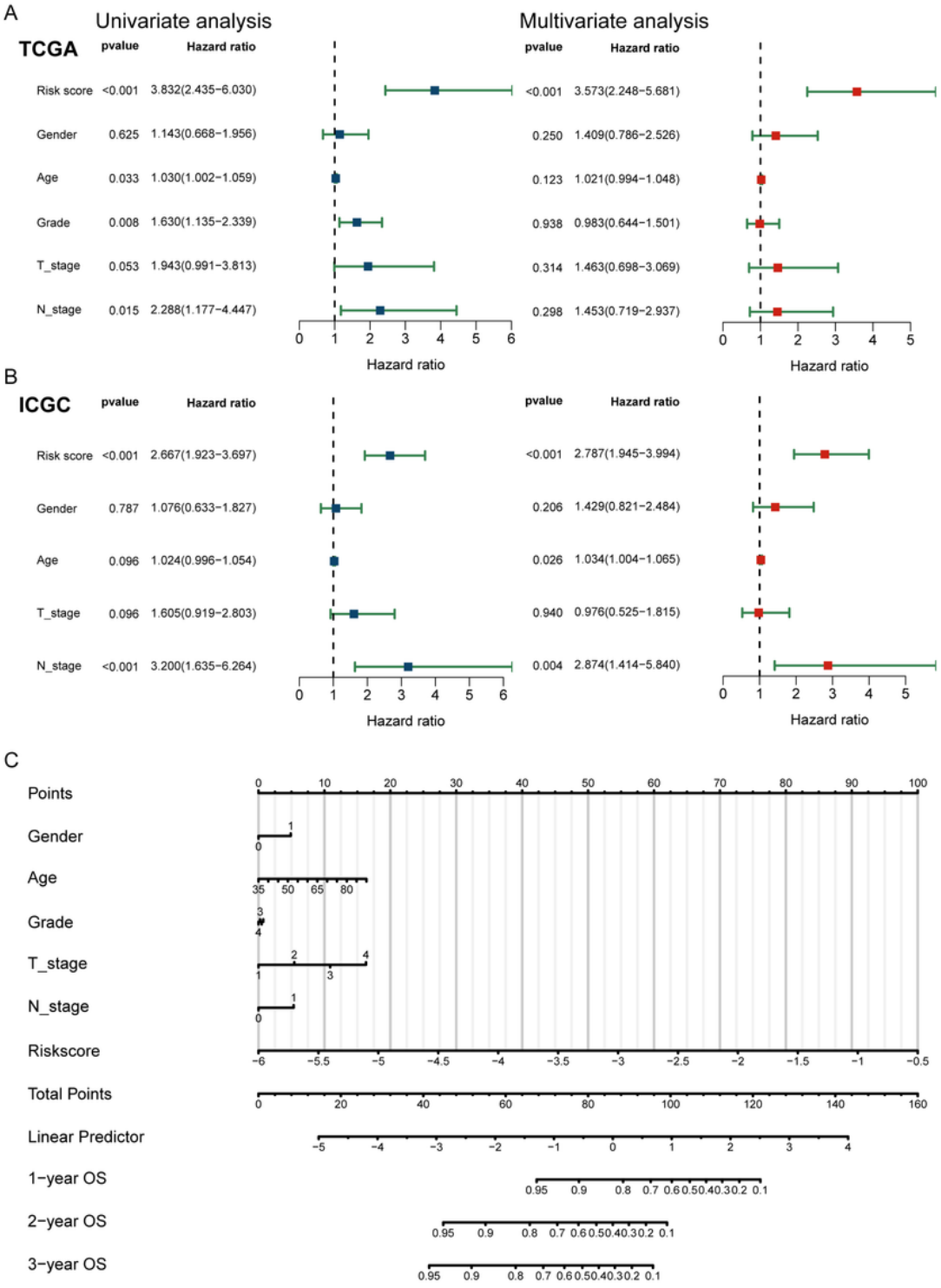


Figure 6

EMT-LPS Was an Independent Prognostic Factor for PAAD Patients. (A,B) Univariate and multivariate analyses revealed that risk score [based on the EMT-related lncRNA prognostic signature (EMT-LPS)] was an independent prognostic predictor in the TCGA and ICGC datasets. (C) Nomogram based on risk score, Gender, Age, T_stage, N_stage and Grade.

