Am J Cancer Res 2021;11(2):495-512
www.ajcr.us /ISSN:2156-6976/ajcr0119937

Original Article

A new 7-gene survival score assay for pancreatic cancer patient prognosis prediction

Lisi Luo[1*], Yufang Li[1*], Chumei Huang[2], Yujing Lin[3], Yonghui Su[4], Hong Cen[4], Yutong Chen[1], Siqi Peng[5,6], Tianyi Ren[1], Rongzhi Xie[1], Linjuan Zeng[1]

[1]Department of Abdominal Oncology, The Cancer Center of The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai 519000, Guangdong Province, China; [2]Department of Gastroenterology, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen 518107, China; [3]Department of Pathology, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, China; [4]Department of General Surgery, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai 519000, Guangdong Province, China; [5]Center for Interventional Medicine, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai 519000, Guangdong Province, China; [6]Guangdong Provincial Key Laboratory of Biomedical Imaging, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai 519000, Guangdong Province, China. [*]Equal contributors.

Received August 9, 2020; Accepted December 18, 2020; Epub February 1, 2021; Published February 15, 2021

Abstract: Gene expression features that are valuable for pancreatic ductal adenocarcinoma (PDAC) prognosis are still largely unknown. We aimed to explore pivotal molecular signatures for PDAC progression and establish an efficient survival score to predict PDAC prognosis. Overall, 163 overlapping genes were identified from three statistical methods, including differentially expressed genes (DEGs), coexpression network analysis (WGCNA), and target genes for miRNAs that were significantly related to PDAC patients' overall survival (OS). Then, according to the optimal value of the cross-validation curve (lambda = 0.031), 7 non-zero coefficients (ARNTL2, DSG3, PTPRR, ANLN, S100A14, ANKRD22, and TSPAN7) were selected to establish a prognostic prediction model of PDAC patients. We further confirmed the expression level of 7 genes using RT-PCR, western blot, and immunohistochemistry staining in PDAC patients' tissues. Our results showed that the ROC curve of the 7-mRNA model indicated good predictive ability for 1- and 2-year OS in three datasets (TCGA: 0.71, 0.69; ICGC: 0.8, 0.74; GEO batch: 0.61, 0.7, respectively). The hazard ratio (HR) of the low-risk group had a similar significant result (TCGA: HR = 0.3723; ICGC: HR = 0.2813; GEO batch: HR = 0.4999; all $P < 0.001$). Furthermore, Log-rank test results in three cohorts showed that the 7-mRNA assay excellently predicted the prognosis and metastasis, especially in TNM stage I&II subgroups of PDAC. In conclusion, the strong validation of our 7-mRNA signature indicates the promising effectiveness of its clinical application, especially in patients with TNM stages I&II.

Keywords: Pancreatic cancer, bioinformatics, molecular signature, a prediction model

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal diseases worldwide [1]. One-half of patients are diagnosed at a distant stage, while approximately thirty percent present with a regional stage [2]. Surgery is regarded as the only treatment for cure, but long-term survival rates after surgical resection are low, and recurrence rates remain high despite adjuvant treatment, with the majority of patients relapsing within two years [3, 4].

Most patients receive chemotherapy after surgery to improve long-term survival; however, there is no uniform standard for postoperative patients who are suitable for chemotherapy in relatively early TNM stages. In contrast, for advanced or relapsed PDAC, even the triple combination strategy mFOLFIRINOX (fluorouracil, leucovorin, irinotecan, and oxaliplatin) resulted in short progression-free survival (PFS, 6.4 months) and overall survival (OS, 11.1 months) among advanced PDAC patients [5]. Although Olaparib exhibits promising effects on PDAC patients with the BRCA1/2 mutation, the mutation rate is only 4-7% [6, 7]. Most targeted therapies have remained disappointing; for example, erlotinib combined with gemcitabine prolonged OS by only 0.5 months compared with gemcitabine alone [8], and clinical trials using AVASTIN or cetuximab obtained negative results [9, 10]. Hence, it is urgently needed to explore pivotal molecular mechanisms for tumor progression and to screen new therapeutic targets.

The characteristics of gene expression and biofunctions in PDAC tissues are still largely unknown. Serum carbohydrate antigen 19-9 (CA19-9), carcinoembryonic antigen (CEA), and carbohydrate antigen 125 (CA125) are commonly used biomarkers for diagnosis, therapeutic response evaluation, and prognosis prediction of PDAC [11]. However, the specificities of these antigens are poor, and plasma protein levels are affected by many factors. It is rational to characterize PDAC using gene expression features at the time of diagnosis, which might be biomarkers for prognosis prediction and targets for treatment.

To obtain gene expression profiles from PDAC samples and to avoid bias in patient selection, we performed bioinformatics analysis using database-based big data for PDAC patients. With growing amounts of data, numerous bioinformatics tools have been developed over the years, which is a vast and complex multidisciplinary research area gaining research attention [12]. The results of bioinformatics analyses are entering clinical practice, guiding diagnosis, prognosis, and treatment. Currently, bioinformatics has become a promising field in cancer research.

In the present study, we combined multiple public databases and used comprehensive tools to analyze data and the clinical features of PDAC. We discussed the gene expression characteristics and underlying mechanisms of PDAC progression and established a valuable prognostic prediction model using a 7-mRNA signature.

Materials and methods

Primers and antibodies

The primer sequences of 7 mRNAs and GAPDH used in this study are listed in Table S1. Eight primary antibodies were purchased, including the anti-PTPRR (GTX111152, GeneTex, CA, USA), anti-ANLN antibody (GTX107742, GeneTex, CA, USA), anti-TSPAN7 (ab211870, Abcam, Cambridge, UK), anti-DSG3 (ab14416, Abcam, Cambridge, UK), anti-ANKRD22 (PA553010, Thermo Fisher Scientific, MA, USA), anti-ARNTL2 antibody (PA563653, Thermo Fisher Scientific, MA, USA), anti-S100A14 (10489-1-AP, Proteintech, Chicago, USA) and anti-GAPDH antibody (5174S, CST, MA, USA). Peroxidase-conjugated secondary antibodies were purchased from Cell Signaling Technology (7074S, 7076S, CST, MA, USA).

Human tissues

Tissue samples were collected from patients who received an operation at the Fifth Affiliated Hospital of Sun Yat-sen University (Zhuhai, China) between October 2013 and May 2020. The protocol was approved by the Institutional Ethical Review Board of the hospital. Among the collected samples, 12 fresh isolated samples including 4 pairs of PDAC and adjacent non-tumor tissue, and 4 benign pancreatic tissues, were used to detect mRNA and protein expression levels by RT-PCR and Western blot assays. Another 74 paraffin-embedded tissue specimens (including 43 PDAC and 31 non-tumor) were used for immunohistochemistry. PDAC patients were 26 males and 17 females, and the median age was 60.5 years old (range 32-85). The majority of patients were classified as stage I&II (Stage I: n = 8, Stage II: n = 8, Stage III: n = 4, Stage IV: n = 3).

PDAC dataset collection and processing

All the expression data and clinical information were open-source and accessible. The gene expression datasets GSE28735, GSE62452, and GSE57495 and their clinical information were downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/). The mRNA/miRNA expression profiles of PDACs and their corresponding clinical information were obtained from The Cancer Genome Atlas (TCGA) public database (https://cancergenome.nih.gov/) and ICGC (https://icgc.org/).

The dataset GSE28735, which includes 45 PDAC and 45 normal samples, was used to screen DEGs and to construct the co-expression networks. From the TCGA data portal, the OS-related miRNA and their target genes were selected. We then identified common mRNAs by 3 independent ways for the further Cox and Lasso regression in TCGA. In addition, we excluded patients who lacked survival information and verified our result in 95 cases of ICGC pancreatic cancer and 170 cases of GEO batch samples by integrating GSE28735, GSE62452, and GSE57495. The enrolled PDAC datasets are described in Table S2.

We conducted a series of processing steps after mRNA/miRNA data downloading, including calculating the raw counts, choosing the tumor samples, mapping to the gene symbol name, and removing the batch effects between GSE28735, GSE62452, and GSE57495 [13].

GEO differential mRNA expression of genes

The series matrix file of GSE28735 was read using the GEO query package in R language software, and the data had already been preprocessed by RNA normalized and log2 transformed by the submitters. If there were multiple probes annotated to the same gene, we chose the probe with the max expression level to represent the expression level of this gene. Bioconductor (version 3.8.0, http://www.bioconductor.org) package Limma [14] was used to identify significantly differentially expressed genes (DEGs) between PDAC and normal samples. *P* values were adjusted for multiple testing using the Benjamini-Hochberg False Discovery Rate (FDR) method. Genes with FDR < 0.05 and |log2 fold change (FC)| ≥ 1 were considered as DEGs.

Weighted gene coexpression network construction and PDAC-related modules

The Weighted Gene Coexpression Network Analysis (WGCNA) was conducted by the WGCNA package [15] in R software. As a previous study showed that the WGCNA analysis was sensitive to batch effects and outlier samples, we performed hierarchical cluster analysis [16]. The module eigengene (ME), which is considered representative of the gene expression profiles, was calculated to identify clinical-associated modules. To find the most tumor-related modules, we conducted Module-Trait Relationships calculations for each module. Then, for genes in the significant tumor-related modules, we calculated the Gene Significance [17] and Gene Module Membership (MM) within the genes, modules, and clinical traits. Finally, we identified the genes in the PDAC-related modules [18, 19].

miRNA profile survival analysis and target genes

The normalized miRNA expression profiles and their clinical profiles of PDAC were obtained from the PDAC data portal. We only included miRNAs that were expressed in at least 10 tumor samples. Then, we used 2 methods, univariate analysis using the Kaplan-Meier curve with log-rank test and multivariate Cox proportional hazards regression analysis, to conduct the survival analysis. To explore the prognostic-correlations of clinical factors, as previously reported [20], patients were divided into high- and low-expression groups based on the gene expression median value and three clinical factors (gender, age, and stage) in the multivariate Cox analysis using the "Survival" package in R. The adjusted *P*-value cutoff of 0.01 was used to identify the prognostic-related miRNAs. The target genes of these selected miRNAs were predicted using the miRwalk database (version 2.0, http://www.zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2).

Acquisition of intersecting genes

Overlapping genes were identified as candidates for the subsequent analysis and were oriented from the differential expression analysis, WGCNA analysis, and prognostic-related-miRNA target genes. A Venn diagram was created using the VennDiagram package in R.

Prognostic model building in TCGA

First, we conducted univariate Cox regression among the candidate genes to screen OS-related genes. Then, Least absolute shrinkage and selection operator (LASSO) regression was performed to further select genes by LASSO penalty, and the optimal values of the penalty parameter (lambda) were identified by 10-times cross-validations using the 'glmnet' package. Finally, we built a prognostic signature model based on the multivariate Cox proportional hazards regression. The risk score, also known as the prognosis index (PI), was established for the gene-based-model according to the following equation: PI = $\Sigma \beta i \times Ei$ [21], where βi represents the coefficient of the involved gene i, and *Ei* refers to the expression level of the corresponding gene. Besides, we employed time-dependent receiver operating characteristic curve analysis to find the best PI cutoff value and to evaluate the model performance by the area under the curve (AUC) values. The best PI cutoff value, which could divide the patients into a high-risk or low-risk survival group, was identified according to the maximal Youden's index (J). Then, Kaplan-Meier (K-M) survival curves were drawn to assess the association between risk score and OS utilizing the survival package [22].

