Importance of gene expression signatures in pancreatic cancer prognosis and the establishment of a prediction model

Xiaokai Yan1,2,*
Haifeng Wan1
Xiangyong Hao2
Tian Lan1
Wei Li1
Lin Xu1,3
Kefei Yuan1,3, *
Hong Wu1,3

1Department of Liver Surgery and Liver Transplantation, West China Hospital, Sichuan University, Chengdu, China; 2Department of General Surgery, Gansu Provincial Hospital, Lanzhou 730000, Gansu Province, China; 3Laboratory of Liver Surgery, West China Hospital, Sichuan University, Chengdu, China

*These authors contributed equally to this work

Background and aim: Pancreatic cancer (PC) is one of the most common tumors with a poor prognosis. The current American Joint Committee on Cancer (AJCC) staging system, based on the anatomical features of tumors, is insufficient to predict PC outcomes. The current study is endeavored to identify important prognosis-related genes and build an effective predictive model.

Methods: Multiple public datasets were used to identify differentially expressed genes (DEGs) and survival-related genes (SRGs). Bioinformatics analysis of DEGs was used to identify the main biological processes and pathways involved in PC. A risk score based on SRGs was computed through a univariate Cox regression analysis. The performance of the risk score in predicting PC prognosis was evaluated with survival analysis, Harrell’s concordance index (C-index), area under the curve (AUC), and calibration plots. A predictive nomogram was built through integrating the risk score with clinicopathological information.

Results: A total of 945 DEGs were identified in five Gene Expression Omnibus datasets, and four SRGs (LYRM1, KNTC1, IGF2BP2, and CDC6) were significantly associated with PC progression and prognosis in four datasets. The risk score showed relatively good performance in predicting prognosis in multiple datasets. The predictive nomogram had greater C-index and AUC values, compared with those of the AJCC stage and risk score.

Conclusion: This study identified four new biomarkers that are significantly associated with the carcinogenesis, progression, and prognosis of PC, which may be helpful in studying the underlying mechanism of PC carcinogenesis. The predictive nomogram showed robust performance in predicting PC prognosis. Therefore, the current model may provide an effective and reliable guide for prognosis assessment and treatment decision-making in the clinic.

Keywords: risk score, nomogram, TCGA, GEO

Introduction

Pancreatic cancer (PC), as one of the most common tumors, is the leading cause of cancer-related death worldwide and has a very poor prognosis.1 Currently, the American Joint Committee on Cancer (AJCC) staging system remains the most widely used predictive model for PC. The system was designed to provide a guide for prognosis assessment and therapeutic decisions.2 However, the AJCC staging system was constructed to assess only the three basic indicators of anatomic spread (including the extent of the tumor, the extent of spread to the lymph nodes, and the presence of metastasis) and is unable to comprehensively elucidate tumor behaviors.3 In fact, PC patients with the same AJCC stage may have different clinical prognosis after receiving the same treatments. Therefore, the current predictive system is not sufficient to predict the outcomes of patients with PC, and refinement is necessary.
Over the past few decades, great efforts have been made to identify the molecular markers of cancer. The importance of gene signatures in the initiation, progression, and prognosis of tumors has been shown in many studies. Thousands of genes can be studied simultaneously with the use of next-generation sequencing and novel microarray technologies, facilitating the investigation of the interaction between gene signatures and tumors. Therefore, an increasing number of researchers are interested in using gene signatures for the risk stratification of patients.

To the best of our knowledge, to date, only two studies have used gene expression signatures to build predictive models for PC. Both the studies assessed the power of their prognostic models in a single dataset, and none of these models was constructed based on both clinicopathological factors and gene signatures. In the current study, we endeavored to identify important prognosis-related genes through a multi-dataset analysis, and built composite predictive models for PC that are more applicable in guiding prognostic assessments and treatment decision-making.

