Identification By Genetic Algorithm Optimized Back Propagation Artificial Neural Network and Validation of A Four-Gene Signature for Diagnosis and Prognosis of Pancreatic Cancer

Zhenchong Li
Department of General Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, School of Medicine, South China University of Technology, Guangzhou 510080

Shanzhou Huang
Department of General Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, School of Medicine, South China University of Technology, Guangzhou 510080

Zuyi Ma
Shantou University Medical College

Dongmei Xia
Department of General Surgery, Hui Ya Hospital of The First Affiliated Hospital, Sun Yat-Sen University, Huizhou, Guangdong 516081

Yifeng Cai
Department of General Surgery, Hui Ya Hospital of The First Affiliated Hospital, Sun Yat-Sen University, Huizhou, Guangdong 516081

Jian Wu
Department of Liver Surgery, The First Affiliated Hospital of Sun Yat-Sen University, Guangzhou, 510080

Hongkai Zhuang
Shantou University Medical College

Zixuan Zhou
Department of General Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, School of Medicine, South China University of Technology, Guangzhou 510080

Shujie Wang
The Second School of Clinical Medicine, Southern Medical University, Guangzhou 510515, Guangdong Province

Chunsheng Liu
Shantou University Medical College

Qi Zhou
Department of General Surgery, Hui Ya Hospital of The First Affiliated Hospital, Sun Yat-Sen University, Huizhou, Guangdong 516081

Chuanzhao Zhang
Department of General Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, School of Medicine, South China University of Technology, Guangzhou 510080

Baohua Hou (hbh1000@126.com)

Department of General Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, School of Medicine, South China University of Technology, Guangzhou 510080

Research Article

Keywords: Pancreatic cancer, Biomarker, Diagnosis, Prognosis, Machine learning

Posted Date: February 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-151851/v1

Abstract

Background: Although some improvements in the management of pancreatic cancer (PC) have been made, no major breakthroughs in terms of biomarker discovery or effective treatment have emerged. Here, we applied artificial intelligence (AI)-based methods to develop a model to diagnose PC and predict survival outcome.

Methods: Multiple bioinformatics methods, including RankProd, were performed to identify differentially expressed genes (DEGs) in PC. A Back Propagation (BP) model was constructed, followed by Genetic Algorithm (GA) filtering and verification of its prognosis capacity in the TCGA cohort. Furthermore, we validated the protein expression of the selected DEGs in 92 clinical PC tissues using immunohistochemistry.

Results: Four candidate genes (*LCN2*, *SLC6A14*, *SPOCK1*, and *VCAN*) were selected to establish a four-gene signature for PC. The gene signature was validated in the TCGA PC cohort, and found to show satisfactory discrimination and prognostic power. Areas under the curve (AUC) values of overall survival were both greater than 0.60 in the TCGA training cohort, test cohort, and the entire cohort. Kaplan-Meier analyses showed that high-risk group had a significantly shorter overall survival and disease-free survival than the low-risk group. Further, the elevated expression of *SLC6A14* and *SPOCK1* in PC tissues was validated in the TCGA+GETx datasets and 92 clinical PC tissues, and was significantly associated with poor survival in PC.

Conclusions: Using RankProd and GA-ANN, we developed and validated a diagnostic and prognostic gene signature that yielded excellent predictive capacity for PC patients' survival.

Introduction

Pancreatic cancer (PC) is one of the most common cancers in the world, with 458,918 new cases and accounting for 4.5% of all cancer-related deaths in 2018, according to GLOBOCAN 2018 (1). Despite the development of new tools for early diagnosis and identification of potential risk factors of PC, its incidence is still increasing and 355,317 new cases are predicted to occur in 2040. There are two main types of PC: pancreatic adenocarcinoma, the most common type that accounts for 85% of cases and occurs in pancreatic exocrine glands; and pancreatic neuroendocrine tumor, which is less common, accounts for less than 5% of cases, and arises in pancreatic endocrine tissue (2). The prognosis of pancreatic adenocarcinoma is very poor: only 24% of patients survive for one year, while 9% survive for five years (3). Despite numerous recent advances in the management of PC, the 5-year survival rate for PC has increased from 6% to only 9% from 2014 to 2018 (3). Surgical resection remains the only potential cure for PC. Surgery, chemotherapy, and radiotherapy have traditionally been used to prolong survival and/or relieve the symptoms of PC. However, there is still no clear cure for advanced-stage patients (4). Thus, there is an urgent need for further research to develop local and systemic treatment, along with the need to evaluate the outcomes of these approaches.

Current diagnostic tests for PC are still non-specific and may miss some early-stage cases (5). Most cases of PC are diagnosed at an advanced stage, and 80–90% of patients have unresectable tumors when diagnosed (5). Early-stage PC is usually clinically asymptomatic, and patients with symptoms attributable to PC mostly have advanced disease. Therefore, it is important to elucidate the mechanisms involved in the transformation of a healthy pancreatic cell into a tumor cell and to identify potential biomarkers expressed at the early stage of PC. Therefore, early detection of PC is vital for selecting an optimal therapeutic approach for patients and prolonging their survival (7, 8).

Several diagnostic tools are available in clinical practice; these include abdominal ultrasonography, tri-phasic CT (criteria for diagnosis and staging) (9, 10) and magnetic resonance imaging (MRI) of the abdomen (11), as well as endoscopic ultrasound-guided fine-needle aspiration cytology (12). Biopsy is of great value for diagnosis, and its sensitivity is about 80% (12). However, there is great room for improvement in sensitivity and accuracy, as well as prognostic prediction. Therefore, a comprehensive analysis of accurate prognostic biomarkers is needed to guide patients' treatment. Second-generation sequencing technologies and high-throughput microarray chips are valuable tools for the discovery of novel cancer biomarkers. However, a high rate of statistical errors has been noted because of the relatively small amount of samples (13). Larger sample sizes and the use of machine-learning algorithm has reduced such errors effectively (13, 14). Many studies also employed integrated various data to increase their sample size and thus identified promising biomarkers (15). Suraj et al. screened differentially expressed genes (DEG) to identify NF-κB and interferon signatures of clear cell renal cell carcinoma by integrating different datasets from kidney tissue microarrays (16). Hou et al. combined RankProd and an artificial neural network to develop a diagnostic and prognostic model, and identified C1QTNF3 as a biomarker for prognostic prediction of prostate cancer (17).

Herein, we applied artificial intelligence (AI)-based methods to develop a model comprising a small gene set that may be used to diagnose PC and predict survival outcome. First, we expanded the sample size by integrating data from various independent datasets using RankProd. Next, we applied a genetic algorithm-artificial neural network (GA-ANN) model to screen for candidate genes and construct a predictive model for PC. RankProd and GA-ANN were used in combination, providing a promising processing approach to discover candidate gene patterns and novel biomarkers for PC diagnosis and prognosis prediction.

Methods

Data collection

Publicly available data for pancreatic carcinoma samples and normal controls were collected from The Cancer Genome Atlas (TCGA) database (http//gdc.cancer.gov/) and the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) and analyzed. The following criteria were applied for the collection of microarray data from GEO: (1) Datasets were created by microarray analysis for genome-wide mRNA expression profiling; (2) Single-channel of the experimental platform was used; (3) All cases were pathologically diagnosed as pancreatic carcinoma; (4) Number of cases or normal controls must be more than 10. All data types were originally Log2(FPKM + 1), and were then converted to Log2(TPM + 1). We used the Affy package of R and gcRNA package to draw and standardize the GEO data. The robust multi-array average (RMA) method was used for quality control and data normalization. If multiple probes corresponded to the same gene, the average expression value of these probes was taken as the expression level of the gene. Clinical data for the pancreatic samples in the TCGA dataset were obtained from cBioPortal (http://www.cbioportal.org/). Data for normal pancreatic tissues from GTEx was downloaded from xenabrowser (https://xenabrowser.net/). A |log2foldchange|>2 and a padj < 0.01 was considered to indicate statistical significance for the DEGs; analysis was conducted using the limma package of R, based on the Benjamini-Hochberg procedure. The expression patterns of DEGs in tumor- and normal tissues were determined by clustering analyses. The flowchart of the data analysis process was shown in Fig. 1.