Validation in ICGC and GEO batch datasets

We used ICGC and GEO batch datasets to examine the predictive ability of the prognostic signature model. For each dataset, HRs, and AUCs at 1, 2, and 5 years of signatures were calculated. Furthermore, Kaplan-Meier survival curves and time-dependent receiver operating characteristic curves were plotted. Also, we further explored the roles of 7 genes in PDAC tumor metastasis on ICGC dataset.

GO enrichment analysis, KEGG and GSEA pathway analysis

Gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Set Enrichment Analysis (GSEA) of the DEGs were performed for the biological functional annotation and pathway enrichment analysis through the "clusterProfiler" package. Terms with FDR < 0.05 were considered significantly enriched.

RNA extraction, cDNA synthesis, and qRT-PCR

Total RNA was extracted using TRIzol (Invitrogen, CA, USA), and RNA concentration was detected according to a 260/280 by ABI Prism 7900 Sequence Detection System (Applied Biosystems, CA, USA). The High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) was used to synthesize the first-strand cDNA. qRT-PCR assays were performed on an ABI Prism 7900 Sequence Detection System using TB Green® Premix Ex Taq™ (Takara, Shiga, Japan). The relative expression levels of 7 mRNA were quantified to GAPDH.

Western blotting

The total protein of 12 tissue samples was extracted using RIPA Lysis buffer with protease inhibitors (PMSF) and quantified using BCA Protein assay kit (Thermo Fisher Scientific, MA, USA), respectively. 20 μg of total protein was separated on 15% or 8% Tris-HCl gel and transferred to a 0.2-μm or 0.45-μm polyvinylidene difluoride membrane (Millipore, MA, USA) according to the molecular weight of the detected protein. All primary and secondary antibodies dilution and incubation were performed following the manufacturers' instructions. Blots were visualized using the Tanon-5200 Multiimaging System (Tanon, China). GAPDH was used as a housekeeping gene for loading control. All experiments were repeated at least three times.

Immunohistochemistry (IHC) staining

The IHC staining kit (Boster, Wuhan, China) was used for immunohistochemistry staining on paraffin-embedded tissue sections. All experiments were performed according to the manufactory's protocol. The results of IHC were interpreted by two independent pathologists who did not know the clinicopathological data of the patients and the results were agreed upon together. The immunoreactive score (IRS) was calculated according to the ratio of positive cells and the staining intensity. The ratio of positive cells was defined as follows: 1 point (< 25% cell staining), 2 points (25-50% cell staining), 3 points (51-75% cell staining), and 4 points (> 75% cell staining). The staining intensity was scored as 0 points (negative), 1 point (weakly positive), 2 points (positive), and 3 points (strongly positive). Images were acquired by using a microscope (Zeiss, Germany) (magnification × 400).

Statistical analysis

Differences between the two groups were assessed using Student's t-test. Continuous data are presented as mean ± SD. Kaplan-Meier survival plots were assessed by the log-rank test. All analyses and graphics were conducted in R (version 3.6.2, https://www.r-proje ct.org/). P < 0.05 was considered significant.

Results

Screening of differentially expressed genes in the GEO dataset

In total, 413 differentially expressed genes (DEGs) in GSE28735 were identified. Among these DEGs, 256 genes were upregulated, and 157 genes were downregulated. The volcano plots of DEGs were drawn using the ggplot2 R package to visualize the results in Figure S1.

Identification of genes in the PDAC-associated modules by WGCNA

Overall, 5078 genes and 89 samples were selected after gene and sample screening and preprocessing, and the removed sample was normal in GSM711957 (**Figure 1A**). We used a power calculation of β = 12 (scale-free R2 = 0.895) (**Figure 1B**, **1C**). All selected gene dendrograms and their corresponding modules are displayed in **Figure 1D**. There were 11 modules, including the special gray module according to the network result (i.e., black, blue, brown, green, gray, magenta, pink, purple, red, turquoise, and yellow modules). According to a previous study, the gray module had no association with clinical traits [23]. Among these 11 modules, red (r = 0.70; P = 2e-14) and turquoise (r = 0.69; P = 5e-14) modules showed positive relationships with PDAC (**Figure 1E**). Furthermore, the genes in the turquoise and red modules showed strong correlations with PDAC (Red: r = 0.76; P = 5.8e-25; Turquoise: r = 0.69; P = 4.3e-170) (**Figure 1F**, **1G**). Finally, we identified turquoise and red modules as the key modules, in which there were 1324 genes, including 1198 and 126 genes, respectively.

Prognosis-associated miRNA and intersecting genes

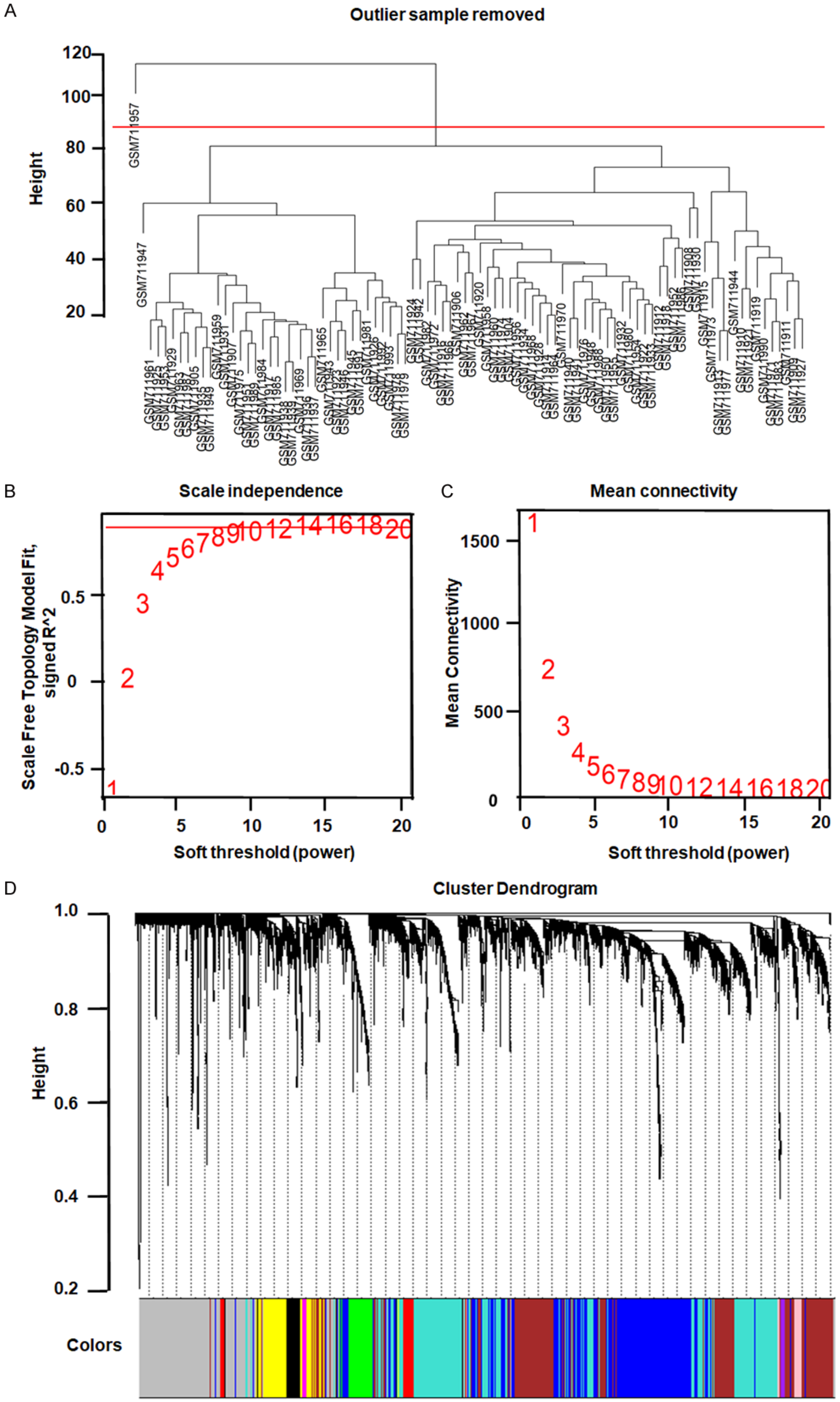
microRNAs are small-non-coding RNAs that can modulate mRNA expression by binding to target-mRNA complementary sequences. A previous study demonstrated that microRNAs can promote tumor cell growth and metastasis by binding to tumor suppressor genes or act as tumor suppressor factors by binding to oncogenes [24]. It is rational to infer that prognosis-associated-miRNA target genes may exert important biofunctions in PDAC progression and serve as valuable predictive markers. Overall, 1881 miRNAs and 183 samples of PDAC were downloaded from the UCSC Xena (http://xena.ucsc.edu/). We acquired 179 tumor samples and 467 miRNAs, which were expressed in at least 10 tumor samples. There were 29 miRNAs and 12 miRNAs identified by the log-rank test and Multivariate Cox proportional hazards regression analysis, respectively, and 10 intersecting miRNAs were described (Figure S2A and Table S3). miRNA-143 was the only microRNA whose high expression correlated with poor prognosis, and 9 other miRNAs exhibited a positive influence on overall survival. Then, 12425 target genes were predicted for these 10 miRNAs by means of the miRwalk database. Finally, 163 overlapping genes were obtained from the above mentioned three analyses (DEG, WGCNA, and 10-miRNAs-target-genes), and this process was performed using R package VennDiagram (Figure S2B).

Identification of a 7-gene signature

We initially collected mRNA expression levels and clinical information of the 163-overlapping genes of 174 patients from TCGA. For univariate Cox analysis, 42 mRNAs remained (Table S4). Then, the cross-validation curve was plotted on these 42 mRNAs to find the optimal lambda (**Figure 2A**). Here, we obtained 7 nonzero coefficients according to the optimal value (lambda = 0.031), which were ARNTL2, DSG3, PTPRR, ANLN, S100A14, ANKRD22, and TSPAN7. Their expression profiles in PDAC are displayed in the heatmap (**Figure 2B**). In addition, the correlations between 10 miRNAs and 7 genes explored in TCGA datasets were described in Figure S3. We then calculated the risk score based on their expression levels and related coefficients: Risk score = 0.233* (expression level of ARNTL2) + 0.078* (expression level of DSG3) + 0.044* (expression level of PTPRR) + 0.029* (expression level of ANLN) + 0.055* (expression level of S100A14) + 0.06* (expression level of ANKRD22) - 0.127* (expression level of TSPAN7). The association between the risk score and survival status in the TCGA dataset was explored in **Figure 2C**, and we observed that alive numbers decreased as PI values increased. The HRs and 95% confidence intervals (CIs) are shown in the forest plot (**Figure 2D**).