**Materials and methods**

**Gene Expression Omnibus (GEO) datasets**

We searched and downloaded mRNA expression profiling data series concerning PC from the GEO (https://www.ncbi.nlm.nih.gov/geo/) using the following keywords: “pancreatic cancer” and “pancreatic ductal adenocarcinoma.” The “Organism” parameter was limited to “Homo sapiens,” and the “study type” parameter was set to “Expression profiling by array.” Ineligible studies were excluded using the following criteria: 1) studies with less than 15 PC samples or non-tumor pancreatic samples; 2) studies using only PC cell lines or xenografts; 3) studies analyzing only blood samples or tumor samples; and 4) studies analyzing only pancreatic endocrine tumors. Finally, five PC datasets (GSE15471, GSE16515, GSE28735, GSE62452, and GSE71729) were selected for further analysis. Probes were matched with the gene names in accordance with the annotation file provided by the manufacturer. If multiple probes matched a single gene, probes were integrated by using the arithmetic mean to account for the expression level of a single gene. The expression data were log2 transformed.

**The Genome Cancer Atlas (TCGA) TCGA dataset**

Transcriptome data (fragments per kilo-base of exon per million fragments) and the corresponding PC clinical information were obtained from TCGA (https://cancer.gov/). After removing patients who died within 3 months and patients without gene expression information, 172 patients with corresponding survival information were retained. Genes expressed in over 80% of samples were retained, and the zero values in the expression matrix were replaced with the minimum non-zero value of the corresponding gene. Then the expression data were log2 transformed.

**Identification of differentially expressed genes (DEGs) and bio-information analysis**

A Significant Analysis of Microarrays (SAM) algorithm was used to identify genes that were differentially expressed between tumor and non-tumor samples via BRB-Array Tools (https://linus.nci.nih.gov/BRB-ArrayTools). A false discovery rate of <0.005 was set as the cutoff criterion. DEGs (including downregulated and upregulated genes) in the five GEO datasets were selected through overlapping analysis, and then functional annotation and pathway enrichment analyses were performed using DAVID software (https://david.ncifcrf.gov/). A protein–protein interaction (PPI) network was established for DEGs using Search Tool for the Retrieval of Interacting Genes (https://string-db.org/) and visualized using Cytoscape 3.6.0.

**Identification of potential prognostic genes**

The expression values of DEGs in the GSE28735, GSE62452, GSE71729, and TCGA datasets were analyzed through a univariate Cox proportional hazard regression model. Genes significantly associated with overall survival (OS) in all these datasets were identified as survival-related genes (SRGs), and a P-value <0.05 was set as the cutoff criterion. Correlation analyses and survival analyses were performed to assess the importance of SRGs in PC progression and prognosis.

A risk score for each dataset was computed through the summation of the gene expression value multiplied by the corresponding coefficient from a univariate Cox regression model (TCGA dataset as a training cohort for risk score, and the other GEO datasets as external validation cohorts). The performance of the risk score in predicting OS was evaluated through a survival analysis, Harrell’s concordance index (C-index), area under the curve (AUC) of the receiver operating characteristic (ROC) curve, and a calibration plot.
comparing predicted vs observed Kaplan–Meier estimates of survival probability.\textsuperscript{20}

Development, comparison, and validation of predictive nomogram

In the TCGA dataset, a predictive nomogram was built on the basis of risk score and clinicopathological information using a backward stepwise Cox proportional hazard model.\textsuperscript{21} The calibration ability of the nomogram was assessed using a calibration plot comparing nomogram-predicted vs observed Kaplan–Meier estimates of survival probability, using 1,000 bootstrap resamples.\textsuperscript{20} We compared the discriminative ability of the nomogram with that of the AJCC stage through the C-index and AUC.\textsuperscript{18} In addition, based on the total point in the nomogram, patients were stratified into three subgroups in the TCGA dataset, including a low-risk group (total point <33.3%), a medium-risk group (total point between 33.3% and 66.6%), and a high-risk group (total point >66.6%), and survival curves for these subgroups were estimated using the Kaplan–Meier method.