GO and pathway enrichment analyses

DEGs were integrated to DAVID 6.7 (https://david-d.ncifcrf.gov/) to perform Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis in order to explore the potential biological functions and pathways. The false discovery rate (FDR) < 0.01 was set as the threshold. Results were visualized using R package ggplot2.

Construction of the GA-BP PC prediction model

Based on the DEGs, we developed the Back Propagation model (BP model) in MATLAB (MathWorks, Massachusetts, USA) with the expression values of the DEGs as the set input variables and the type of samples as the output variables (cancer or normal tissues). A training set was developed with 39 microarray samples and 39 other samples were used as a test set both from GSE15471 by random assignment. The model consisted of five layers with the number of DEGs as the input layer (each layer represents the expression value of one probe) and one node as the output layer (the type of samples). The learning goal was set at 0.1 and the learning rate was 0.1. According to the optimization by Genetic Algorithm (GA), the initial population number was 30 and the maximum evolutionary generation number was 100. The useful input variables were randomly selected by GA-ANN to maintain stable computational accuracy at each round of calculation. Thus, the input variables could be reduced by nearly half per round. The candidate input variables (probes) were obtained after five rounds of calculations.

Diagnostic assay for gene signature in independent dataset

The predictive ability and stability were assessed in both training and test sets. Logistic regression was used to assay the relative risk and diagnostic capacity of genes from a GEO dataset. Using the R package “glm”, the dataset was randomly divided into 10 groups, with nine groups as training sets and one group as the test set. Then, we combined the gene expression data to construct a linear model. Each sample was assigned a logistic regression coefficient and composite index. Next, the area under curve (AUC) of the receiver operating characteristic (ROC) curves and accuracy (ACC) was used to evaluate the performance of the model using R package “ROCR”. ACC equals (TP + TN)/N. TN represents true negative, TP represents true positive, and N represents sample numbers. Finally, the diagnosis model was confirmed for other datasets used in the study.

Prognostic index of gene signature in prognosis of survival of PC

Patients with overall survival (OS) information from TCGA were divided into a training set (89 cases) and test set (88 cases). Patients with disease-free survival (DFS) information from TCGA were divided into a training set (69 cases) and a test set (68 cases). A prognostic index (PI) was constructed as a comprehensive indicator of the candidate genes for each PC patient in the BP model. The PI was calculated by a linear combination of the gene expression values weighted by Cox regression coefficients. The calculation formula for PI was defined as follows: Risk Score = $\sum_i(\beta_i * X_i)$

βi is the Cox regression coefficient of the ith variable and Xi is the value of the ith variable. For the form of PI, X_i is the expression value of each mRNA after log2-transformation and βi is the univariate Cox regression coefficient of the ith mRNA.

Patients, samples, and follow-up

We obtained tissue samples for validation from patients who had been diagnosed with PDAC pathologically after surgery at Guangdong Provincial People’s Hospital or the First Affiliated Hospital of Sun Yat-sen University from 2015 to 2017. All patients enrolled into the study were followed up until August 2020. Overall survival (OS) was

defined as the duration between surgical resection and death or the last follow-up. Disease-free survival (DFS) was defined as the duration from surgery to tumor recurrence or metastasis.

Immunohistochemistry and reagents

Formalin-fixed paraffin-embedded specimens were used for immunohistochemistry. After deparaffinization, hydration, and blocking, the samples were mixed with the primary antibody and incubated overnight at 4°C (dilution ratio 1:1,000). The staining was compared and evaluated as previously described (18). The following antibodies were purchased from Abcam: SLC6A14 (ab254786) and SPOCK1 (ab229935).

Statistical analysis

Chi-square tests and Fisher's exact tests were used to compare categorical data. Log-rank tests and Kaplan-Meier analyses were performed to assess the predictive ability of the prognostic model. Data processing and analysis was performed using SPSS 22.0 software (IBM, USA). $P < 0.05$ was considered to indicate significance. The symbols * and ** have been used to represent $p < 0.05$ and $p < 0.01$, respectively, in the figures.

Results

Description of Microarray Data and Processing Methodology

GSE15471, GSE62165, GSE62452, and GSE78229 datasets from the GEO database were included in this study (Table 1). The GSE15471 dataset contains data for 39 PC and 39 normal pancreatic tissue samples, without survival information. The GSE62165 dataset consists of data for 118 PC and 13 normal pancreatic tissue samples, without survival information. The GSE62452 dataset is comprised of data for 69 PC and 61 normal pancreatic tissue samples, with OS information for 65 cases. There are data for 177 PC samples and only four normal pancreatic tissues in the TCGA dataset. Data for 177 cases with OS information and 137 cases with DFS information were obtained from the TCGA dataset. GTEx has data for 167 normal pancreatic tissues.

Table 1
Characteristics of datasets used in the study.

Dataset	Tumor	Normal	OS	DFS
GSE15471	39	39	-	-
GSE62165	118	13	-	-
GSE62452	69	61	65	-
GSE78229	49	-	49	-
TCGA	177	4	177	137
GTEx	-	167	-	-

OS, overall survival. DFS, disease-free survival. TCGA, The Cancer Genome Atlas. GTEx, The Genotype-Tissue Expression.

Identification and Functional Enrichment of DEGs

GSE15471 and GSE62165 datasets were used to identify DEGs by comparing pancreatic carcinoma tissues with normal pancreatic tissues (Fig. 2). Heatmap analysis showed that DEGs were differentially expressed between

normal and cancer tissues (Fig. 2A **and B**). Below the cut off, we identified 150 DEGs (125 upregulated genes and 25 downregulated genes) from GSE15471 and 571 DEGs (363 upregulated genes and 208 downregulated genes) from GSE62165, as shown by the volcano plot gene expression profiles (Fig. 2C **and D**). Furthermore, we identified 103 co-regulated (82 co-upregulated and 21 co-downregulated) DEGs between the two datasets (Fig. 2E). GO and KEGG analyses were performed to identify functions of the DEGs (Fig. 3A **and B**). GO terms revealed that DEGs were enriched in biological processes (BP) including “extracellular structure organization; extracellular matrix organization; skeletal system development; collagen fibril organization; collagen metabolic process; bone development; chondrocyte development; ossification; cartilage development; connective tissue development”. Cell component mainly comprised “extracellular matrix; extracellular matrix component; collagen-containing extracellular matrix; endoplasmic reticulum lumen; fibrillar collagen trimer; complex of collagen trimers; collagen trimer; banded collagen fibril; basement membrane; secretory granule lumen”. Molecular functions of DEGs were mostly enriched in “extracellular matrix structural constituent conferring tensile strength; glycosaminoglycan binding; proteoglycan binding; extracellular matrix structural constituent; collagen binding; heparin binding; integrin binding; sulfur compound binding; protease binding; platelet-derived growth factor binding”. KEGG pathway enrichment analysis showed that the DEGs were mainly related to “Protein digestion and absorption; ECM-receptor interaction; Focal adhesion; Amoebiasis; Human papillomavirus infection; PI3K-Akt signaling pathway; AGE-RAGE signaling pathway in diabetic complications; Relaxin signaling pathway; Small cell lung cancer; Rheumatoid arthritis”

GA-BP screening for gene signature and diagnostic capacity

We input the 103 co-regulated genes as characteristic variables to establish the BP model, followed by GA filtering. After five rounds of modeling, the predictive accuracy of both BP and GA-BP for diagnosis of PC reached 100%, with high modeling speed (Fig. 4A **and B**). The process of training and testing are shown in Table 2. Then, we obtained four genes as a minimum candidate gene list to diagnose whether the pancreatic sample was normal or tumor tissue (Fig. 4C **and D**); these four genes were as follows: *LCN2*, *SLC6A14*, *SPOCK1*, and VCAN. The four-candidate-gene model yielded an 86.96% diagnostic accuracy for normal pancreatic tissue. The ROC revealed that the AUC of the model was as high as 0.9565 (Fig. 4E). We performed 10 rounds of cross-validation to test the stability of the model (Fig. 4F). The results showed that the sensitivity and specificity were above 0.750.

Table 2
Table of parameter of BP, accuracy rate of prediction and number of genes filtered from GA-BP.