Predictive ability of the 7-mRNA score model

In the TCGA database, the prognosis index (PI) for each sample and ROC curve at 1, 2, and 5 years were created by the time ROC package in R, and their corresponding AUCs were 0.71, 0.69, and 0.66, respectively. Based on the maximal Youden's index, we found that the best cutoff of the PI was 0.403, which was used to divide all cases into high risk (N = 62) and low risk group (N = 112). The analysis of risk score (PI) was displayed. The K-M curves


A
Outlier sample removed
Height
120
100
80
60
40
20
GSM711957
GSM711947
GSM711965
GSM711981
GSM711984
GSM711959
GSM711961
GSM711929
GSM711969
GSM711924
GSM711942
GSM711906
GSM711920
GSM711970
GSM711932
GSM711908
GSM711930
GSM711915
GSM711973
GSM711944
GSM711919
GSM711911
GSM711927
B
Scale independence
Scale Free Topology Model Fit, signed R^2
0.5
0
-0.5
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
0
5
10
15
20
Soft threshold (power)
C
Mean connectivity
Mean Connectivity
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
0
5
10
15
20
Soft threshold (power)
D
Cluster Dendrogram
Height
1.0
0.8
0.6
0.4
0.2
Colors

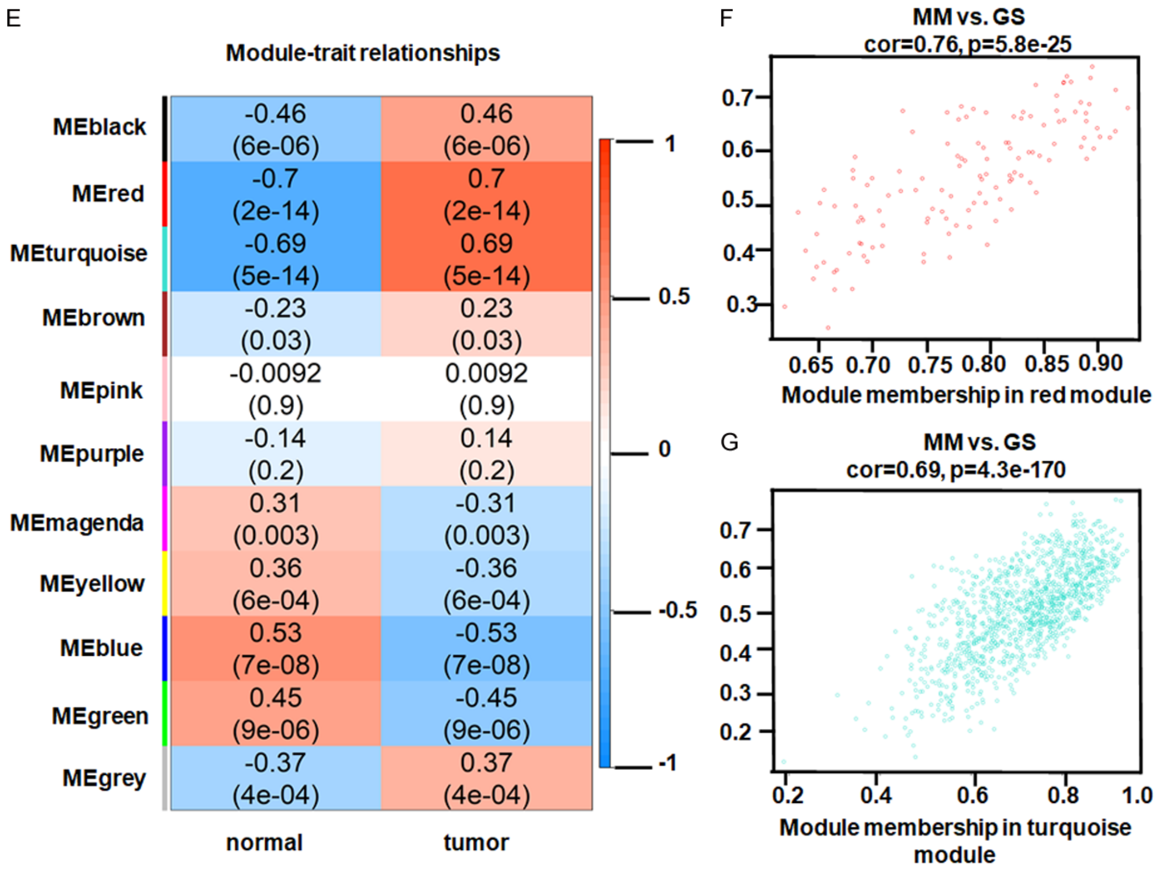


Figure 1. Identification of PDAC-related module genes by WGCNA analysis. A. Sample clustering dendrogram and a clear outlier sample (GSM711957) was removed. B. The optimal power was 12 when the red guide line was set to 0.89. C. The mean connectivity of different soft-threshold powers. D. Clustering of dendrogram and corresponding modules. E. Heatmap of the correlation between ME and GS. F, G. Dot plot of the correlation between MM and GS in the red and turquoise modules, respectively. ME, Module Eigengene; MM, Gene Module Membership; GS, Gene Significance.

indicated that the patients with low-risk scores had a significantly longer survival time (median: low risk 30.4 months vs. high risk 11.3 months, HR = 0.3723, 95% CI: 0.2461-0.5633, P = 1.133e-06) (**Figure 3A**). We used ICGC and GEO batch groups for further external validation of this 7-gene-based signature (**Figure 3B**, **3C**). The AUCs of the three datasets were all larger than 0.6, indicating that our 7-gene assay had an excellent performance at predicting OS. The K-M survival analysis returned the same result, in that the patients with low-risk score had a significantly longer survival time: ICGC (HR = 0.2813, 95% CI: 0.1643-0.4815, P = 9.133e-07) and GEO series (HR = 0.4999, 95% CI: 0.3407-0.7333, P = 0.000391).

Univariate and multivariate analyses of the 7-mRNA score model

To verify the model's predictive performance, we stratified the 7-mRNA score model by clinical factors: gender, age, and TNM stage in TCGA. First, we performed COX univariable analysis, which showed three variables associated with OS, including our 7-mRNA signature, Pathologic node, and age. Then, the multivariate modeling results further verified that the 7-mRNA risk score was indeed associated with OS. Additionally, the result implied that PI was the most significant prognostic-related factor, even including TNM stage and other clinical factors (Risk-low: HR = 0.410, 95% CI: 0.27-0.63, P < 0.0001; **Table 1**).

Stratification analysis based on clinicopathological characteristics

To further delineate the most suitable patient for the clinical application of the prognostic model, we evaluated the predictive ability of the 7-mRNA signature in different TNM stages, which is the most important clinicopathological characteristic affecting prognosis. As shown in

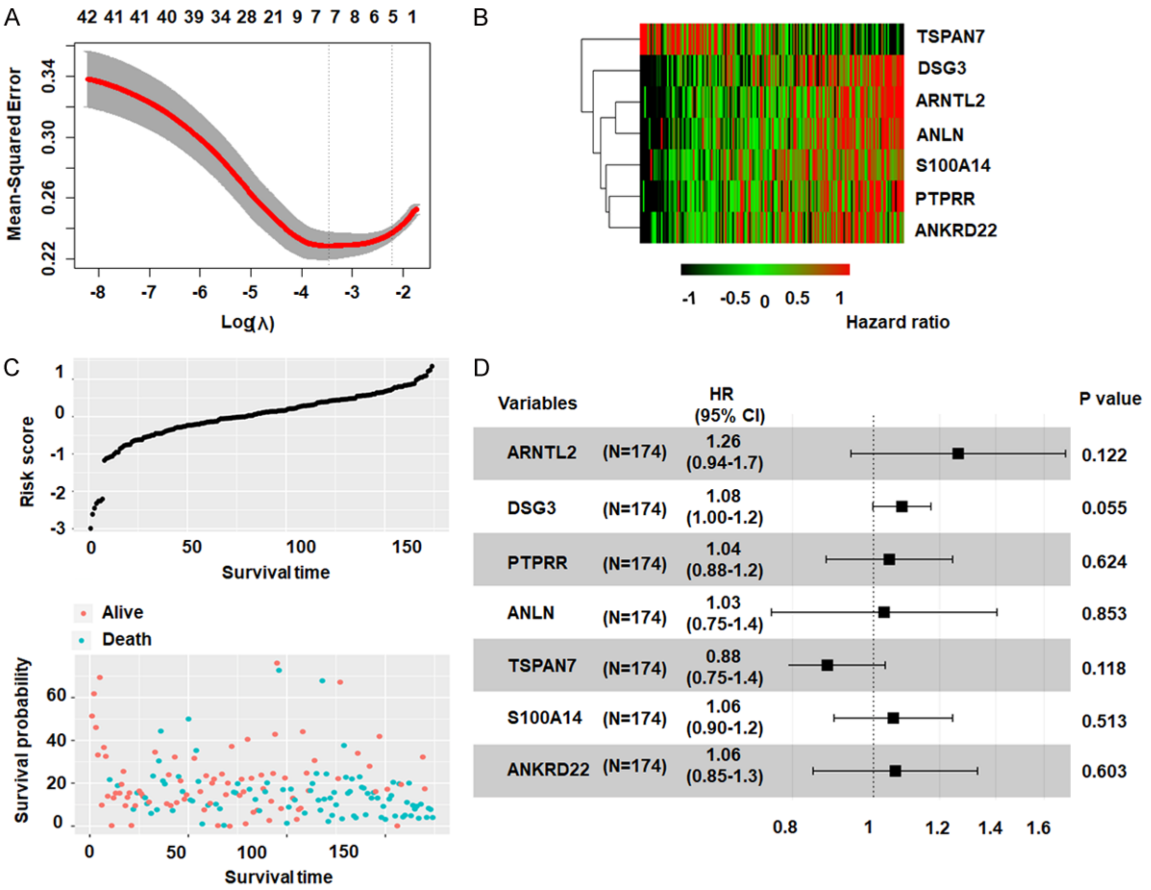


Figure 2. Seven-mRNAs in the prognostic model in the TCGA dataset. A. Cross-validation to find the optimal lambda in the LASSO regression. Here, the first dotted vertical line was adopted, with lambda = 0.031. B. The 7-signature expression heatmap in TCGA PDAC. C. Risk score analysis of the 7-gene-based signature. From top to bottom: risk score and patients' survival probability. D. Forest plot of multivariate Cox regression of the 7 mRNAs in the model.

Figure 4A, 167 cases of TCGA TNM stage I&II divided by the 7-mRNA score exhibited statistically significant differences in survival probability (HR = 0.5454, 95% CI: 0.3557-0.8363, P = 0.0048). Overall, 87 ICGC patients with TNM stage I&II also showed significant differences in low and high-risk groups (HR = 0.3314, 95% CI: 0.1913-0.5740, P = 3.6433e-05) (**Figure 4B**), and 111 TNM stage I&II patients in the GEO batch also verified the predictive ability of the 7-mRNA score (HR = 0.5315, 95% CI: 0.3291-0.8583, P = 0.0086) (**Figure 4C**). Because metastasis is the most important factor that affects patients' survival, we then further explore the roles of 7-gene in tumor metastasis in PDAC. As shown in **Figure 4D**, the association between 7-mRNA model expression in 47 cases of ICGC metastasis was analyzed by Kaplan-Meier survival analysis, and the results indicated patients with a high 7-mRNA signature score exhibited a significantly high risk of metastasis (HR = 0.1548, 95% CI: 0.0731-0.3277, P = 7.3018e-08).