Statistical analysis

SAM analysis was performed using BRB-Array Tools. All other statistical analyses were completed using R (https://www.r-project.org/, v3.3.4). A \textit{P}-value <0.05 (two-sided) was considered to indicate statistical significance. A chi-square or Fisher’s exact test was used to assess differences in categorical variables. Student’s \textit{t}-test or a non-parametric Mann–Whitney \textit{U}-test was used to detect differences in continuous variables between two groups. ANOVA or the Kruskal–Wallis test was used to detect the differences in continuous variables between multiple groups. OS was assessed using the log-rank test. HR and 95% CIs were estimated using a Cox regression model. Box plots were constructed using the R package “ggplot2.”\textsuperscript{22} The ROC curve was plotted using the R package “qROC.”\textsuperscript{23} A heat-map was plotted using the R package “gplots.”\textsuperscript{24} The survival analysis and Cox proportional hazard regression analysis were carried out using the R package “survival.”\textsuperscript{25} The C-index and nomogram were completed using the R package “rms.”\textsuperscript{26}

Ethics statement

All datasets (GSE15471, GSE16515, GSE28735, GSE62452, GSE71729, and TCGA) are freely available as public resources. Therefore, additional approval by an ethics committee was not needed in this study.

Results

Identification of DEGs

A total of 9,886, 3,961, 2,276, 3,732, and 1,605 genes differentially expressed between tumor and non-tumor tissues were identified after the SAM analysis of GSE15471, GSE16515, GSE28735, GSE62452, and GSE71729 datasets, respectively (Figure S1A–E). A total of 945 DEGs were found in the five GEO datasets through overlapping analysis (Figure 1A and B; Table S1), including 389 downregulated genes and 556 upregulated genes in tumor samples compared with non-tumor samples. Distinct expression patterns of the 945 DEGs in the five GEO datasets were presented through hierarchical clustering analysis (Figure S2A–E).

Functional annotation analysis, pathway enrichment analysis, and PPI network for DEGs

In Gene ontology (GO) biological process analysis, the 945 DEGs were found to be principally enriched in zinc II ion transmembrane import, wound healing, regulation of lipid catalytic process, regulation of fibroblast migration, positive regulation of synapse assembly, positive regulation of cell growth, as well as other biological processes (Figure 1C). Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis showed that the DEGs were mainly associated with salmonella infection, pyruvate metabolism, proteoglycans in cancer, PI3K-Akt signaling pathway, pathways in cancer, pancreatic secretion, p53 signaling pathway, and other biological pathways (Figure 1D). PPI network was constructed to evaluate the interactive relationships among the DEGs (Figure S3).

Identification of SRGs and the correlation of SRGs with clinicopathological information

Among the 945 DEGs, a total of 64, 190, 136, and 596 genes associated with OS were identified in the GSE28735, GSE62452, GSE71729, and TCGA datasets, respectively. We also found four SRGs (\textit{LYRM1}, \textit{KNTC1}, \textit{IGF2BP2}, and \textit{CDC6}) in the four datasets (Figure 2A) through overlapping analysis. Correlation analysis was performed to determine the association between the expression levels of SRGs and clinicopathological information, including tissues types (normal, tumor, and metastatic samples) (Figure 2B), histological grade (Figure 2C and D), the extent of the tumor (PT) (Figure 2E), tumor subtype (Figure 2F), AJCC stage
Among the SRGs, KNTC1, IGF2BP2, and CDC6 were significantly associated with tissue types, histological grade, PT, and tumor subtype \((P < 0.05)\); LYRM1 was significantly differentially expressed in normal, tumor, and metastatic tissues \((P < 0.05)\).