	Circle 1	Circle 2	Circle 3	Circle 4	Circle 5	Circle 5
Training(normal/tumor)	39(17/22)	39(16/23)	39(16/23)	39(16/23)	39(16/23)	39(16/23)
Testing(normal/tumor)	39(22/17)	39(23/16)	39(23/16)	39(23/16)	39(23/16)	39(16/23)
Testing results of BP						
Normal (TN)	86.36%	86.96%	89.96%	82.61%	86.96%	91.30%
Tumor (TP)	88.24%	93.75%	87.5%	93.75%	100%	100%
Time cost for modeling	2.23s	2.07s	2.30s	2.23s	1.65s	1.37s
Testing results of GA-BP	Population = 30 Generation = 100	Population = 30 Generation = 100	Population = 20 Generation = 100	Population = 10 Generation = 100	Population = 8 Generation = 100	
Normal (TN)	81.82%	78.26%	82.61%	91.30%	86.96%	
Tumor (TP)	94.12%	81.25%	93.75%	81.25%	100%	
Time cost for modeling	0.27s	0.28s	0.25s	0.33s	0.19s	
Candidate genes	57	33	15	8	4	

BP, Back Propagation. GA, Genetic Algorithm. TN, normal pancreatic tissues. TP, pancreatic carcinoma tissues.

Then, the diagnosis capacity of the model was validated in external datasets (Table 3). The model demonstrated high accuracy of diagnosing normal and tumor tissues. Apart from GSE62452, the ACC and AUC are both higher than 87%. The model also showed better capacity for prediction of overall survival than for other clinical characteristics (Table 4 **and** Table 5).

Table 3
Validation of the model's diagnosis capacity in external datasets

Dataset	TN	TP	ACC	AUC
GSE62165	92.31%	86.44%	87.02%	0.8932
GSE15471_GSE62165	90.38%	91.72%	91.39%	0.9096
GSE62452	80.33%	68.12%	73.85%	0.7406
TCGA_GTEx	98.25%	76.27%	87.07%	0.8719

TN, normal pancreatic tissues. TP, pancreatic carcinoma tissues. ACC, Accuracy. AUC, Area Under Curve. TCGA, The Cancer Genome Atlas. GTEx, The Genotype-Tissue Expression.

Table 4
The beta, p-values and hazard ratio coefficients of 4 genes in survival prediction model for the TCGA cohort.

Genes	Overall survival				Disease-free survival			
	beta	Hazard ratio	95% CI	p.value	beta	Hazard ratio	95% CI	p.value
LCN2	0.37	1.4	0.95–2.2	0.086	0.51	1.7	1.1–2.6	0.027
SLC6A14	0.62	1.9	1.2–2.8	0.0035	0.48	1.6	1-2.5	0.033
SPOCK1	0.26	1.3	0.86-2	0.22	0.20	1.2	0.79–1.9	0.37
VCAN	0.21	1.2	0.81–1.9	0.32	0.33	1.4	0.88–2.2	0.16

Table 5
Predictive value of factors for Overall survival (OS) and Disease-free survival (DFS).

Prognostic factors	OS			DFS		
	Training	Test	Entire	Training	Test	Entire
Risk score	**0.6296**	0.6158	**0.6237**	0.5933	**0.6609**	**0.6024**
Age	0.5519	0.6097	0.5755	0.5189	0.4662	0.4908
TMN stage (T)	0.5474	0.5669	0.5607	0.4964	0.6517	0.5767
TMN stage (M)	0.4835	0.5078	0.4971	0.4545	0.5345	0.4991
TMN stage (N)	0.5561	0.6285	0.5890	0.5203	0.5734	0.5477
Grade	0.5446	0.5511	0.5572	0.6412	0.5287	0.5895
Stage	0.5446	0.5553	0.5473	0.5111	0.6193	0.5674

Prognostic prediction capacity of the four-gene signature

We first verified the prognosis prediction capacity of the four-gene signature for OS: 177 samples from the TCGA database with OS information were divided into a training cohort (89 cases) and a test cohort (88 cases). The risk score was calculated in the training cohort with a median value of 9.1566. Cases with risk scores higher than 9.1566 were identified as belonging to the high-risk group. Cases with risk scores lower than 9.1566 were defined as belonging to the low-risk group. The high-risk group (45 cases) showed a significantly poorer survival than the low-risk group (44 cases) in the training cohort (p = 0.031, Fig. 5A). The AUC of the ROC is 0.6296 (Fig. 5D). Then, the risk model was validated in the testing cohort (p = 0.012, Fig. 5B) and the entire TCGA cohort (p = 0.0012, Fig. 5C), both of which showed inferior survival outcome in the high-risk groups with AUC higher than 0.600 of the ROC (Fig. 5E **and F**).

We next tested the prognosis prediction capacity of the four-gene signature for DFS: 137 samples from the TCGA database with DFS information were divided into a training cohort (69 cases) and a test cohort (68 cases). The risk score was calculated in the training cohort with a medium value of 10.5966. Thirty-five cases with risk scores higher than 10.5966 were identified as belonging to the high-risk group; 34 cases with risk scores lower than 10.5966 were defined as belonging to the low-risk group. The high-risk group had a significantly poorer survival than the low-risk group in the training cohort (p = 0.025, Fig. 6A). The AUC of the ROC is 0.5933 (Fig. 6D). Then, the risk model was validated in the testing cohort (p = 0.017, Fig. 6B) and the entire TCGA cohort (p = 0.0039, Fig. 6C),

both of which showed inferior survival outcomes in the high-risk groups with AUC of the ROC higher than 0.600 (Fig. 6E **and F**). Besides, univariate analyses were performed and it's showed that risk score, age, T stage, N stage, grade and AJCC stage were associated with OS and DFS of patients (Table 6).

Table 6
Univariate and multivariate analysis for Overall survival (OS) and Disease-free survival (DFS).

	Univariate analysis	
	Hazard ratio (95%CI)	p.value
Prognostic factors for OS		
Risk score (High vs Low)	2 (1.3-3)	0.0014
Age (< 70 vs > = 70)	1.6 (1.1–2.4)	0.016
TMN stage (T1-T2 vs T3-T4)	2.1 (1.1-4)	0.019
TMN stage (M0 vs M1)	1.1 (0.33–3.5)	0.89
TMN stage (N0 vs N1-N1b)	2.2 (1.3–3.6)	0.0026
Grade (G1-G2 vs G3-G4)	2.4 (1.1–5.3)	0.025
Stage (I-II vs III-IV)	1.6 (1-2.4)	0.038
Prognostic factors for DFS		
Risk score (High vs Low)	1.5 (0.94–2.3)	0.09
Age (< 70 vs > = 70)	1.9 (1.2–2.9)	0.0064
TMN stage (T1-T2 vs T3-T4)	2.2 (1.2–4.2)	0.017
TMN stage (M0 vs M1)	0.94 (0.23–3.9)	0.93
TMN stage (N0 vs N1-N1b)	1.8 (1.1–2.9)	0.018
Grade (G1-G2 vs G3-G4)	2.8 (1.3–6.1)	0.012
Stage (I-II vs III-IV)	1.8 (1.1–2.8)	0.012

Validation of Candidate Genes in PC Tissues

The expression of 4 candidate genes in the TCGA cohort and GTEx samples is shown in Fig. 7A and **Figure S1**. All were expressed at higher levels in PC tissues than in normal pancreatic tissues. The high expression of SLC6A14 and SPOCK1 is associated with inferior OS (Fig. 7B). Subsequently, we detected the protein expression of these two genes in 92 clinical PC tissues and paired normal pancreatic tissues by IHC. The results showed that the expression of SLC6A14 and SPOCK1 in tumor tissues was significantly higher than that in normal tissues (Fig. 7C **and D**). SLC6A14 was overexpressed in 71 cases out of 92 PC samples, while SPOCK1 was overexpressed in 59 cases. We also noticed that patients with high levels of SLC6A14 ($p = 0.032$) and SPOCK1 ($p = 0.009$) expression in tumor tissue had poorer survival than those with low-level protein expression (Fig. 7E). Overall, these data suggest that SLC6A14 and SPOCK1 were constantly overexpressed in PC.