Validation of expression level of 7 genes in clinical patients' tissue

Firstly, for the mRNA level of 7 genes, TSPAN7 is the unique one that decreased in PDAC, another 6-gene expression was notably increased in the tumor tissues than in the paired normal tissues (**Figure 5A**). Then, we examined the protein expression by western blot in 12 samples (4 paired of PDAC and adjacent non-tumor tissues, and 4 benign pancreatic tissues). In general, increased expression of ANLN, DSG3, PTPRR, ARNTL2, S100A14, and ANKRD22 protein was detected in PDAC compared to matched normal tissues. The TSPAN7 expression in tumors was very low (**Figure 5B**). Similarly, the immunoreactive score (IRS) of 7 genes in PDAC and normal tissues by IHC staning was consistent with the mRNA results

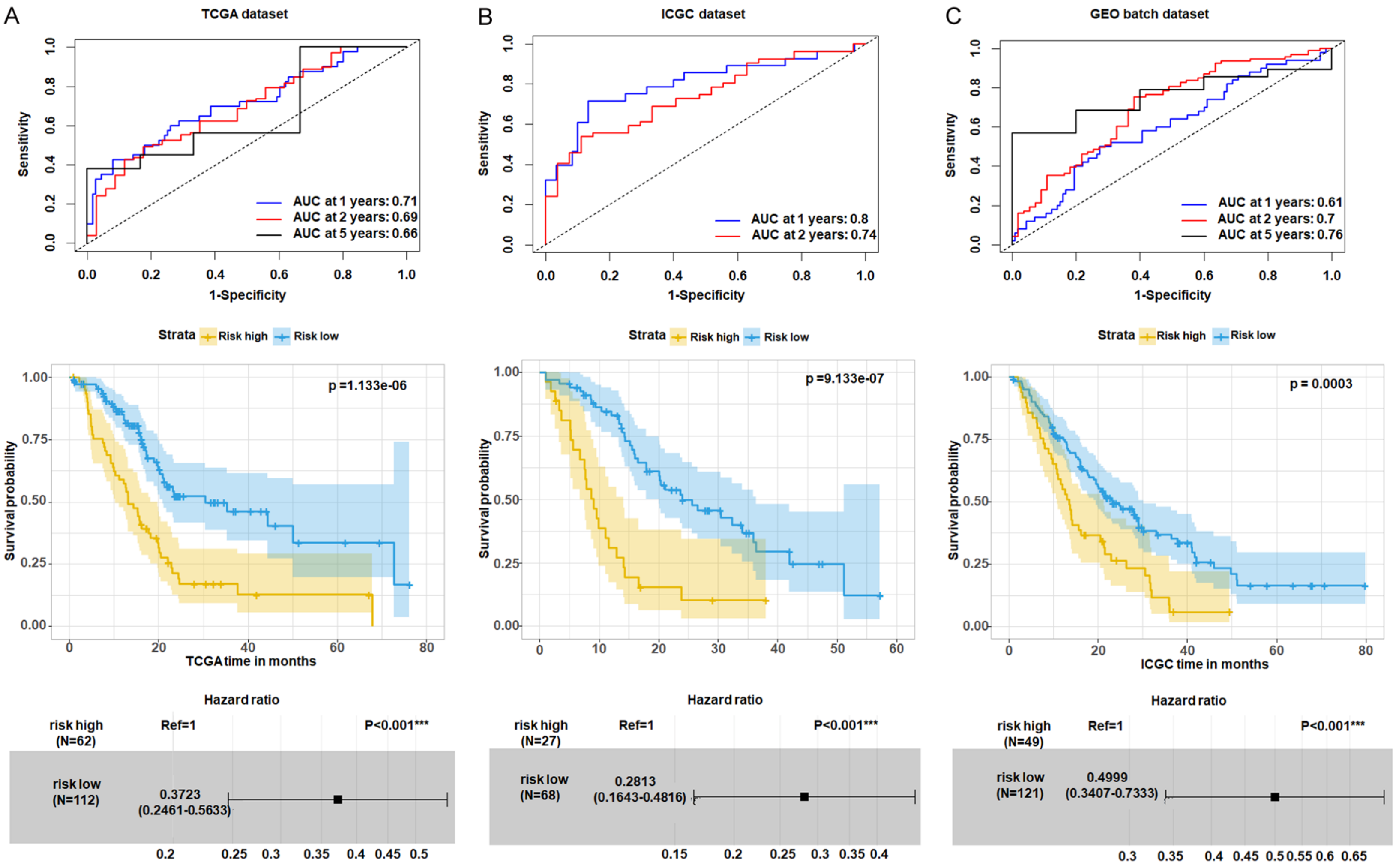


Figure 3. Comparison of the 7-mRNA model in the TCGA, ICGC, and GEO batch databases. Survival prediction of ROC curves; Kaplan-Meier curves of low-risk and high-risk score groups; and results of PI's HR, 95% CI, and *p*-value in the TCGA (A) ICGC (B), and GEO batch datasets (C), which showed similar significant results.

Table 1. Univariable and multivariable Cox regression of the 7-mRNA risk score, TNM, stage, age, and gender on 174 TCGA PDAC dataset

variable		Univariable Cox				Multiple variable Cox			
		Coefficient	HR	95% CI	*P* value	Coefficient	HR	95% CI	*P* value
Ref	risk high	-	1.00	-	-	-	1.00	-	-
	risk low	-0.990	0.370	0.25-0.56	< 0.0001	-0.900	0.410	0.27-0.63	< 0.0001
Ref	T1		1.00	-	-	-	1.00	-	-
	T2	-0.150	0.860	0.19-4	0.848	-0.230	0.800	0.14-4.4	0.793
	T3	0.490	1.630	0.4-6.65	0.495	0.030	1.030	0.15-7.26	0.978
	T4	-0.140	0.870	0.08-9.62	0.910	-0.120	0.890	0.08-10.07	0.922
Ref	N0	-	1.00	-	-	-	1.00	-	-
	N1	0.720	2.060	1.23-3.46	0.006	0.490	1.640	0.88-3.04	0.121
	NX	0.530	1.710	0.39-7.56	0.482	1.050	2.850	0.55-14.75	0.211
Ref	M0	-	1.00	-	-	-	1.00	-	-
	M1	-0.390	0.680	0.16-2.82	0.595	-0.140	0.870	0.12-6.49	0.895
	MX	-0.120	0.890	0.58-1.35	0.579	-0.040	0.960	0.62-1.47	0.838
Ref	stage i	-	1.00	-	-	-	1.00	-	-
	Stage ii	0.700	2.020	0.93-4.39	0.077	0.060	1.070	0.24-4.75	0.933
	Stage iii	0.090	1.090	0.13-8.91	0.936	-		-	-
	Stage iv	0.300	1.360	0.28-6.6	0.706	-		-	-
age		0.020	1.020	1-1.05	0.031	0.020	1.020	1-1.05	0.037
Ref	female	-	1.00	-	-	-	1.00	-	-
	male	-0.220	0.800	0.53-1.21	0.298	-0.060	0.950	0.62-1.44	0.795

(**Figure 5C**). Notably, we detected the positive expression of ANKRD22 in almost all tumor tissues and negative expression in non-tumor tissues. Moreover, TSPAN7 was only expressed in the benign pancreas (**Figure 5D**).

Results of GO function enrichment, KEGG and GSEA pathway analysis

To explore the potential biological processes and pathways involved in PDAC progression, we performed GO, KEGG, and GSEA analysis on DEGs of GSE28735. As shown in **Figure 6**, extracellular matrix and structure organization were the main Biological process (BP) terms. For the 256 overexpressed genes, the majority of terms were related to adhesion, collagen binding, metastasis, and matrix activity in molecular function (MF), and extracellular matrix (ECM) was enriched in cellular components (CC) (**Figure 6A**). For the 157 down-regulated genes, multicellular organismal homeostasis and serine-type peptidase activity were the main terms of BP and MF, respectively, and platelet alpha granule was enriched in CC (**Figure 6B**). Four KEGG pathways were enriched by upregulated genes, including ECM-receptor interaction, focal adhesion and PI3K-AKT signaling pathway. Meanwhile, the downregulated genes were mainly enriched in complement and coagulation cascade pathways (**Figure 6C**). As revealed in the GESA pathway analysis, among the 16 upregulated pathways, the PI3K-AKT signaling pathway was the most significantly enriched pathway (**Figure 6D**). The integrating results of GO, KEGG, and GSEA analysis were consistent with our expectation that ECM remodeling and collagen adhesion may promote tumor development and metastasis.

Discussion

We integrated and analyzed multiple databases to characterize the gene expression profile of PDAC tissues and the potential mechanisms underlying PDAC progression. GSE28735 is a commonly used dataset in pancreatic cancer, and many scholars have obtained DEGs of tumors and adjacent non-tumor for follow-up research, such as analysis of biological information, biological function, or molecular mechanism. However, DEGs alone may not be sufficient to accurately screen out genes for pancreatic cancer diagnosis, prognosis, or targeted therapy. Therefore, we combined three methods for gene signature screening, which

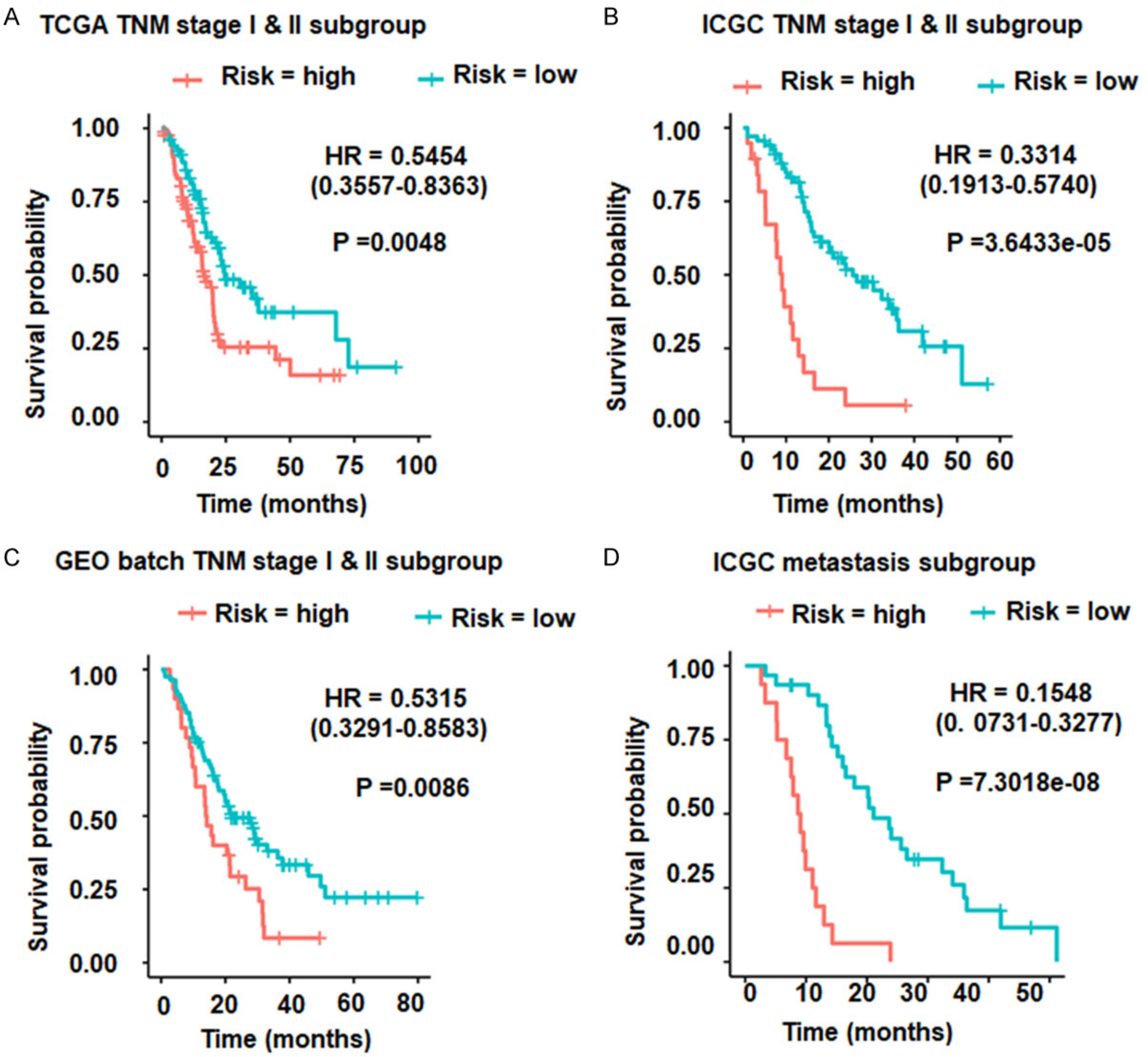


Figure 4. Kaplan-Meier curve to evaluate the 7-mRNA-assay's predictive ability in 167 TCGA (A), 92 ICGC (B), and 111 GEO batch group (C) of TNM stage I&II PDAC patients, and 47 ICGC metastasis cohorts (D). The green and red curves represent low-risk and high-risk groups, respectively. The differences between them were tested by the two-sided log-rank method.

should meet the requirements: (i) differences in expression between tumors and adjacent non-tumor achieved through the DEG method; (ii) a cluster of genes closely related to pancreatic cancer acquired by WGCNA analysis; (iii) possible functional genes in PDAC, which reminds us of microRNAs and their target genes, a hot topic in the field of cancer research.

miRNAs play important roles in diseases by regulating target genes and have become biomarkers for diagnosis and prognosis in cancers [25]. We found that a portion of the 1881 microRNAs was only detected in individual pancreatic cancer tissues. Therefore, we screened microRNAs that were expressed in at least 10 pancreatic cancer specimens for survival analysis. In total, 10 miRNAs were screened and had a significant effect on the prognosis of patients with pancreatic cancer. Because miRNAs exert biofunctions via their modulation of target genes, we supposed that coding gene mRNA expression levels might more tightly and precisely correlate with cancer progression. From 12425 target genes, we predicted 10 miRNAs using the miRwalk database. Among those, 163 genes were selected through overlapping DEG, PDAC-related genes of WGCNA analysis, and prognostic-related miRNA target genes.