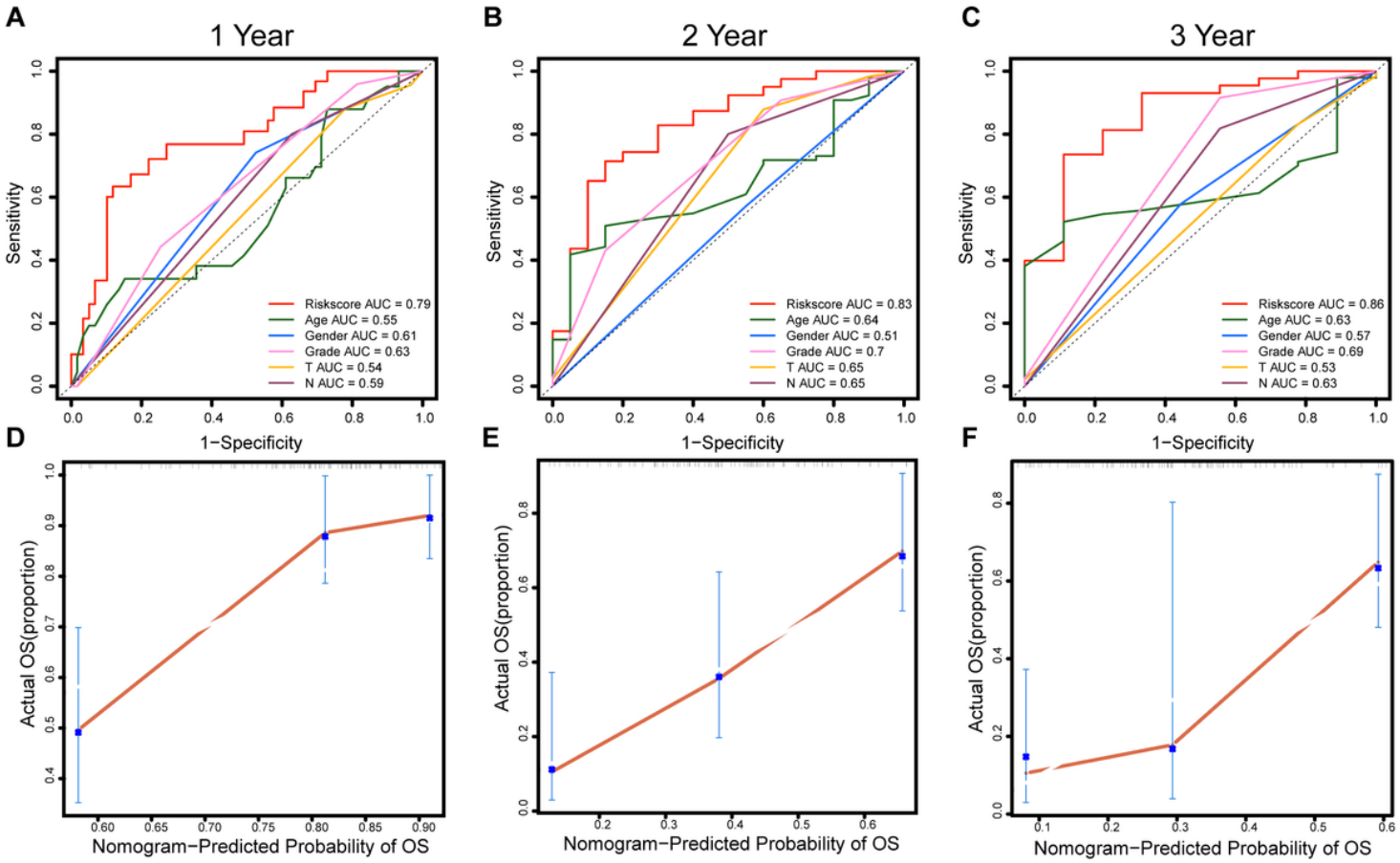


Figure 7

Construction of the EMT-LPS-Based Calibration in TCGA dataset. (A–C) Calibration plots of the nomogram for predicting the probability of OS at 1, 2, and 3 years in the TCGA dataset. (D–F) Time-dependent receiver operating characteristic (ROC) curves for the nomogram, Rskscore, Age, Grade,T stage and N stage in the TCGA dataset (for predicting 1, 2, and 3-year OS).

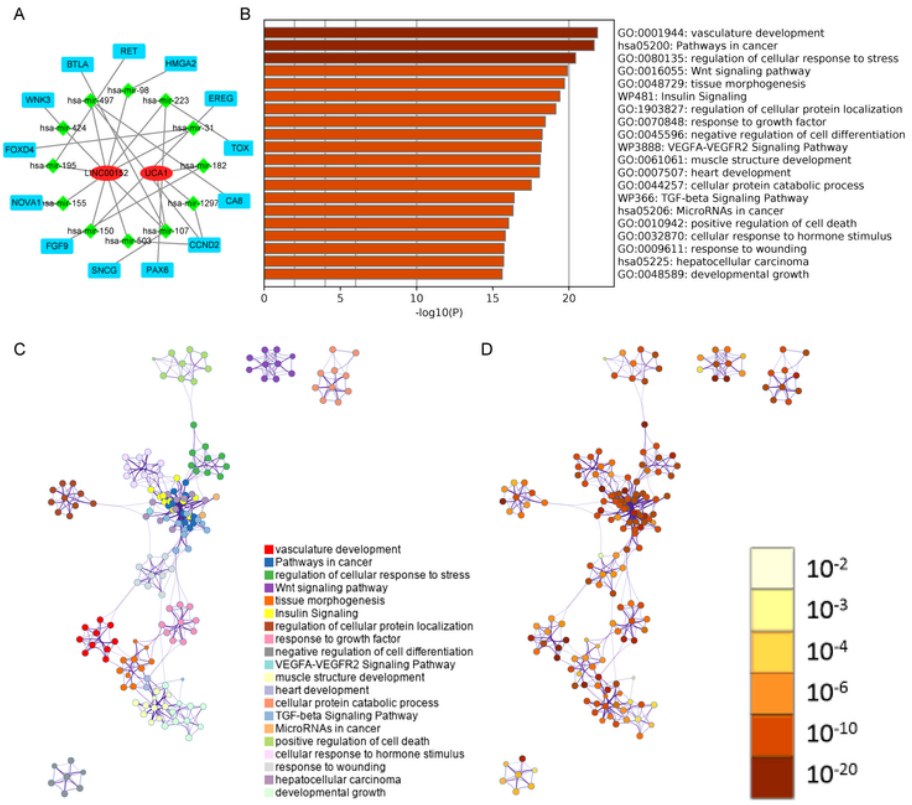


Figure 8

(A) The result of the ceRNA Network and Functional Enrichment Analysis. The ceRNA network of the two EMT-related lncRNAs (red) and their target miRNAs (green) and mRNAs (blue). (B) Heatmap of enriched terms across the 929 mRNAs, colored according to p-value. Network of enriched terms colored according to (C) cluster ID (nodes with the same cluster ID are typically close to each other) and (D) p-value (terms with more genes tend to have higher p-values).

Supplementary Files

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

- SupplementaryMaterial.docx