Discussion

Despite consistent progress in the diagnosis and management of PC, few breakthroughs for effective biomarkers and treatment strategies have emerged (19). Identification of the biological and molecular mechanisms as well as discovery of timely diagnostic, prognostic, and therapeutic biomarkers for PC is therefore necessary (20). Various studies have attempted to identify biomarkers and construct predictive models to diagnose or predict the survival of PC. Cheng et al. identified diagnostic and prognostic biomarkers for PC by a comprehensive analysis, which may promote proliferation and migration of PC cells (21). Wu et al. developed a nine-gene signature based on GEO and TCGA datasets and constructed a nomogram combining the gene signature and clinical prognostic features to predict OS in PC (22). Our present study provides a different approach to selecting candidate biomarkers and establishing a diagnosis and survival prediction model. To screen potential biomarkers that may enable diagnosis and prognostic prediction for PC, RankProd and GA-ANN were used to construct a model. Finally, four DEGs (*LCN2*, *SLC6A14*, *SPOCK1*, and *VCAN*) were identified, and a four-gene signature was developed by our data processing system. Both AUCs and Kaplan-Meier analyses of the gene signature for OS and DFS showed a stably high value for diagnosis and prognosis of PC.

In the current study, the four candidate genes *LCN2*, *SLC6A14*, *SPOCK1*, and *VCAN* were selected for further study. Studies have shown that these four genes are vital in cancer diagnosis and progression, especially in PC. The role of *LCN2* in PC was contradictory. Its expression is increased in pancreatic neoplasia, and this up-regulated level is correlated with malignant progression to PC (23–25); the increased expression of *LCN2* has also been observed in various mouse models of PC (26). However, LCN2 depletion was also found in poorly differentiated PC (mesenchymal-like) and considered to be essential for invasion and metastasis (27). This down-regulation may be brought about by the activation of the EGFR signaling pathway, which inhibits E-cadherin and epithelial-to-mesenchymal transition (EMT) (28). LCN2 is also reported to inhibit angiogenesis and cause hypovascular conditions in tumor microenvironment (29). Thus, a therapeutic strategy involving the inhibition of LCN2-induced hypovascularity may potentially enhance the delivery of chemotherapeutic drugs and improve treatment effectiveness. As a member of the SLC6 family, SLC6A14 is a Na^+- and Cl^--dependent solute transport molecule that activates the transport of neutral and basic amino acids (30). Previous studies have revealed that the expression of SLC6A14 is increased in cervical cancer, colorectal cancer, breast cancer, as well as PC (31). In our study, we also confirmed its increased expression in PC tissue and its correlation with poor survival of PC patients. It has been shown that blockade of SLC6A14 with either α-methyl-L-tryptophan (α-MT), a pharmacological inhibitor, or shRNA-mediated gene silencing causes amino acid starvation, inhibits the mTORC1 signaling pathway, and decreases PC cell growth and proliferation in vitro and in vivo (30). Thus, α-MT exhibited convincing specificity and potency as a pharmacological blocker of SLC6A14 and drug target for PC therapy. Gemcitabine-based chemotherapy is the main treatment for PC patients with or without surgery. Resistance to gemcitabine is a growing challenge to the effective treatment of PC because of the down-regulation of drug transporters SLC29A1 (ENT1) and SLC28A1 (CNT1) (32, 33). Because the expression of SLC6A14 is increased in PC, amino acid-based prodrug forms of gemcitabine could be used as substrates for SLC6A14 to enhance the chemotherapeutic sensitivity of gemcitabine in this form of cancer. SPARC (Osteonectin), Cwcv and Kazal like Domains Proteoglycan 1 (SPOCK1), one of the Ca^{2+}-binding proteoglycan family members, was shown to be highly expressed in several cancer types (34). Studies have indicated that SPOCK1-mediated EMT regulates proliferation and invasion in various malignancies (34). A recent study has shown that SPOCK1 induces EMT to promote PC metastasis and inactivates the PI3K/Akt signaling pathway to attenuate PC cell apoptosis, in vitro and in vivo (35). Our work also showed that the expression level of SPOCK1 was up-regulated and associated with a shorter overall survival in PC.

VCAN, an ECM macromolecule, induces several biological activities such as apoptosis and is known to accumulate in several types of cancers (36). It has been reported that VCAN interacts with numerous ECM components including hyaluronan, fibronectin, thrombospondin 1, and fibrillin to create an active biopolymer that affects cell morphosis, adhesion, proliferation, and migration (37, 38). However, studies of VCAN in PC are limited; therefore, further research is needed to investigate the roles of the post-translational modifications of VCAN.

Although the gene signature presented here possessed satisfactory predictive value, there are a few limitations to our study. First, PC is a heterogeneous tumor at the genetic and molecular level; therefore, the established model needs to be further validated in other clinical studies. To better understand the potential roles of *LCN2*, *SLC6A14*, *SPOCK1*, and *VCAN* in PC, further experiments and in vitro or in vivo studies are necessary to validate our results and explore the underlying molecular mechanisms.

In conclusion, the application of RankProd and GA-ANN enabled the identification of novel biomarkers for the diagnosis and prognostic prediction of PC in this study. Using this data processing approach, we developed and validated a prognostic gene signature that showed excellent predictive capacity for patient survival in PC.

Declarations

Acknowledgments

Not applicable.

Declaration

The study was executed according to the World Medical Association Declaration of Helsinki. All methods performed in our study were according to the relevant guidelines and regulations approved by the Medical Research Ethics Committee of Guangdong Provincial People's Hospital and the First Affiliated Hospital of Sun Yat-sen University. All patients from Guangdong Provincial People's Hospital and the First Affiliated Hospital of Sun Yat-sen University who enrolled into the study provided written informed consent.

Authors' contributions

Conceptualization, Zhenchong Li, Shanzhou Huang and Zuyi Ma; Data curation, Zhenchong Li; Formal analysis, Zhenchong Li, Hongkai Zhuang, Zixuan Zhou, Shujie Wang and Chunsheng Liu; Funding acquisition, Qi Zhou, Chuanzhao Zhang and Baohua Hou; Investigation, Zhenchong Li, Dongmei Xia, Yifeng Cai and Jian Wu; Methodology, Zhenchong Li, Shanzhou Huang, Zuyi Ma, Dongmei Xia, Yifeng Cai and Jian Wu; Software, Hongkai Zhuang, Zixuan Zhou, Shujie Wang and Chunsheng Liu; Writing – original draft, Zhenchong Li, Shanzhou Huang, Zuyi Ma, Qi Zhou, Chuanzhao Zhang and Baohua Hou; Writing – review & editing, Zhenchong Li, Shanzhou Huang, Zuyi Ma, Qi Zhou, Chuanzhao Zhang and Baohua Hou.

Funding

This study was supported by Special Funding from the PhD doctorate program of Guangdong Provincial People's Hospital (2020bq09), the Science and Technology Program of Huizhou (2018Y305, 2019C0602009), Funding for the Construction of Key Specialty in Huizhou (Qi Zhou), National Natural Science Foundation of China (Project No.: 82072635, 81702783, 82072637 and 81672475), and Guangdong Medical Science and Technology Fund (Project No.: 201707010323).

Availability of data and materials

The data that supports the findings of this study are available from TCGA (https://portal.gdc.cancer.gov/repository) and GEO databases (GEO, https://www.ncbi.nlm.nih.gov/geo/).

Ethics approval and consent to participate

No ethics approval was required for this work. All utilized public data sets were generated by others who obtained ethical approval.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflicts of interest.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394-424.
2. Hidalgo M, Cascinu S, Kleeff J, Labianca R, Lohr JM, Neoptolemos J et al. Addressing the challenges of pancreatic cancer: future directions for improving outcomes. PANCREATOLOGY 2015;15(1):8-18.
3. McGuire S. World Cancer Report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. ADV NUTR 2016;7(2):418-419.
4. Mohammed S, Van Buren GN, Fisher WE. Pancreatic cancer: advances in treatment. World J Gastroenterol 2014;20(28):9354-9360.
5. De La Cruz MS, Young AP, Ruffin MT. Diagnosis and management of pancreatic cancer. AM FAM PHYSICIAN 2014;89(8):626-632.
6. Agarwal B, Correa AM, Ho L. Survival in pancreatic carcinoma based on tumor size. PANCREAS 2008;36(1):e15-e20.
7. Avgerinos DV, Bjornsson J. Malignant neoplasms: discordance between clinical diagnoses and autopsy findings in 3,118 cases. APMIS 2001;109(11):774-780.
8. Sens MA, Zhou X, Weiland T, Cooley AM. Unexpected neoplasia in autopsies: potential implications for tissue and organ safety. ARCH PATHOL LAB MED 2009;133(12):1923-1931.
9. Klauss M, Schobinger M, Wolf I, Werner J, Meinzer HP, Kauczor HU et al. Value of three-dimensional reconstructions in pancreatic carcinoma using multidetector CT: initial results. World J Gastroenterol 2009;15(46):5827-5832.
10. Wong JC, Lu DS. Staging of pancreatic adenocarcinoma by imaging studies. Clin Gastroenterol Hepatol 2008;6(12):1301-1308.
11. Vincent A, Herman J, Schulick R, Hruban RH, Goggins M. Pancreatic cancer. LANCET 2011;378(9791):607-620.