Meanwhile, the four SRGs were analyzed using X-tile to select the best cutoff values for OS, and on this basis, patients were divided into low- and high-expression groups. Kaplan–Meier survival analysis showed that all SRGs were significantly correlated with patient OS \((P < 0.05)\) in the four datasets (Figure 3A–D).

Collectively, these results indicate that the identified SRGs play important roles in the development and progression of PC.

**Performance assessment of risk score in predicting outcome**

As described previously, the risk score was computed through the summation of the gene expression value multiplied by the
corresponding coefficient (coefficients were obtained from the TCGA dataset through a univariate COX analysis): Risk score = (−0.4705 × expression value of LYRM1) + (0.3707 × expression value of KNTC1) + (0.4106 × expression value of IGF2BP2) + (0.4623 × expression value of CDC6).

Then, we stratified patients into low- and high-risk groups in accordance with the median risk scores in the GSE28735, GSE62452, GSE71729, and TCGA datasets. The Kaplan–Meier survival curves of both groups were notably different in the four datasets (P < 0.05) (Figure 4A1–D1).

The power of the risk score in predicting OS was assessed through C-index and ROC analysis. The C-index of the risk score in the TCGA, GSE71729, GSE62452, and GSE28735 datasets was 0.640 (95% CI, 0.572–0.708), 0.601 (95% CI, 0.531–0.671), 0.648 (95% CI, 0.558–0.738), and 0.689 (95% CI, 0.573–0.805), respectively (Figure S5). The ROC analysis of the risk score is shown in Figure 4A2–D2, and all AUC values at the 3-year point in the four datasets are greater than 0.70.

In addition, relatively good agreement was observed between the expected and observed outcomes for 1-, 2-, and 3-year OS in the calibration curves of risk score (Figure 4A3–D3).

In summary, these results indicate that the risk score shows relatively good performance in predicting the OS of PC patients.

**Assessment of prognostic factors in PC patients**

After removing patients for whom important clinical information was not available (including age, sex, malignancy history, diabetes history, pancreatitis history, tumor size, tumor site, tumor subtype, histological grade, residual tumor, AJCC stage, radiation treatment, and targeted therapy), 95 patients...
Figure 3 Survival analysis of SRGs in four datasets.

Notes: Survival curves of LYRM1 (A1), KNTC1 (B1), IGF2BP2 (C1), and CDC6 (D1) in GSE28735, GSE62452, GSE71729, and TCGA datasets.

Abbreviations: SRGs, survival-related genes; TCGA, The Genome Cancer Atlas.

were retained. Univariate and multivariate adjusted Cox regression analyses were performed to identify prognostic factors for OS. As shown in Table 1, the unadjusted univariate analysis indicated that risk score ($P<0.001$), age ($P=0.013$), tumor size ($P=0.022$), tumor subtype ($P=0.001$), histological grade ($P=0.016$, G3 and G4 vs G1), AJCC stage ($P=0.002$ [IIIB vs I]), $P=0.006$ [III and IV vs I]), radiation treatment ($P=0.014$), and targeted therapy ($P=0.014$) were significantly associated with OS, while the multivariate adjusted Cox regression analysis showed that risk score, age, tumor size, tumor subtype, radiation treatment, and targeted therapy served as significant independent risk factors ($P<0.05$).

Development, comparison, and validation of predictive nomogram

To build a more applicable and individualized predictive model, a predictive nomogram integrating clinical information and gene signatures was constructed based on the
95 patients with complete clinical information in TCGA. Through a stepwise Cox proportional hazard analysis, risk score, age, sex, tumor subtype, tumor size, residual tumor, radiation treatment, and targeted therapy were selected to establish a nomogram model (Figure 5A). The calibration plot for predicting 1-, 2-, and 3-year OS (Figure 5B) showed that the nomogram model performed well with the ideal prediction model.