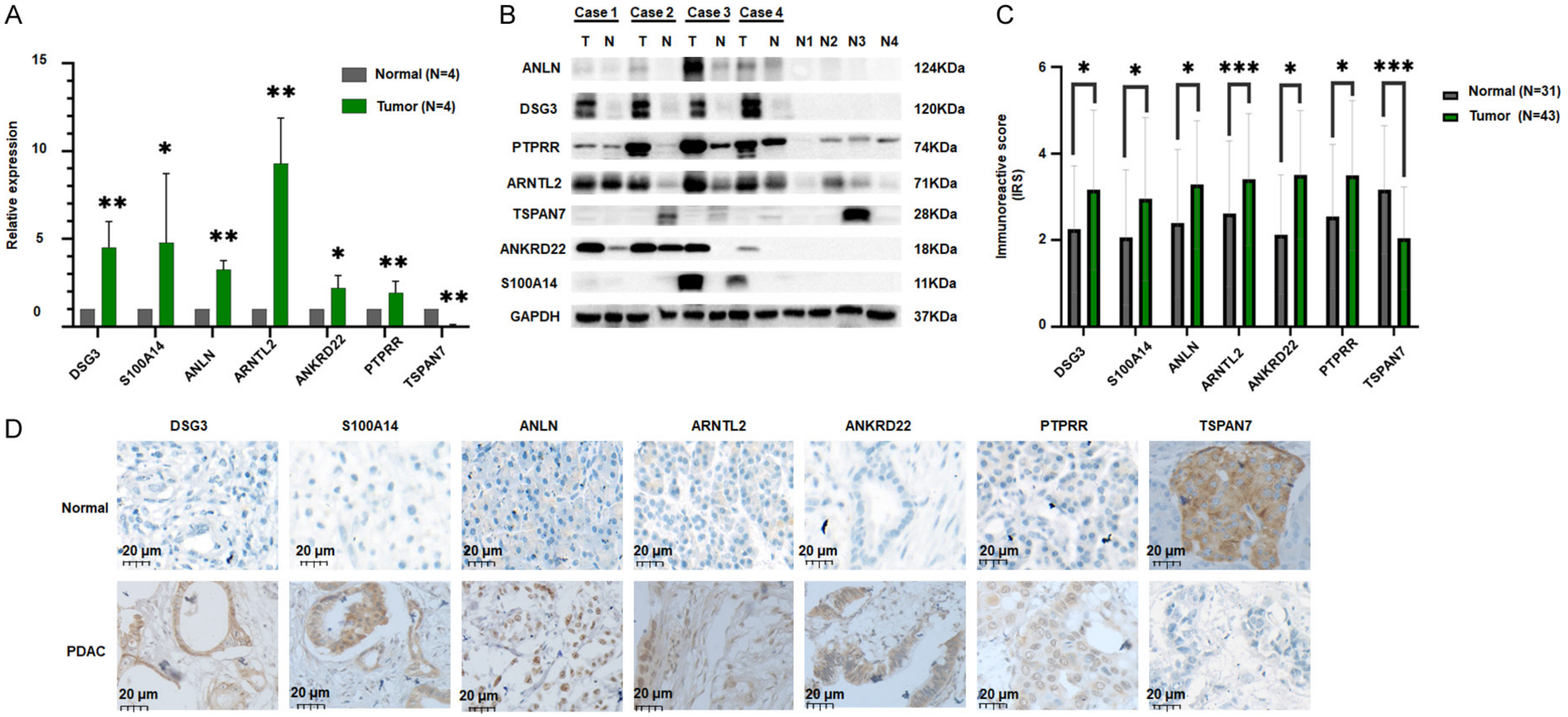
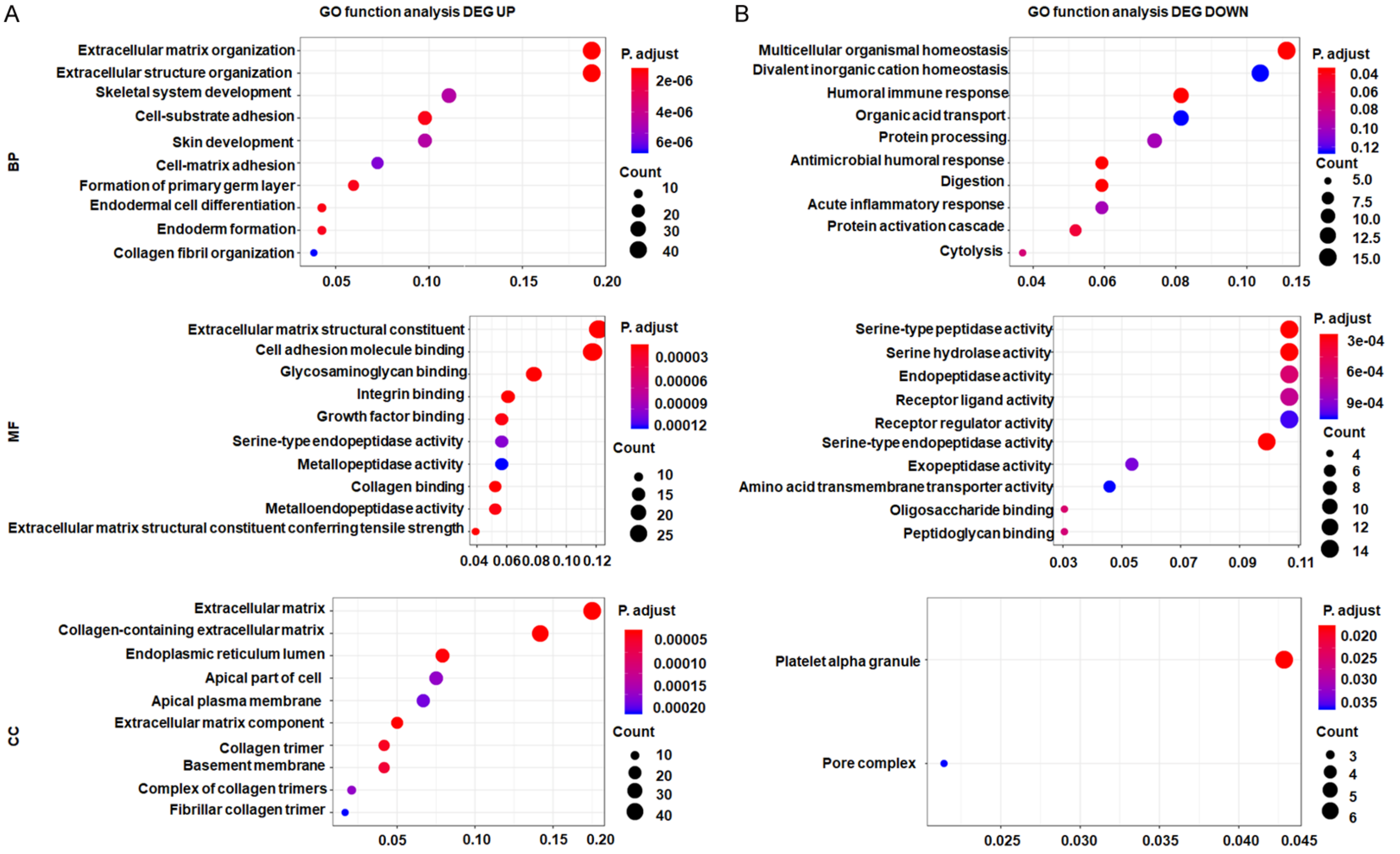


Figure 5. The 7 gene expression characteristics were observed in pancreatic cancer tissues. A. The 7 genes mRNA expression level was detected by RT-PCR, and β-actin was used as an internal control. B. Western blot assay was used to detect the protein level of the 7 genes in 12 tissue samples. GAPDH was used as an internal control. Case 1-4, 4 paired of PDAC and adjacent non-tumor samples. N, non-tumor tissue. T, tumor tissue. C. Immunoreactive score (IRS) of the 7 genes in PDAC samples and normal tissues. IRS = (percentage of cells of staining score) + (staining intensity score). A total score of 0-6 was evaluated. D. Representative IHC images of the 7 genes expression in PDAC and non-tumor tissues (400 × magnification).


A
GO function analysis DEG UP
BP
Extracellular matrix organization
Extracellular structure organization
Skeletal system development
Cell-substrate adhesion
Skin development
Cell-matrix adhesion
Formation of primary germ layer
Endodermal cell differentiation
Endoderm formation
Collagen fibril organization
0.05
0.10
0.15
0.20
P. adjust
2e-06
4e-06
6e-06
Count
10
20
30
40
MF
Extracellular matrix structural constituent
Cell adhesion molecule binding
Glycosaminoglycan binding
Integrin binding
Growth factor binding
Serine-type endopeptidase activity
Metallopeptidase activity
Collagen binding
Metalloendopeptidase activity
Extracellular matrix structural constituent conferring tensile strength
0.04 0.06 0.08 0.10 0.12
P. adjust
0.00003
0.00006
0.00009
0.00012
Count
10
15
20
25
CC
Extracellular matrix
Collagen-containing extracellular matrix
Endoplasmic reticulum lumen
Apical part of cell
Apical plasma membrane
Extracellular matrix component
Collagen trimer
Basement membrane
Complex of collagen trimers
Fibrillar collagen trimer
0.05
0.10
0.15
0.20
P. adjust
0.00005
0.00010
0.00015
0.00020
Count
10
20
30
40
B
GO function analysis DEG DOWN
Multicellular organismal homeostasis
Divalent inorganic cation homeostasis
Humoral immune response
Organic acid transport
Protein processing
Antimicrobial humoral response
Digestion
Acute inflammatory response
Protein activation cascade
Cytolysis
0.04
0.06
0.08
0.10
0.15
P. adjust
0.04
0.06
0.08
0.10
0.12
Count
5.0
7.5
10.0
12.5
15.0
Serine-type peptidase activity
Serine hydrolase activity
Endopeptidase activity
Receptor ligand activity
Receptor regulator activity
Serine-type endopeptidase activity
Exopeptidase activity
Amino acid transmembrane transporter activity
Oligosaccharide binding
Peptidoglycan binding
0.03
0.05
0.07
0.09
0.11
P. adjust
3e-04
6e-04
9e-04
Count
4
6
8
10
12
14
Platelet alpha granule
Pore complex
0.025
0.030
0.035
0.040
0.045
P. adjust
0.020
0.025
0.030
0.035
Count
3
4
5
6