12. Harewood GC, Wiersema MJ. Endosonography-guided fine needle aspiration biopsy in the evaluation of pancreatic masses. AM J GASTROENTEROL 2002;97(6):1386-1391.
13. Nanni L, Brahnam S, Lumini A. Combining multiple approaches for gene microarray classification. BIOINFORMATICS 2012;28(8):1151-1157.
14. Bengio S, Bengio Y. Taking on the curse of dimensionality in joint distributions using neural networks. IEEE Trans Neural Netw 2000;11(3):550-557.
15. Del CF, Jankevics A, Eisinga R, Heskes T, Hong F, Breitling R. RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. BIOINFORMATICS 2017;33(17):2774-2775.
16. Peri S, Devarajan K, Yang DH, Knudson AG, Balachandran S. Meta-analysis identifies NF-kappaB as a therapeutic target in renal cancer. PLOS ONE 2013;8(10):e76746.
17. Hou Q, Bing ZT, Hu C, Li MY, Yang KH, Mo Z et al. RankProd Combined with Genetic Algorithm Optimized Artificial Neural Network Establishes a Diagnostic and Prognostic Prediction Model that Revealed C1QTNF3 as a Biomarker for Prostate Cancer. EBIOMEDICINE 2018;32:234-244.
18. Huang S, Zhang C, Sun C, Hou Y, Zhang Y, Tam NL et al. Obg-like ATPase 1 (OLA1) overexpression predicts poor prognosis and promotes tumor progression by regulating P21/CDK2 in hepatocellular carcinoma. Aging (Albany NY) 2020;12(3):3025-3041.
19. Kamisawa T, Wood LD, Itoi T, Takaori K. Pancreatic cancer. LANCET 2016;388(10039):73-85.
20. Singhi AD, Koay EJ, Chari ST, Maitra A. Early Detection of Pancreatic Cancer: Opportunities and Challenges. GASTROENTEROLOGY 2019;156(7):2024-2040.
21. Cheng Y, Wang K, Geng L, Sun J, Xu W, Liu D et al. Identification of candidate diagnostic and prognostic biomarkers for pancreatic carcinoma. EBIOMEDICINE 2019;40:382-393.
22. Wu M, Li X, Zhang T, Liu Z, 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.
23. Bartsch DK, Gercke N, Strauch K, Wieboldt R, Matthäi E, Wagner V et al. The Combination of MiRNA-196b, LCN2, and TIMP1 is a Potential Set of Circulating Biomarkers for Screening Individuals at Risk for Familial Pancreatic Cancer. J CLIN MED 2018;7(10).
24. Slater EP, Fendrich V, Strauch K, Rospleszcz S, Ramaswamy A, Mätthai E et al. LCN2 and TIMP1 as Potential Serum Markers for the Early Detection of Familial Pancreatic Cancer. TRANSL ONCOL 2013;6(2):99-103.
25. Kaur S, Sharma N, Krishn SR, Lakshmanan I, Rachagani S, Baine MJ et al. MUC4-mediated regulation of acute phase protein lipocalin 2 through HER2/AKT/NF-κB signaling in pancreatic cancer. CLIN CANCER RES 2014;20(3):688-700.
26. Gomez-Chou SB, Swidnicka-Siergiejko AK, Badi N, Chavez-Tomar M, Lesinski GB, Bekaii-Saab T et al. Lipocalin-2 Promotes Pancreatic Ductal Adenocarcinoma by Regulating Inflammation in the Tumor Microenvironment. CANCER RES 2017;77(10):2647-2660.
27. Hanai J, Mammoto T, Seth P, Mori K, Karumanchi SA, Barasch J et al. Lipocalin 2 diminishes invasiveness and metastasis of Ras-transformed cells. J BIOL CHEM 2005;280(14):13641-13647.
28. Tong Z, Chakraborty S, Sung B, Koolwal P, Kaur S, Aggarwal BB et al. Epidermal growth factor down-regulates the expression of neutrophil gelatinase-associated lipocalin (NGAL) through E-cadherin in pancreatic cancer cells. CANCER-AM CANCER SOC 2011;117(11):2408-2418.

29. Venkatesha S, Hanai J, Seth P, Karumanchi SA, Sukhatme VP. Lipocalin 2 antagonizes the proangiogenic action of ras in transformed cells. MOL CANCER RES 2006;4(11):821-829.
30. Coothankandaswamy V, Cao S, Xu Y, Prasad PD, Singh PK, Reynolds CP et al. Amino acid transporter SLC6A14 is a novel and effective drug target for pancreatic cancer. Br J Pharmacol 2016;173(23):3292-3306.
31. Sikder M, Yang S, Ganapathy V, Bhutia YD. The Na(+)/Cl(-)-Coupled, Broad-Specific, Amino Acid Transporter SLC6A14 (ATB(0,+)): Emerging Roles in Multiple Diseases and Therapeutic Potential for Treatment and Diagnosis. AAPS J 2017;20(1):12.
32. Bhutia YD, Hung SW, Patel B, Lovin D, Govindarajan R. CNT1 expression influences proliferation and chemosensitivity in drug-resistant pancreatic cancer cells. CANCER RES 2011;71(5):1825-1835.
33. Andersson R, Aho U, Nilsson BI, Peters GJ, Pastor-Anglada M, Rasch W et al. Gemcitabine chemoresistance in pancreatic cancer: molecular mechanisms and potential solutions. Scand J Gastroenterol 2009;44(7):782-786.
34. Sun LR, Li SY, Guo QS, Zhou W, Zhang HM. SPOCK1 Involvement in Epithelial-to-Mesenchymal Transition: A New Target in Cancer Therapy? CANCER MANAG RES 2020;12:3561-3569.
35. Li J, Ke J, Fang J, Chen JP. A potential prognostic marker and therapeutic target: SPOCK1 promotes the proliferation, metastasis, and apoptosis of pancreatic ductal adenocarcinoma cells. J CELL BIOCHEM 2020;121(1):743-754.
36. Wight TN, Kinsella MG, Evanko SP, Potter-Perigo S, Merrilees MJ. Versican and the regulation of cell phenotype in disease. Biochim Biophys Acta 2014;1840(8):2441-2451.
37. Du WW, Yang W, Yee AJ. Roles of versican in cancer biology--tumorigenesis, progression and metastasis. HISTOL HISTOPATHOL 2013;28(6):701-713.
38. Wu YJ, La Pierre DP, Wu J, Yee AJ, Yang BB. The interaction of versican with its binding partners. CELL RES 2005;15(7):483-494.