We compared the predictive power of the nomogram model, AJCC stage and risk score: the C-index (Figure 5C) of the nomogram was 0.804 (95% CI, 0.740–0.868), which is significantly greater than that of the AJCC stage (0.609 [95% CI, 0.536–0.683], \(P<0.001\)) and risk score (0.645 [95% CI, 0.558–0.732], \(P<0.001\)). The AUC of the nomogram at 1 year (Figure 5D) was 0.833 (95% CI, 0.731–0.935), which is superior compared with that of the AJCC stage (0.572 [95% CI, 0.464–0.680], \(P<0.001\)) and risk score (0.707 [95% CI, 0.574–0.840], \(P=0.026\)). The AUC of the nomogram at 2 years (Figure 5E) was 0.888 (95% CI, 0.797–0.978), which is superior to that of the AJCC stage (0.757 [95% CI, 0.636–0.878], \(P=0.039\)) and risk score (0.686 [95% CI, 0.543–0.829], \(P=0.005\)). In addition, based on the total point of the nomogram, we stratified patients into low-, medium-, and high-risk groups (cutoff points were selected at each tertile point). Then, Kaplan–Meier analysis revealed that scoring using the nomograms effectively discriminated the risk groups in PC (\(P<0.0001\)) (Figure 5F).

**Discussion**

In the past few decades, large amounts of data have been generated via high-throughput methods, such as microarrays and next-generation sequencing technologies, which significantly facilitates investigations of the interaction between gene signatures and disease. Meanwhile, an increasing number of
studies tend to identify biomarkers through the analysis of multiple data sources, which often provides stronger evidence than a single data source. In the current study, to enhance the strength of our results, we identified DEGs and SRGs in PC via a joint analysis of six different data sources.

Through GO biological process and KEGG analyses of the DEGs, the main biological processes and pathways involved in human PC were identified (Figure 1C and D). Many previous studies have reported that the PI3K-Akt and p53 signaling pathways play important roles in cell cycle arrest, cell invasion, proliferation, angiogenesis, and metastasis in PC, which is consistent with our results.27–33 Therefore, the biological processes and pathways reported here are worth further study to increase our understanding of the mechanism underlying carcinogenesis and progression in PC.