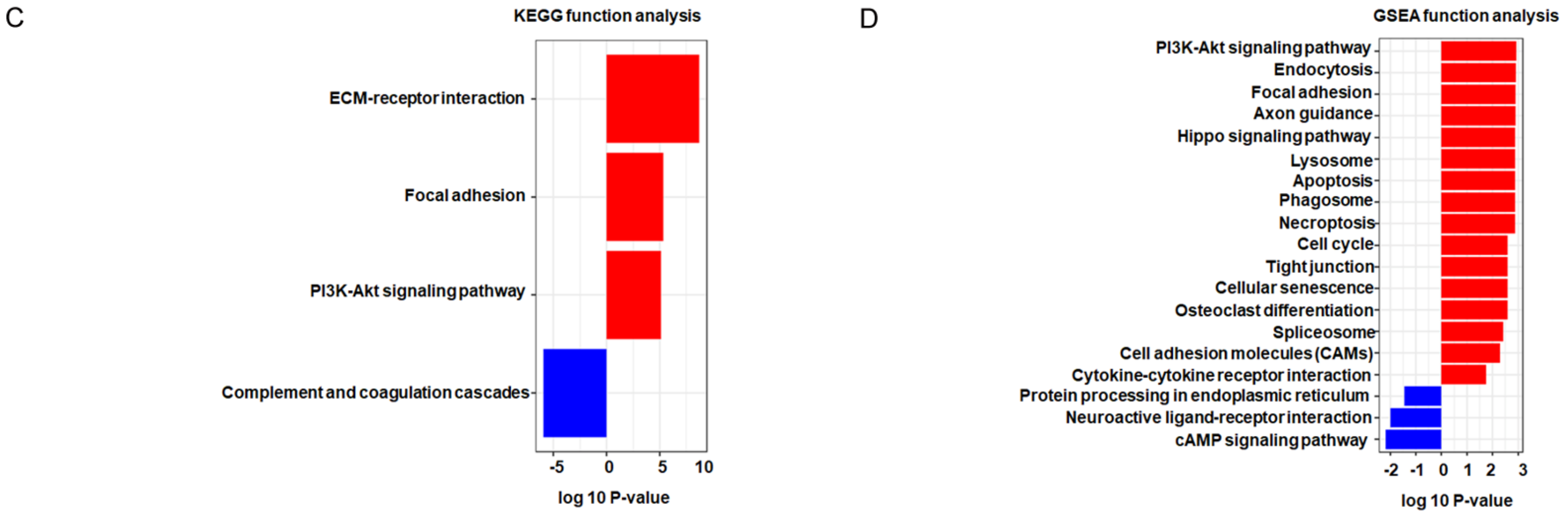


Figure 6. Top 10 terms of functional enrichment of 413 DEGs. (A) GO function analysis of upregulated DEGs. (B) GO function analysis of downregulated DEGs. Signal pathway enrichment of 413 DEGs in KEGG (C) and GSEA (D). The red bar represents upregulated pathways, and the blue bar represents downregulated pathways.

Among these 163 genes, we used univariate Cox and LASSO regression analysis to define 7 non-zero coefficient genes that are tightly related to OS and established the prognostic prediction model, thereafter named the 7-gene survival score assay of PDAC. It is worth noting that, even though combined with clinical factors, including grade stage, TNM stage, and age, our 7-mRNA signature still proved to be a powerful independent predictive indicator. In addition, the good validation of the prediction model in the ICGC testing dataset and GEO batch dataset indicate the potential effectiveness for clinical application.

To explore the potential mechanism of the 7 genes in PDAC, we first learned about the major overexpressed DEGs in pancreatic cancer enriched in extracellular matrix structural constituents, cell adhesion molecule binding, glycosaminoglycan binding, and integrin binding. These functions are closely related to tumor metastasis [26]. Distance dissemination occurs in 90% of PDAC cases and is the main cause of death [27]. Some of the signaling pathways suggested in the KEGG and GSEA pathway enrichment have been confirmed in previous studies, such as the PI3K-Akt signaling pathway [28].

Among these 7 genes, ARNTL2, DSG3, PTPRR, ANLN, S100A14, and ANKRD22 were significantly overexpressed in PDAC tissues. ARNTL2 is an important clock gene often affected by human immune-inflammatory conditions and involved in type 1 diabetes [29]. It also has been described as a candidate gene for cancer metastasis [30]. PTPRR expression was found to be inversely correlated with disease prognosis by suppressing activation of the Wnt/β-catenin pathway in ovarian cancer [31]. ANLN and S100A14 play pivotal roles in epithelial mesenchymal-transformation (EMT) of cancers [32, 33]. ANKRD22 has been reported to enhance cell cycle progression in non-small cell lung cancer by upregulating the expression of E2F1 [34]. Another important function of ANKRD22 is its role in glucose metabolism, which we will discuss in the following section. Desmoglein 3 (Dsg3) exerts pro-survival activity on kidney epithelial cells by suppressing reactive oxygen species beyond the traditional desmosomal adhesion and tissue integrity [35]. Meanwhile, its role in cancer progression is still controversial [36]. TSPAN7 is downregulated in PDAC tissues and is favorable to patient prognosis. The exact biofunctions and molecular mechanism of TSPAN7 in PDAC are unclear; however, studies in mice bearing myelomas have indicated that TSPAN7 acts as a tumor suppresser [37].

Recently, a series of PDAC prognostic gene signatures were found. Chen et al. reported a 15-gene prognostic model using a PDAC cohort from Moffitt Cancer Center and Stratford microarray cohort dataset [38]. Wu et al. established a nine-gene signature (MET, KLK10, COL17A1, CEP55, ANKRD22, ITGB6, ARNTL2, MCOLN3, and SLC25A45) associated with PDAC prognosis [39]. Zhou et al. identified a 6-mRNA signature (KYNU, MET, INPP4B, IGF2BP3, ANKRD22, and TOP2A) using a similar public dataset (high risk group in TCGA: HR = 2.86, 95% CI: 1.89-4.32, P < 0.0001) [40]. Caba O et al. identified a four-gene predictor set (ANKRD22, CLEC4D, VNN1, and IRAK3) by transcriptional profiling of peripheral blood in 18 pancreatic adenocarcinoma patients and 18 controls [41]. Compared to the previously defined signatures, our 7-gene signature provides a precise and superior prediction model: (i) multiple statistical methods were combined for target gene selection, which found 163 filtered genes closely related to the prognosis and biological function of pancreatic cancer patients; (ii) multiple prediction indicators, including AUC values under the ROC curve, K-M plot and HR results, proved that our 7-gene model is a powerful predictor of OS; (iii) multi-platform PDAC data, including microarray, miRNA, and transcriptomics from GEO, TCGA, and ICGC provided consistent results, which implies that the 7-mRNA model is an efficient prediction model for PDAC.

Noticeably, there was one gene (ANKRD22) in common among our 7-mRNA model and the models of Wu et al., Zhou et al., and Caba O et al. A recent study showed that ANKRD22 promotes glycolysis by promoting the metabolic reprogramming process in colorectal cancer cells [42], which reminds us that severe metabolic stress occurs in PDAC because of desmoplasia and poor vascularity [43]. ANKRD22 might be involved in the special glucose mechanisms that maintain a high glycolysis rate in PDAC, a process well known as the Warburg effect, a hallmark of PDAC [44]. Besides databases, the results of clinical samples confirmed the overexpression of ANKRD22 in PDAC. We suppose that ANKRD22 is a new potential therapeutic target for PDAC.

In the present study, we carefully removed the batch effects between different GEO platforms and used the statistical method to select the best PI cutoff rather than an arbitrary median cutoff. However, our research still has some limitations. First, we could not explore or validate our 7-mRNA model with TNM or grade stage due to the lack of clinical information of most GEO series. Second, we need to collect sufficient clinical cases, including tissue samples and clinicopathological information, to validate the retrospective predictive efficacy of the 7-gene survival score assay. Third, and most importantly, we need further experiments in prospective PDAC cohorts to validate the prognostic classification and whether it is possible to predict benefits from chemotherapy. Further research is needed to explore their functions in PDAC and to identify whether these 7 genes are promising therapeutic targets.

Conclusion

In summary, we established a new 7-mRNA signature validated in multiple databases. It may be valuable for clinical applications for prognostic predictions in TNM stage I&II PDAC patients.

Acknowledgements

This work was supported by National Natural Science Foundation of China (grant no. 8207-3209, 81872348), Natural Science Foundation of Guangdong Province, China (grant no. 2020A151501808). The Fundamental Research Funds for the Central Universities (grant no. 19ykzd08).

Disclosure of conflict of interest

None.

Address correspondence to: Rongzhi Xie and Linjuan Zeng, Department of Abdominal Oncology, The Cancer Center of The Fifth Affiliated Hospital, Sun Yat-sen University, 52 Mei Hua East Road, Zhuhai 519000, Guangdong, China. Tel: +86-756-252-8027 (Office); Fax: +86-756-252-8166; E-mail: xierzhi@mail.sysu.edu.cn (RZX); zenglinj@mail.sysu.edu.cn (LJZ)

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and 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: 394-424.

[2] Siegel RL, Miller KD and Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019; 69: 7-34.

[3] Barnes CA, Chavez MI, Tsai S, Aldakkak M, George B, Ritch PS, Dua K, Clarke CN, Tolat P, Hagen C, Hall WA, Erickson BA, Evans DB and Christians KK. Survival of patients with borderline resectable pancreatic cancer who received neoadjuvant therapy and surgery. Surgery 2019; 166: 277-285.

[4] Ma SJ, Oladeru OT, Miccio JA, Iovoli AJ, Hermann GM and Singh AK. Association of timing of adjuvant therapy with survival in patients with resected stage I to II pancreatic cancer. JAMA Netw Open 2019; 2: e199126.

[5] Conroy T, Desseigne F, Ychou M, Bouche O, Guimbaud R, Becouarn Y, Adenis A, Raoul JL, Gourgou-Bourgade S, de la Fouchardiere C, Bennouna J, Bachet JB, Khemissa-Akouz F, Pere-Verge D, Delbaldo C, Assenat E, Chauffert B, Michel P, Montoto-Grillot C and Ducreux M. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. N Engl J Med 2011; 364: 1817-1825.

[6] Friedenson B. BRCA1 and BRCA2 pathways and the risk of cancers other than breast or ovarian. MedGenMed 2005; 7: 60.

[7] Golan T, Kindler HL, Park JO, Reni M, Mercade TM and Hammel P. Geographic and ethnic heterogeneity in the BRCA1/2 pre-screening population for the randomized phase III POLO study of olaparib maintenance in metastatic pancreatic cancer (mPC). J Clin Oncol 2018; 36: 4115.

[8] Moore MJ, Goldstein D, Hamm J, Figer A, Hecht JR, Gallinger S, Au HJ, Murawa P, Walde D, Wolff RA, Campos D, Lim R, Ding K, Clark G, Voskoglou-Nomikos T, Ptasynski M and Parulekar W. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. J Clin Oncol 2007; 25: 1960-1966.