Figures

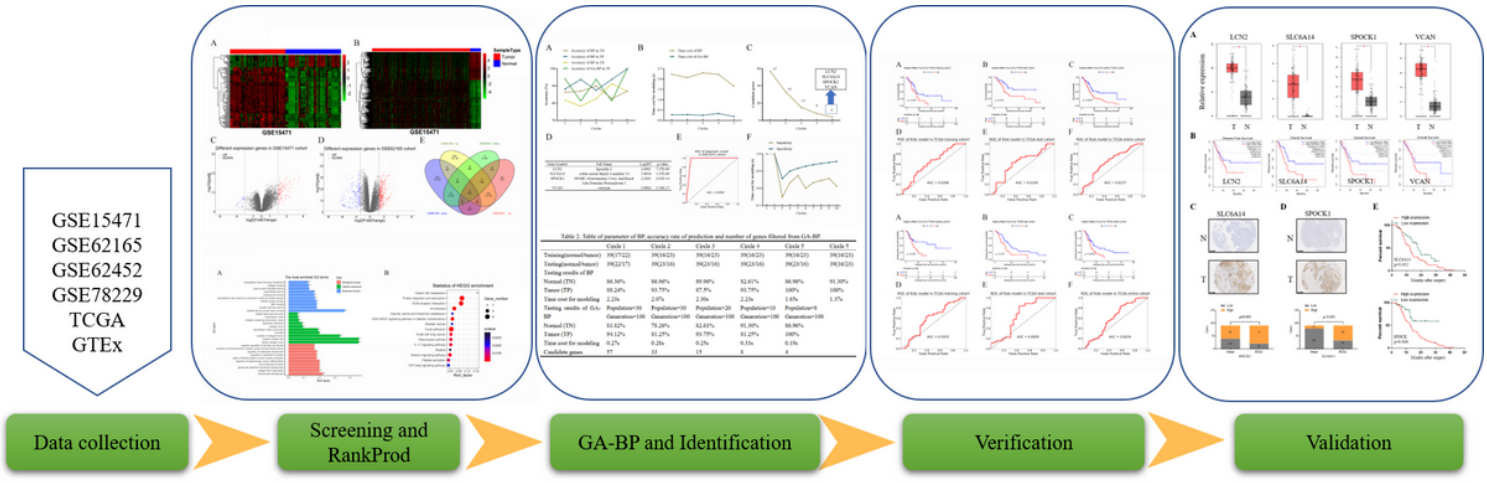


Figure 1

Flowchart of the data analysis process and validation of candidate genes.

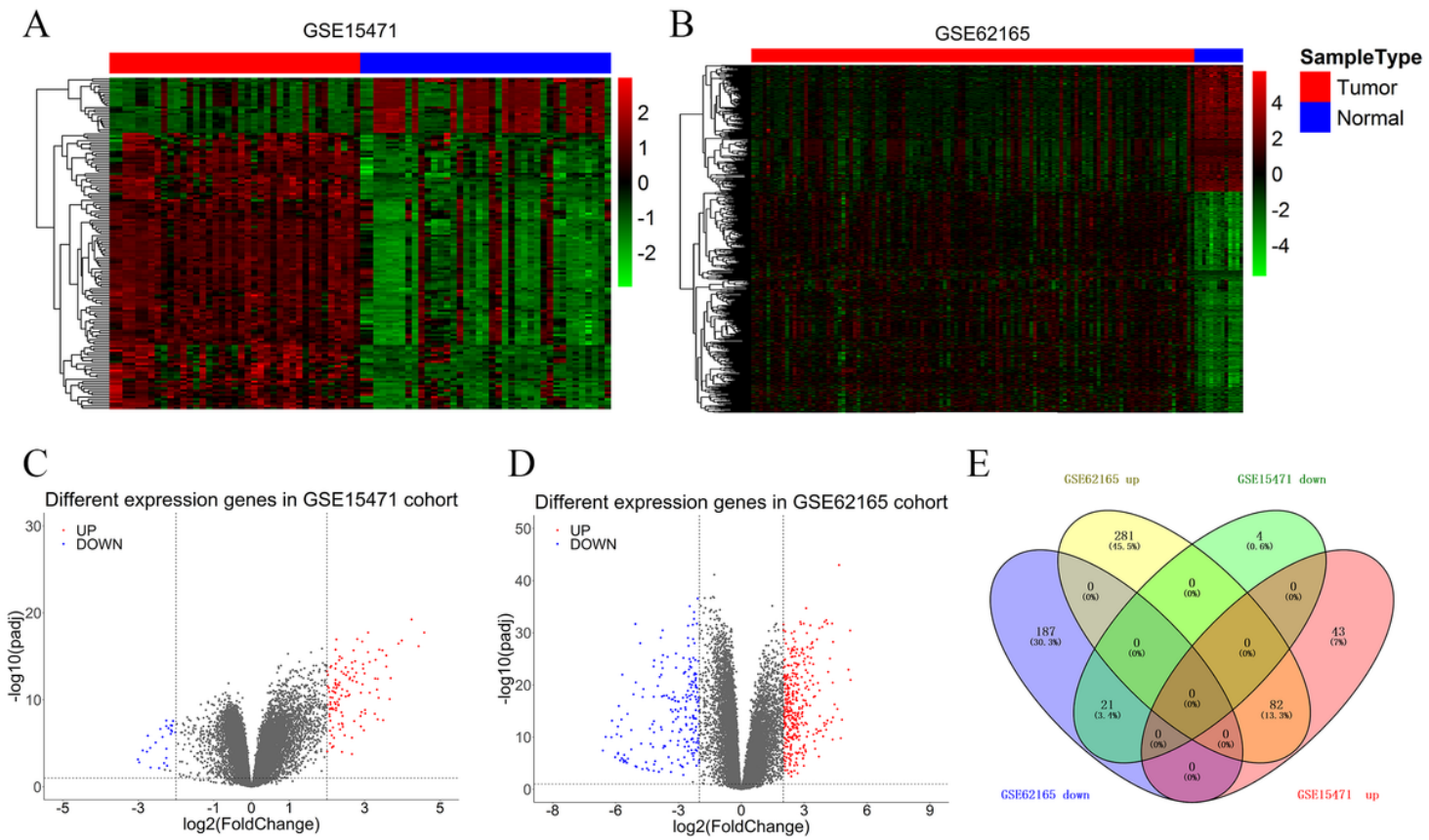


Figure 2

Identification of differentially expressed genes (DEGs) in PC. (A-B) Heatmap of DEGs in GSE15471 (A) and GSE62165 datasets (B). (C-D) Volcano plot of DEGs in GSE15471 (C) and GSE62165 datasets (D). (E) Venn diagram to identify co-regulated DEGs between GSE15471 and GSE62165 datasets.