### Table 1 Cox regression analysis of risk factors associated with overall survival in the TCGA dataset

<table>
<thead>
<tr>
<th>Variables</th>
<th>Unadjusted HR (95% CI)</th>
<th>P-value</th>
<th>Adjusted 1a HR (95% CI)</th>
<th>P-value</th>
<th>Adjusted 2b HR (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk score</td>
<td>1.667 (1.261–2.205)</td>
<td>&lt;0.001</td>
<td>1.539 (1.098–2.157)</td>
<td>0.012</td>
<td>1.524 (1.082–2.146)</td>
<td>0.016</td>
</tr>
<tr>
<td>Age</td>
<td>1.035 (1.007–1.063)</td>
<td>0.013</td>
<td>1.041 (1.008–1.076)</td>
<td>0.015</td>
<td>1.043 (1.012–1.075)</td>
<td>0.006</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1.045 (0.608–1.797)</td>
<td>0.874</td>
<td>0.694 (0.355–1.358)</td>
<td>0.286</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malignancy history</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.993 (0.421–2.341)</td>
<td>0.988</td>
<td>2.046 (0.737–5.686)</td>
<td>0.170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes history</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1.019 (0.534–1.945)</td>
<td>0.954</td>
<td>0.776 (0.373–1.616)</td>
<td>0.498</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pancreatitis history</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1.006 (0.429–2.361)</td>
<td>0.988</td>
<td>1.469 (0.533–4.054)</td>
<td>0.458</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumor size</td>
<td>1.225 (1.030–1.458)</td>
<td>0.022</td>
<td>1.394 (1.120–1.735)</td>
<td>0.003</td>
<td>1.284 (1.051–1.569)</td>
<td>0.014</td>
</tr>
<tr>
<td>Tumor site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>4.064 (0.982–16.830)</td>
<td>0.053</td>
<td>1.439 (0.268–7.730)</td>
<td>0.672</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tail</td>
<td>2.301 (0.420–12.620)</td>
<td>0.337</td>
<td>0.852 (0.128–5.680)</td>
<td>0.869</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumor subtype</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Othersc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDAC</td>
<td>6.849 (2.103–22.300)</td>
<td>0.001</td>
<td>5.760 (1.436–23.109)</td>
<td>0.014</td>
<td>4.412 (1.226–15.886)</td>
<td>0.023</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td>2.207 (0.832–5.857)</td>
<td>0.111</td>
<td>0.651 (0.189–2.240)</td>
<td>0.496</td>
<td>0.773 (0.254–2.351)</td>
<td>0.650</td>
</tr>
<tr>
<td>G3 and G4</td>
<td>3.427 (1.263–9.294)</td>
<td>0.016</td>
<td>1.140 (0.304–4.272)</td>
<td>0.846</td>
<td>1.040 (0.317–3.414)</td>
<td>0.949</td>
</tr>
<tr>
<td>Residual tumor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R0</td>
<td>1.724 (0.994–2.992)</td>
<td>0.053</td>
<td>1.889 (0.973–3.670)</td>
<td>0.060</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AJCC Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IIA</td>
<td>1.964 (0.546–7.069)</td>
<td>0.302</td>
<td>0.996 (0.194–5.119)</td>
<td>0.996</td>
<td>1.180 (0.310–4.485)</td>
<td>0.808</td>
</tr>
<tr>
<td>IIB</td>
<td>5.185 (1.820–14.771)</td>
<td>0.002</td>
<td>1.125 (0.255–4.962)</td>
<td>0.877</td>
<td>1.756 (0.579–5.326)</td>
<td>0.320</td>
</tr>
<tr>
<td>III and IV</td>
<td>11.391 (1.977–65.628)</td>
<td>0.006</td>
<td>5.353 (0.502–57.097)</td>
<td>0.165</td>
<td>8.755 (1.292–59.344)</td>
<td>0.026</td>
</tr>
<tr>
<td>Radiation treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.387 (0.182–0.822)</td>
<td>0.014</td>
<td>0.378 (0.160–0.896)</td>
<td>0.027</td>
<td>0.411 (0.175–0.964)</td>
<td>0.041</td>
</tr>
<tr>
<td>Targeted therapy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.506 (0.295–0.869)</td>
<td>0.014</td>
<td>0.317 (0.157–0.640)</td>
<td>0.001</td>
<td>0.403 (0.214–0.759)</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Notes: *Adjusted covariates include all the indicators above; **Adjusted covariates include the prognostic factors from an unadjusted COX analysis; including neuroendocrine carcinoma, colloid carcinomas, acinar cell carcinoma and adenocarcinoma not otherwise specified. Bold number indicates statistical significance.

Abbreviations: AJCC, the current American Joint Committee on Cancer stage; PDAC, pancreatic ductal adenocarcinoma; TCGA, The Genome Cancer Atlas.
Survival analyses and correlation analyses indicated that the SRGs (LYRM1, KNTC1, IGF2BP2, and CDC6) were significantly associated with PC prognosis. CDC6 is an essential gene required for DNA replication, which has been reported as overexpressed in various types of cancer.34–36 High expression of CDC6 could trigger tumor-like transformation, apoptosis attenuation, genomic instability, cell proliferation, and epithelial-to-mesenchymal transition37–39 and has been associated with poor prognosis in epithelial ovarian cancer.35 CDC6 depletion could result in increased cell death and attenuate tumor migration and invasion.35–40 IGF2BP2 is a post-transcriptional regulatory factor implicated in mRNA localization, stability, and translational control. In previous studies, IGF2BP2 has been confirmed as upregulated in different cancer types41–44 and is associated with tumor carcinogenesis, invasion, and prognosis.43,45,46 Although the function of Homo sapiens LYRM1 and KNTC1 have not yet been studied in cancer, these two genes have been reported to participate in the regulation of cell division, proliferation, and apoptosis,37–40 which may affect tumor development and progression. However, the roles of LYRM1, KNTC1, IGF2BP2, and CDC6 in PC are still unclear, and further