[9] Kindler HL, Niedzwiecki D, Hollis D, Sutherland S, Schrag D, Hurwitz H, Innocenti F, Mulcahy MF, O'Reilly E, Wozniak TF, Picus J, Bhargava P, Mayer RJ, Schilsky RL and Goldberg RM. Gemcitabine plus bevacizumab compared with gemcitabine plus placebo in patients with advanced pancreatic cancer: phase III trial of the Cancer and Leukemia Group B (CALGB 80303). J Clin Oncol 2010; 28: 3617-3622.

[10] Philip PA, Benedetti J, Corless CL, Wong R, O'Reilly EM, Flynn PJ, Rowland KM, Atkins JN, Mirtsching BC, Rivkin SE, Khorana AA, Goldman B, Fenoglio-Preiser CM, Abbruzzese JL and Blanke CD. Phase III study comparing gemcitabine plus cetuximab versus gemcitabine in patients with advanced pancreatic adenocarcinoma: Southwest Oncology Group-directed intergroup trial S0205. J Clin Oncol 2010; 28: 3605-3610.

[11] Chen Y, Gao SG, Chen JM, Wang GP, Wang ZF, Zhou B, Jin CH, Yang YT and Feng XS. Serum CA242, CA199, CA125, CEA, and TSGF are biomarkers for the efficacy and prognosis of cryoablation in pancreatic cancer patients. Cell Biochem Biophys 2015; 71: 1287-1291.

[12] Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J and Lopez R. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res 2010; 38: W695-699.

[13] Qi L, Yao Y, Zhang T, Feng F, Zhou C, Xu X and Sun C. A four-mRNA model to improve the prediction of breast cancer prognosis. Gene 2019; 721: 144100.

[14] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.

[15] Langfelder P and Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.

[16] Langfelder P and Horvath S. Fast R functions for robust correlations and hierarchical clustering. J Stat Softw 2012; 46: i11.

[17] Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh IT, Ravdin P, Bugarini R, Baehner FL, Davidson NE, Sledge GW, Winer EP, Hudis C, Ingle JN, Perez EA, Pritchard KI, Shepherd L, Gralow JR, Yoshizawa C, Allred DC, Osborne CK and Hayes DF. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. Lancet Oncol 2010; 11: 55-65.

[18] Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S and Geschwind DH. Functional organization of the transcriptome in human brain. Nat Neurosci 2008; 11: 1271-1282.

[19] Giulietti M, Occhipinti G, Principato G and Piva F. Weighted gene co-expression network analysis reveals key genes involved in pancreatic ductal adenocarcinoma development. Cell Oncol (Dordr) 2016; 39: 379-388.

[20] Song J, Xu Q, Zhang H, Yin X, Zhu C, Zhao K and Zhu J. Five key lncRNAs considered as prognostic targets for predicting pancreatic ductal adenocarcinoma. J Cell Biochem 2018; 119: 4559-4569.

[21] Zhang J, Bing Z, Yan P, Tian J, Shi X, Wang Y and Yang K. Identification of 17 mRNAs and a miRNA as an integrated prognostic signature for lung squamous cell carcinoma. J Gene Med 2019; 21: e3105.

[22] Goh TS, Ha M, Lee JS, Jeong DC, Jung ES, Han ME, Kim YH and Oh SO. Prognostic significance of EIF4G1 in patients with pancreatic ductal adenocarcinoma. Onco Targets Ther 2019; 12: 2853-2859.

[23] Su Q, Zhu EC, Qu YL, Wang DY, Qu WW, Zhang CG, Wu T and Gao ZH. Serum level of co-expressed hub miRNAs as diagnostic and prognostic biomarkers for pancreatic ductal adenocarcinoma. J Cancer 2018; 9: 3991-3999.

[24] Frampton AE, Castellano L, Colombo T, Giovannetti E, Krell J, Jacob J, Pellegrino L, Roca-Alonso L, Funel N, Gall TM, Ahmad R, Habib NA, Knosel T, Stebbing J and Jiao LR. Integrated molecular analysis to investigate the role of microRNAs in pancreatic tumour growth and progression. Lancet 2015; 385 Suppl 1: S37.

[25] Carter JV, Galbraith NJ, Yang D, Burton JF, Walker SP and Galandiuk S. Blood-based microRNAs as biomarkers for the diagnosis of colorectal cancer: a systematic review and meta-analysis. Br J Cancer 2017; 116: 762-774.

[26] Du Q, Wang W, Liu T, Shang C, Huang J, Liao Y, Qin S, Chen Y, Liu P, Liu J and Yao S. High expression of integrin alpha3 predicts poor prognosis and promotes tumor metastasis and angiogenesis by activating the c-Src/extracellular signal-regulated protein kinase/focal adhesion kinase signaling pathway in cervical cancer. Front Oncol 2020; 10: 36.

[27] Heeke S, Mograbi B, Alix-Panabieres C and Hofman P. Never travel alone: the crosstalk of circulating tumor cells and the blood microenvironment. Cells 2019; 8: 714.

[28] Hoxhaj G and Manning BD. The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism. Nat Rev Cancer 2020; 20: 74-88.

[29] Lebailly B, Boitard C and Rogner UC. Circadian rhythm-related genes: implication in autoimmunity and type 1 diabetes. Diabetes Obes Metab 2015; 17 Suppl 1: 134-138.

[30] Brady JJ, Chuang CH, Greenside PG, Rogers ZN, Murray CW, Caswell DR, Hartmann U, Connolly AJ, Sweet-Cordero EA, Kundaje A and Winslow MM. An arntl2-driven secretome enables lung adenocarcinoma metastatic self-sufficiency. Cancer Cell 2016; 29: 697-710.

[31] Wang Y, Cao J, Liu W, Zhang J, Wang Z, Zhang Y, Hou L, Chen S, Hao P, Zhang L, Zhuang M, Yu

Y, Li D and Fan G. Protein tyrosine phosphatase receptor type R (PTPRR) antagonizes the Wnt signaling pathway in ovarian cancer by dephosphorylating and inactivating beta-catenin. J Biol Chem 2019; 294: 18306-18323.

[32] Xu J, Zheng H, Yuan S, Zhou B, Zhao W, Pan Y and Qi D. Overexpression of ANLN in lung adenocarcinoma is associated with metastasis. Thorac Cancer 2019; 10: 1702-1709.

[33] Al-Ismaeel Q, Neal CP, Al-Mahmoodi H, Almutairi Z, Al-Shamarti I, Straatman K, Jaunbocus N, Irvine A, Issa E, Moreman C, Dennison AR, Emre Sayan A, McDearmid J, Greaves P, Tulchinsky E and Kriajevska M. ZEB1 and IL-6/11-STAT3 signalling cooperate to define invasive potential of pancreatic cancer cells via differential regulation of the expression of S100 proteins. Br J Cancer 2019; 121: 65-75.

[34] Yin J, Fu W, Dai L, Jiang Z, Liao H, Chen W, Pan L and Zhao J. ANKRD22 promotes progression of non-small cell lung cancer through transcriptional up-regulation of E2F1. Sci Rep 2017; 7: 4430.

[35] Li X, Ahmad US, Huang Y, Uttagomol J, Rehman A, Zhou K, Warnes G, McArthur S, Parkinson EK and Wan H. Desmoglein-3 acts as a pro-survival protein by suppressing reactive oxygen species and doming whilst augmenting the tight junctions in MDCK cells. Mech Ageing Dev 2019; 184: 111174.

[36] Brown L and Wan H. Desmoglein 3: a help or a hindrance in cancer progression? Cancers (Basel) 2015; 7: 266-286.

[37] Cheong CM, Chow AW, Fitter S, Hewett DR, Martin SK, Williams SA, To LB, Zannettino AC and Vandyke K. Tetraspanin 7 (TSPAN7) expression is upregulated in multiple myeloma patients and inhibits myeloma tumour development in vivo. Exp Cell Res 2015; 332: 24-38.

[38] Chen DT, Davis-Yadley AH, Huang PY, Husain K, Centeno BA, Permuth-Wey J, Pimiento JM and Malafa M. Prognostic fifteen-gene signature for early stage pancreatic ductal adenocarcinoma. PLoS One 2015; 10: e0133562.

[39] Wu M, Li X, Zhang T, Liu Z and Zhao Y. Identification of a nine-gene signature and establishment of a prognostic nomogram predicting overall survival of pancreatic cancer. Front Oncol 2019; 9: 996.

[40] Zhou C, Zhao Y, Yin Y, Hu Z, Atyah M, Chen W, Meng Z, Mao H, Zhou Q, Tang W, Wang P, Li Z, Weng J, Bruns C, Popp M, Popp F, Dong Q and Ren N. A robust 6-mRNA signature for prognosis prediction of pancreatic ductal adenocarcinoma. Int J Biol Sci 2019; 15: 2282-2295.

[41] Caba O, Prados J, Ortiz R, Jimenez-Luna C, Melguizo C, Alvarez PJ, Delgado JR, Irigoyen A, Rojas I, Perez-Florido J, Torres C, Perales S, Linares A and Aranega A. Transcriptional profiling of peripheral blood in pancreatic adenocarcinoma patients identifies diagnostic biomarkers. Dig Dis Sci 2014; 59: 2714-2720.

[42] Pan T, Liu J, Xu S, Yu Q, Wang H, Sun H, Wu J, Zhu Y, Zhou J and Zhu Y. ANKRD22, a novel tumor microenvironment-induced mitochondrial protein promotes metabolic reprogramming of colorectal cancer cells. Theranostics 2020; 10: 516-536.

[43] Yan L, Raj P, Yao W and Ying H. Glucose metabolism in pancreatic cancer. Cancers (Basel) 2019; 11: 1460.

[44] Yang J, Ren B, Yang G, Wang H, Chen G, You L, Zhang T and Zhao Y. The enhancement of glycolysis regulates pancreatic cancer metastasis. Cell Mol Life Sci 2020; 77: 305-321.

Table S1. Sequences of primers used in this study

Name	Sequence (5'-3')
ARNTL2 forward	ACTTGGTGCTGGTAGTATTGGA
ARNTL2 reverse	TGTTGGACTCGAATCATCAAGG
PTPRR forward	ACCTATCGCCCATCACATTACA
PTPRR reverse	GCGGTGGTAGCTTTGATCTCA
DSG3 forward	GCAAAAACGTGAATGGGTGAAA
DSG3 reverse	TCCAGAGATTCGGTAGGTGATT
ANLN forward	ATGTCTTCGTGGCCGATTTGA
ANLN reverse	CTCTGACAGTGAGTTTCCTGTTT
S100A14 forward	GAGACGCTGACCCCTTCTG
S100A14 reverse	CTTGGCCGCTTCTCCAATCA
ANKRD22 forward	AGGGCATGTGAGAATCGTTTC
ANKRD22 reverse	GTAGCATTCGTACAAGAGCCTC
TSPAN7 forward	ACCAAACCTGTGATAACCTGTCT
TSPAN7 reverse	AGGGAGATATAGGTGCCCAGA
GAPDH forward	GAAATCCCATCACCATCTTCCAGG
GAPDH reverse	GAGCCCCAGCCTTCTCCATG

Table S2. Enrolled datasets of PDACs in this study

Dataset	Normal	Tumor	Patients with survival information	Patients with stage information	Platform
GSE28735	45	45	42	0	Affymetrix Gene 1.0 ST
GSE62452	61	69	65	65	Affymetrix Gene 1.0 ST
GSE57495	0	63	63	63	Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray
TCGA	4	178	178	178	Illumina HiSeq V2
ICGC	0	95	95	92	Illumina HiSeq