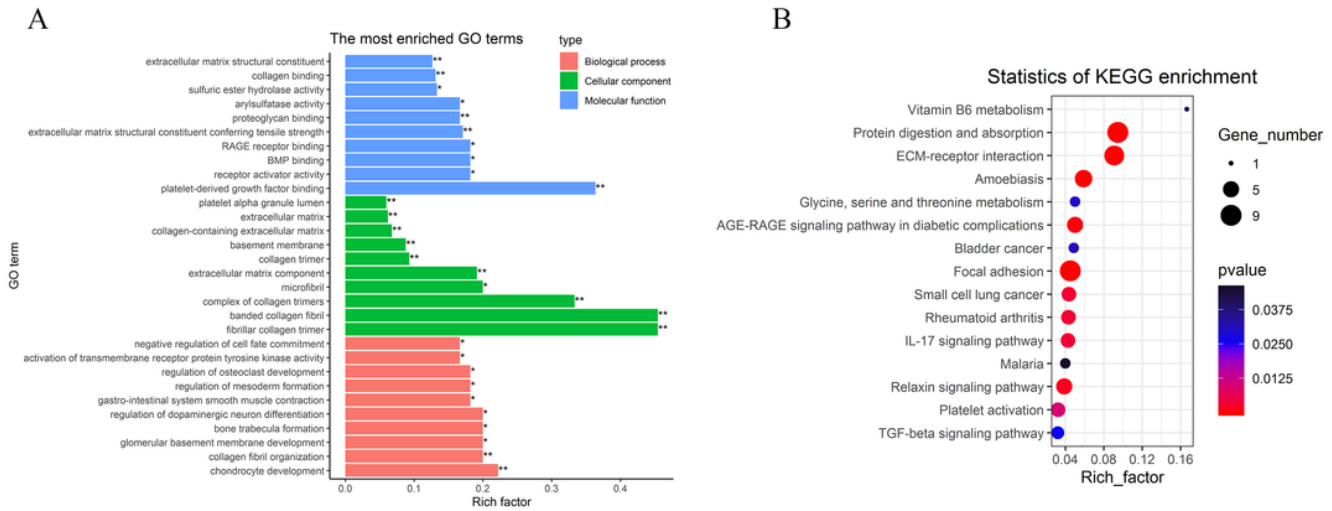
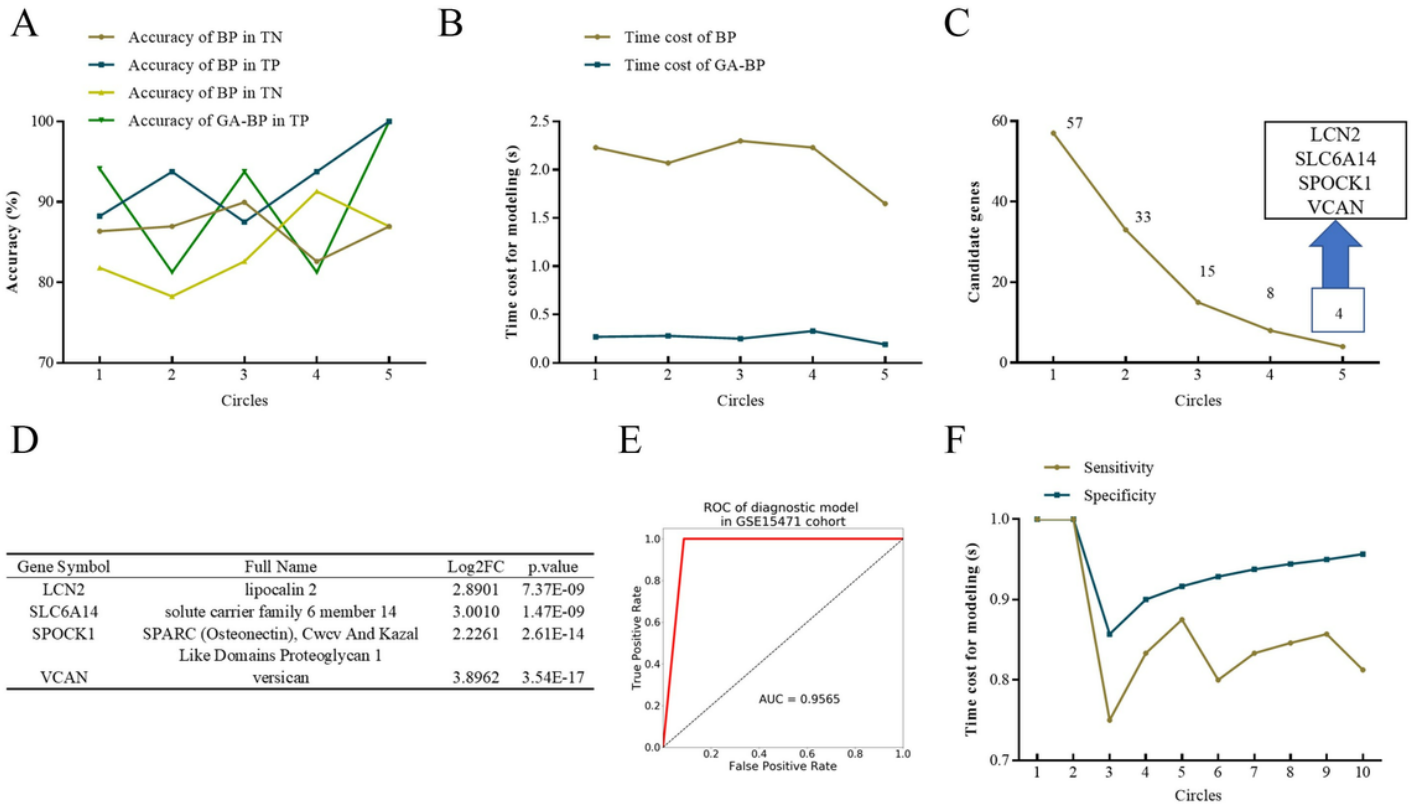


Figure 3

Gene Ontology (A) and Kyoto Encyclopedia of Genes and Genomes (B) pathway analysis to explore functions of the DEGs.



Gene Symbol	Full Name	Log2FC	p.value
LCN2	lipocalin 2	2.8901	7.37E-09
SLC6A14	solute carrier family 6 member 14	3.0010	1.47E-09
SPOCK1	SPARC (Osteonectin), Cwcv And Kazal Like Domains Proteoglycan 1	2.2261	2.61E-14
VCAN	versican	3.8962	3.54E-17

Figure 4

GA-BP to screen for candidate genes. (A) Predictive accuracy of each round of modeling of BP and GA-BP for diagnosis of PC. (B) Time cost for each round of modeling of BP and GA-BP. (C-D) Four candidate genes (LCN2, SLC6A14, SPOCK1, and VCAN) were obtained for diagnosis of PC. (E) Receiver Operating Characteristic Curve (ROC) revealed that the Area Under Curve (AUC) of the model was 0.9565. (F) Cross-validation for 10 times to test the sensitivity and specificity of the model.

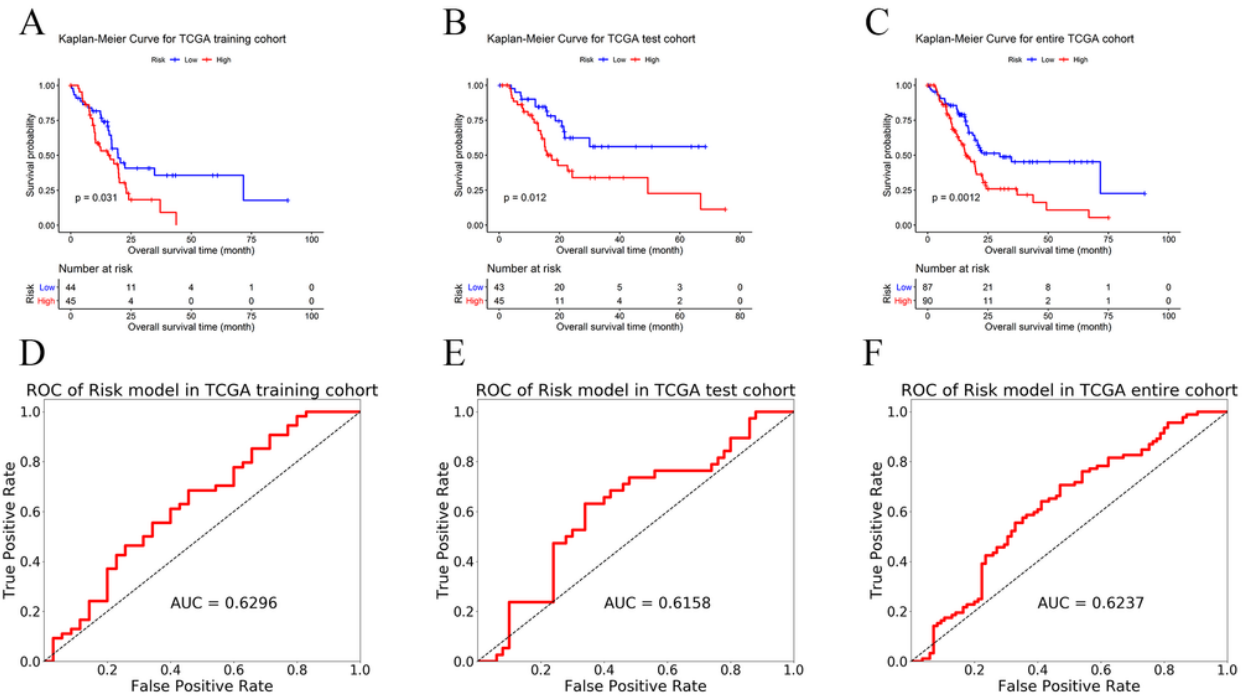


Figure 5

Prognosis prediction capacity of the gene signature for overall survival (OS) in pancreatic cancer (PC). (A-C) Kaplan-Meier analyses of OS in the training cohort (A), testing cohort (B), and entire TCGA cohort (C). (D-F) Receiver operating characteristic (ROC) of gene signature for OS. Area under the curve (AUC) was 0.6296 in the training cohort (D), 0.6158 in the testing cohort (E), and 0.6237 in the entire TCGA cohort (F).