**Figure 5** Performance of the nomogram in predicting prognosis in the TCGA dataset.

**Notes:** (A) Nomogram for predicting 1-, 2-, and 3-year OS in PC patients. (B) Calibration plot for 1-, 2-, and 3-year OS of the nomogram. (C) Comparison of the predictive power of the nomogram model, AJCC stage, and risk score, as assessed using C-index. (D, E) Comparison of the predictive power of the nomogram model, AJCC stage, and risk score by AUC at 1 and 2 years. (F) Kaplan–Meier analysis of risk groups stratified using total point of the proposed nomogram. Other, including neuroendocrine and risk score by aUC at 1 and 2 years. (G) Kaplan–Meier analysis of survival curves for different risk scores. (H) Comparison of the predictive power of the nomogram model, AJCC stage, and risk score by aUC at 1 and 2 years. (I) Kaplan–Meier analysis of survival curves for different risk scores. (J) Comparison of the predictive power of the nomogram model, AJCC stage, and risk score by aUC at 1 and 2 years. (K) Kaplan–Meier analysis of survival curves for different risk scores.

**Abbreviations:** AJCC, the American Joint Committee on Cancer; AUC, area under the curve; C-index, concordance index; OS, overall survival; PC, pancreatic cancer; PDAC, pancreatic ductal adenocarcinoma; TCGA, The Genome Cancer Atlas.
study of their underlying mechanism in PC and potential therapeutic applications is warranted.

The current results demonstrated that the risk score based on the SRGs showed a relatively good and consistent performance in predicting OS in PC patients in the TCGA dataset and the other three validation cohorts (C-indexes of risk score were more than 0.60 and the AUC values at 3-year were more than 0.70 in the four datasets). However, a predictive model based on gene signatures or clinicopathological information alone may be unable to comprehensively elucidate tumor behaviors and their underlying mechanisms. Therefore, a composite and more effective predictive model integrating clinical and gene information is needed.

To the best of our knowledge, a predictive nomogram for PC based on both clinical factors and gene signatures has not been previously reported. In the current study, we generated an effective prognostic nomogram via integrating clinical factors as well as risk score in a TCGA dataset. Good agreement was observed in the calibration curve of our nomogram between the predicted and observed outcomes (Figure 5B). The nomogram demonstrated a greater C-index and AUC values than those of the AJCC stage and risk score (Figure 5C–E). Therefore, our predictive nomogram may facilitate clinicians in predicting the individual risk of patient death and provide guidance for patient assessment and therapeutic decision-making.

However, there are some limitations in the current study. First, we studied the roles of SRGs through data mining only, and no experimental data on the molecular mechanisms of these genes in PC have been reported. Therefore, further experimental studies may enhance our understanding of the biological behavior of PC. Second, the nomogram was developed and validated in a single dataset, and therefore the performance of our model needs to be further validated in independent external datasets with complete gene and clinical information.

Conclusion

The current study identified four new biomarkers that are significantly associated with PC carcinogenesis, progression, and prognosis, which may be helpful in studying underlying carcinogenesis mechanisms and potential therapeutic applications in PC. The predictive nomogram showed robust performance in predicting PC prognosis. Therefore, our model may provide an effective and reliable guide to prognosis assessment and treatment decision-making in the clinic.

Acknowledgments

This work was supported by grants from the Natural Science Foundation of China (81672882 and 81502441), the Science and Technology Support Program of Sichuan Province (2017SZ0003), and Gansu Province Science Foundation for Youths (18JR3RA058).

Author contributions

All authors contributed to data analysis, drafting and revising the article, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Disclosure

The authors report no conflicts of interest in this work.

References