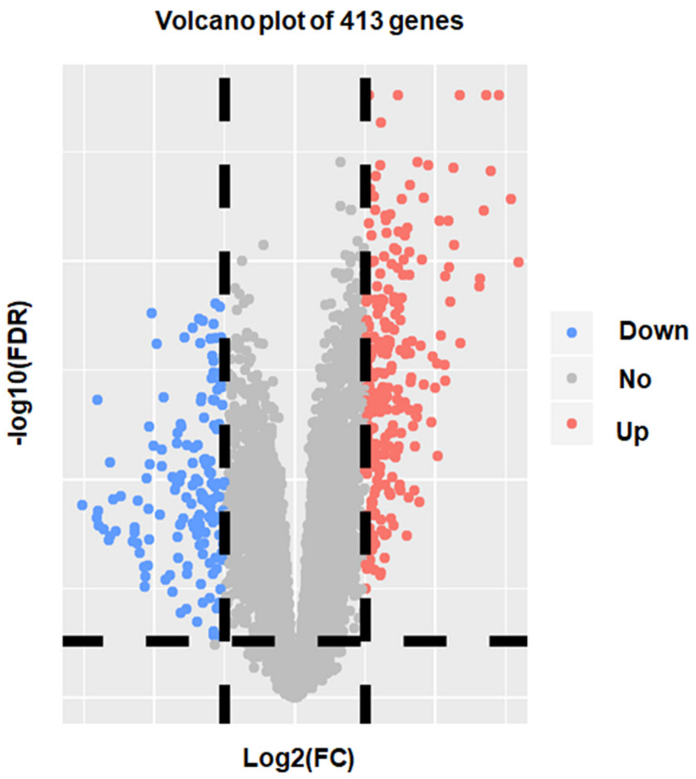


Figure S1. Volcano plot of differentially expressed genes (DEGs) in GSE28735. Genes with FDR < 0.05 and |log2 fold change (FC)| ≥ 1 are considered DEGs. Red, upregulated genes; Gray, non-differential genes; Blue, downregulated genes.

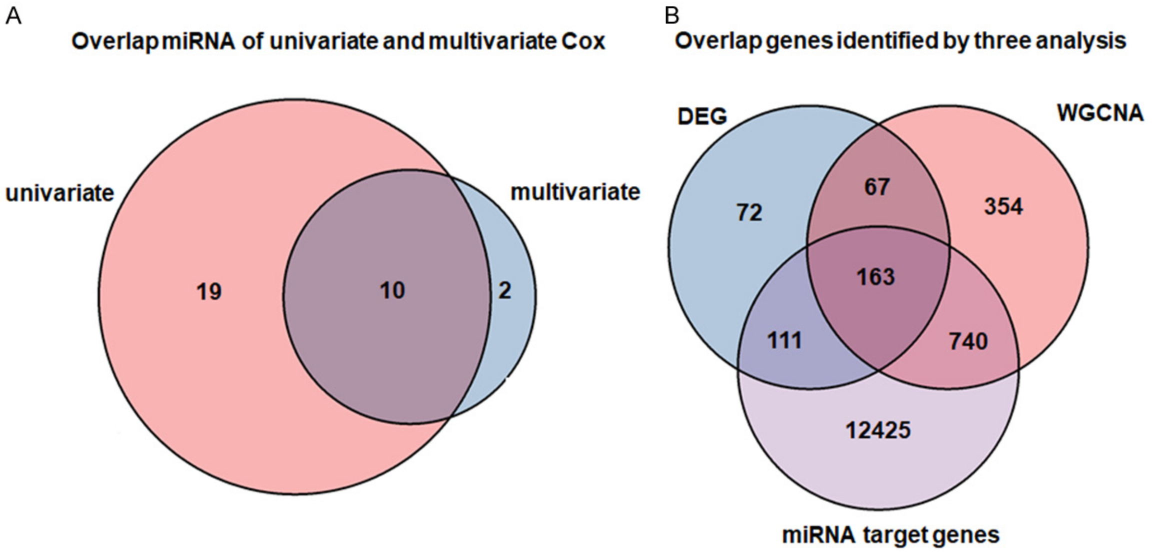


Figure S2. Overlapping genes displayed by Venn diagram. A. Left and right panels show survival-related miRNAs of Logistic and Cox regression methods, respectively. B. Overlapping genes were derived from three analyses: differentially expressed genes (DEGs), Weighted Gene Coexpression Network Analysis (WGCNA), and miRNA target genes.

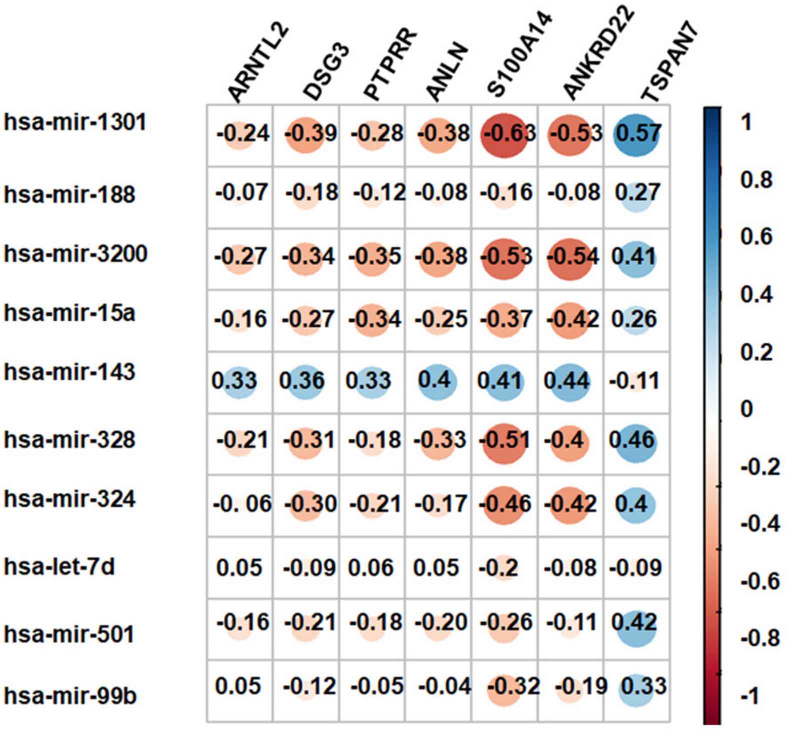
Table S3. Logistic and Cox regression result of 10 survival-related miRNAs

miRNA	Logistic regression *P*-value	Cox regression *P*-value	
		HR (95% CI)	*P*-value
hsa-mir-1301	1.520e-05	2.315 (1.464-3.660)	< 0.001
hsa-mir-188	1.531e-05	2.405 (1.539-3.758)	< 0.001
hsa-mir-3200	3.773e-05	2.196 (1.409-3.422)	0.001
hsa-mir-15a	1.065e-04	1.829 (1.168-2.866)	0.008
hsa-mir-143	4.222e-04	0.500 (0.324-0.772)	0.002
hsa-mir-328	7.613e-04	1.954 (1.244-3.069)	0.004
hsa-mir-324	1.160e-03	1.831 (1.177-2.847)	0.007
hsa-let-7d	2.651e-03	1.870 (1.211-2.889)	0.005
hsa-mir-501	3.907e-03	1.887 (1.236-2.880)	0.003
hsa-mir-99b	7.239e-03	1.936 (1.227-3.054)	0.005

Table S4. 42 mRNAs significantly associated with the overall survival in the 174 TCGA PDAC dataset

	coef	HR (95% CI for HR)	*p*.value
IL1RAP	-0.85	0.43 (0.28-0.66)	0.00014
ARNTL2	-0.83	0.44 (0.28-0.68)	0.00024
INPP4B	-0.74	0.48 (0.31-0.73)	0.00065
AK4	-0.69	0.5 (0.33-0.77)	0.0014
ITGA3	-0.69	0.5 (0.33-0.77)	0.0017
GPRC5A	-0.67	0.51 (0.34-0.78)	0.0021
S100A16	-0.66	0.52 (0.34-0.79)	0.0023
EFNB2	-0.67	0.51 (0.33-0.79)	0.0023
CENPF	-0.66	0.52 (0.34-0.79)	0.0025
MKI67	-0.61	0.54 (0.36-0.82)	0.0041
DSG3	-0.6	0.55 (0.36-0.84)	0.005
PTPRR	-0.6	0.55 (0.36-0.84)	0.0057
APOL1	-0.58	0.56 (0.37-0.85)	0.0066
HEPH	-0.57	0.56 (0.37-0.86)	0.0073
KRT19	-0.56	0.57 (0.38-0.87)	0.009
GJB2	-0.56	0.57 (0.38-0.87)	0.0094
ANLN	-0.55	0.58 (0.38-0.88)	0.01
SERPINB5	-0.54	0.58 (0.39-0.89)	0.011
TSPAN7	0.54	1.7 (1.1-2.6)	0.011
ECT2	-0.54	0.58 (0.38-0.89)	0.012
SULF2	-0.54	0.58 (0.39-0.89)	0.012
KRT6A	-0.52	0.6 (0.39-0.9)	0.015
DCBLD2	-0.51	0.6 (0.39-0.91)	0.016
DDX60	-0.51	0.6 (0.4-0.91)	0.016
KYNU	-0.51	0.6 (0.39-0.91)	0.016
LOXL2	-0.51	0.6 (0.39-0.91)	0.017
TPX2	-0.5	0.61 (0.4-0.92)	0.019
S100A14	-0.51	0.6 (0.4-0.92)	0.019
PADI1	-0.49	0.61 (0.4-0.92)	0.02
CDH3	-0.48	0.62 (0.41-0.94)	0.024
FERMT1	-0.49	0.62 (0.4-0.94)	0.024
STYK1	-0.47	0.63 (0.41-0.95)	0.028

CXCL5	-0.46	0.63 (0.41-0.95)	0.029
FGD6	-0.45	0.63 (0.42-0.96)	0.033
ANKRD22	-0.46	0.63 (0.42-0.97)	0.034
ITGA2	-0.45	0.64 (0.42-0.97)	0.037
PKM	-0.44	0.64 (0.42-0.97)	0.037
SDR16C5	-0.43	0.65 (0.43-0.98)	0.042
SLC2A1	-0.43	0.65 (0.43-0.99)	0.042
MXRA5	-0.43	0.65 (0.43-0.99)	0.044
SLC6A14	-0.42	0.66 (0.43-0.99)	0.046
ERO1A	-0.42	0.66 (0.43-1)	0.05



	ARNTL2	DSG3	PTPRR	ANLN	S100A14	ANKRD22	TSPAN7
hsa-mir-1301	-0.24	-0.39	-0.28	-0.38	-0.63	-0.53	0.57
hsa-mir-188	-0.07	-0.18	-0.12	-0.08	-0.16	-0.08	0.27
hsa-mir-3200	-0.27	-0.34	-0.35	-0.38	-0.53	-0.54	0.41
hsa-mir-15a	-0.16	-0.27	-0.34	-0.25	-0.37	-0.42	0.26
hsa-mir-143	0.33	0.36	0.33	0.4	0.41	0.44	-0.11
hsa-mir-328	-0.21	-0.31	-0.18	-0.33	-0.51	-0.4	0.46
hsa-mir-324	-0. 06	-0.30	-0.21	-0.17	-0.46	-0.42	0.4
hsa-let-7d	0.05	-0.09	0.06	0.05	-0.2	-0.08	-0.09
hsa-mir-501	-0.16	-0.21	-0.18	-0.20	-0.26	-0.11	0.42
hsa-mir-99b	0.05	-0.12	-0.05	-0.04	-0.32	-0.19	0.33

Figure S3. The correlations between 10 miRNAs and 7 genes.