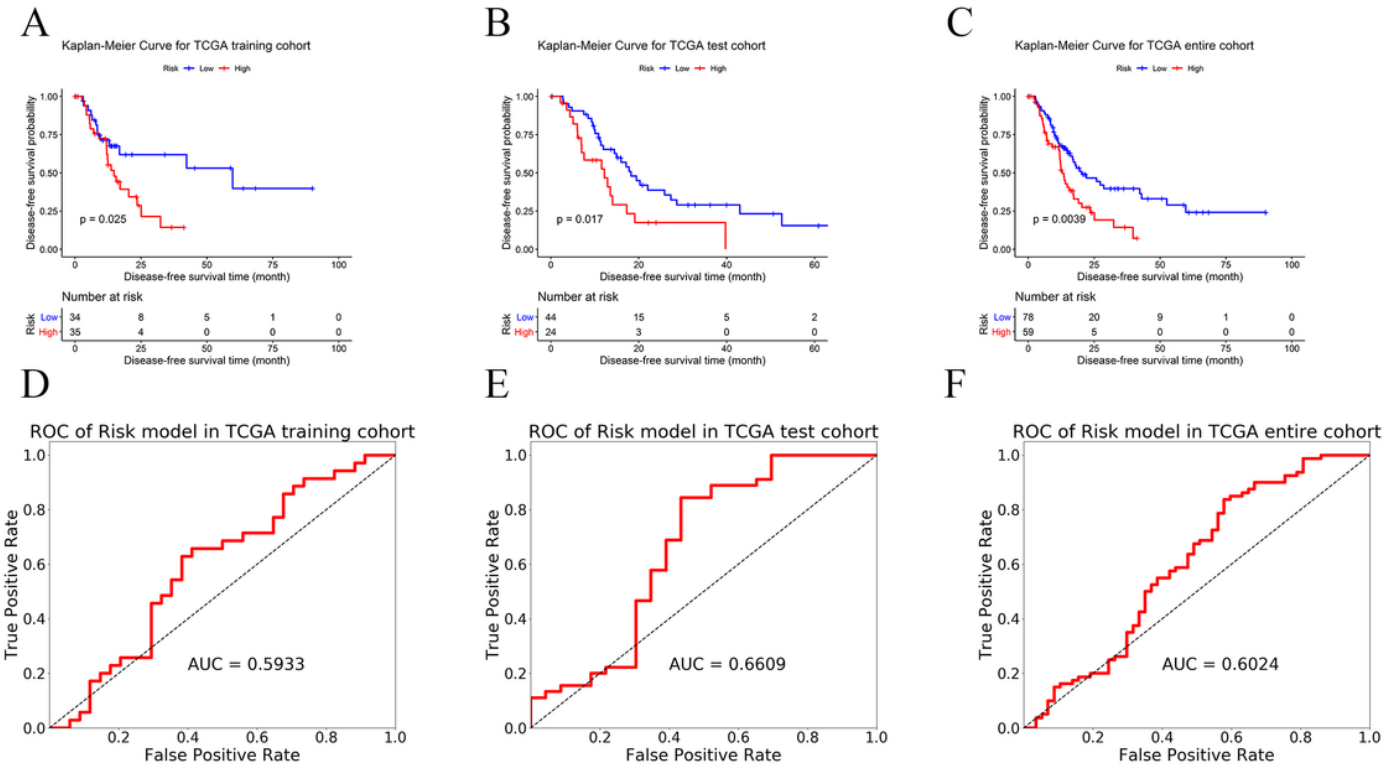


Figure 6

Prognosis capacity of the gene signature for disease-free survival (DFS) in pancreatic cancer (PC). (A-C) Kaplan-Meier analyses of DFS in the training cohort (A), testing cohort (B), and entire TCGA cohort (C). (D-F) Receiver operating characteristic (ROC) of gene signature for DFS. Area under the curve (AUC) was 0.5933 in the training cohort (D), 0.6609 in the testing cohort (E), and 0.6024 in the entire TCGA cohort (F).

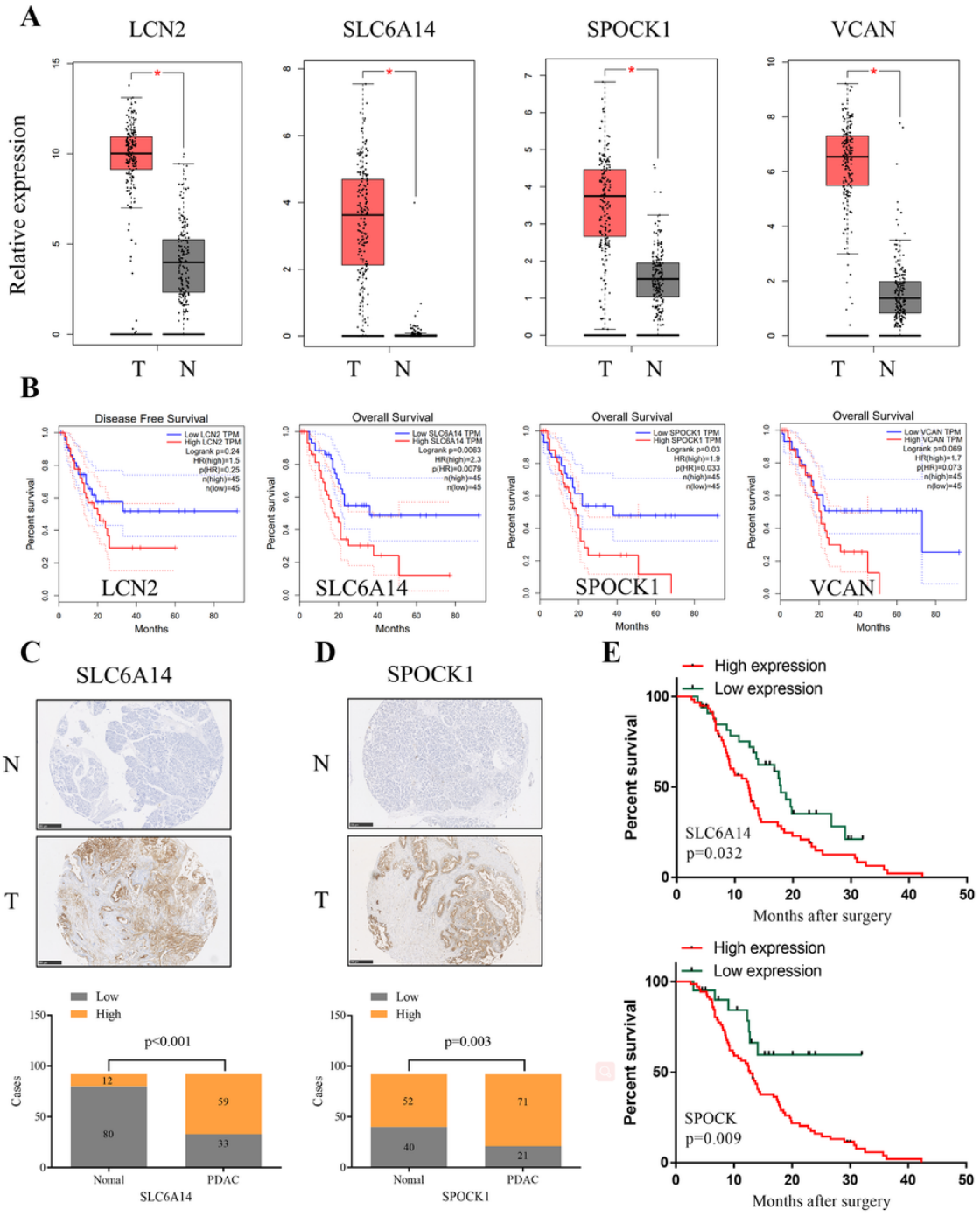


Figure 7

Validation of the four candidate genes in pancreatic cancer (PC) tissues. (A) The expression of the four candidate genes in the TCGA PC cohort and GTEx pancreas samples. (B) Kaplan-Meier analyses for the four candidate genes of overall survival (OS) in the TCGA cohort. (C-D) Protein expression of SLC6A14 (C) and SPOCK1 (D) in 92 clinical

PC tissues and paired normal pancreatic tissues using immunohistochemistry. (E) Kaplan-Meier analyses for SLC6A14 and SPOCK1 of (OS) in the clinical cohort.

Supplementary Files

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

- FigureS1.tiff20210116164142391.png