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\textsuperscript{a,b,c,1}, Zuyi Ma\textsuperscript{a,b,d,1}, Qi Zhou\textsuperscript{e,f,1}, Shujie Wang\textsuperscript{a,b,g,1}, Qian Yan\textsuperscript{a,b,c}, Hongkai Zhuang\textsuperscript{a,b,d}, Zixuan Zhou\textsuperscript{a,b,c}, Chunsheng Liu\textsuperscript{a,b,d}, Zhongshi Wu\textsuperscript{a,b,g}, Jinglin Zhao\textsuperscript{h}, Shanzhou Huang\textsuperscript{a,b,c,g,1}, Chuanzhao Zhang\textsuperscript{a,b,c,g,1}, Baohua Hou\textsuperscript{a,b,c,g,**,***}

\textsuperscript{a} Department of General Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China
\textsuperscript{b} Heyuan People's Hospital, Heyuan 517000, China
\textsuperscript{c} School of Medicine, South China University of Technology, Guangzhou 510065, Guangdong Province, China
\textsuperscript{d} Shantou University of Medical College, Shantou 515000, China
\textsuperscript{e} Department of General Surgery, Hui Ya Hospital of The First Affiliated Hospital, Sun Yat-Sen University, Huizhou, Guangdong 516081, China
\textsuperscript{f} Department of Liver Surgery, The First Affiliated Hospital of Sun Yat-Sen University, Guangzhou 510080, China
\textsuperscript{g} The Second School of Clinical Medicine, Southern Medical University, Guangzhou 510515, Guangdong Province, China
\textsuperscript{h} Department of Gastrointestinal Surgery, Jiangmen Central Hospital, Affiliated Jiangmen Hospital of Sun Yat-sen University, Jiangmen, China

ARTICLE INFO

Keywords:
Pancreatic cancer
Biomarker
Diagnosis
Prognosis
Machine learning

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 Limma Package, 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. Finally, in vitro studies were performed to assess the function of SLC6A14 and SPOCK1 on pancreatic ductal adenocarcinoma (PDAC) cells proliferation and apoptosis.

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. In PDAC cell line, SLC6A14 or SPOCK1 knockdown inhibited cells proliferation, migration and promoted cells apoptosis.

Conclusions: Using Limma Package and GA-ANN, we developed and validated a diagnostic and prognostic gene signature that yielded excellent predictive capacity for PC patients' survival. In vitro studies were further conducted to verify the functions of SLC6A14 and SPOCK1 in PC progression.

* Corresponding authors.
** Corresponding authors.
*** Corresponding authors.

E-mail addresses: hshanzh@163.com (S. Huang), zhangchuanzhao@gdph.org.cn (C. Zhang), bbb1000@126.com (B. Hou).

1 These authors contributed equally to the article and should be considered as co-first authors.
1. 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 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 [6, 7].

Several diagnostic tools are available in clinical practice; these include abdominal ultrasonography, tri-phasic CT (criteria for diagnosis and staging) [8, 9] and magnetic resonance imaging (MRI) of the abdomen [10], as well as endoscopic ultrasound-guided fine-needle aspiration cytology [11]. Biopsy is of great value for diagnosis, and its sensitivity is about 80% [11]. 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 [12]. Larger sample sizes and the use of machine-learning algorithm has reduced such errors effectively [12, 13].

Many studies also employed integrated various data to increase their sample size and thus identified promising biomarkers [14]. 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 [15]. Hou et al. developed a diagnostic and prognostic model, and identified C1QTNF3 as a biomarker for prognostic prediction of prostate cancer via an artificial neural network [16].

Herein, we applied artificial intelligence (AI)-based methods to develop a model comprising a small gene set that may be used to diagnose and predict survival outcome. First, we expanded the sample size by integrating data from various independent datasets using Limma Package. Next, we applied a genetic algorithm-artificial neural network (GA-ANN) model to screen for candidate genes and construct a predictive model for PC. Limma Package 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.

2. Methods

2.1. 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 were originally Log2(FPKM+1), and were then converted to Log2(TPM+1). We used the Affy package of R and gcrlRNA 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.

2.2. 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.

2.3. 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.

2.4. 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. TP 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.

2.8. Immunohistochemistry and reagents

Genes expression and pathological stage in PC. index.html) was also used to determine the relationship between these a better prediction effect. GEPIA database (http://gepia.cancer-pku.cn/index.html) was used in the study.

2.7. ROC curve analysis and pathological stage analysis

And area under the curve (AUC) of the receiver operating character-2.5. 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 (\beta_i x_i).

\( \beta_i \) is the Cox regression coefficient of the ith variable and \( x_i \) 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 \( \beta_i \) is the univariate Cox regression coefficient of the ith mRNA.

2.6. Patients, samples, and follow-up

We obtained tissue samples for validation from patients who had been diagnosed with PDAC pathologically after surgery at Guangdong Pro-vincial 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.

2.5. Prognostic index of gene signature in prognosis of survival of PC

2.4. RT-qPCR and western blot

RT-qPCR and Western Blot were conducted as previously described [21]. RT-qPCR was performed with the following primer pairs:

\[ \text{Forward: 5'-CGGCAAUAAAGGCAAACCA-3', Reverse: 5'-CTCAATACTCCTCCGAC-3'} \]

\[ \text{Forward: 5'-GGGGAGACACCTCGGTTGAA-3', Reverse: 5'-CTTCACCGGCTCCGTAA-3'} \]

\[ \text{Forward: 5'-GGGAGGTTGCTGGGCAAA-3', Reverse: 5'-GAGGTCCTCCTCCGTG-3'} \]

The antibodies used for western blot (SLC6A14[NBP2-93247], SPOCK1 [ab229935]) were respectively purchased form Abcam and Novus.

2.11. CCK-8 and transwell assays

For CCK-8 assays, PANc-1 cells were seeding in 96-well plates with 1.5 \( \times \) \( 10^4 \) cells per well and cultured for 0, 24, 48 and 72 h. The CCK-8 solution was added into the cells and incubated for 2 h at 37 °C. Then PANc-1 cells were measured by OD450. For Transwell assays, PANc-1 cells (1.0 \( \times \) \( 10^5 \)) were resuspended in DMEM with 10% FBS in the upper chamber, and the lower was added DMEM with 20% FBS. After one day, migrated cells in the lower chambers were fixed in 4% paraformaldehyde and then stained with 0.1% crystal violet. Migrated cells were imaged using an inverted microscope and calculated from three different fields.

2.12. Flow cytometry assay of apoptosis

Trypsin Solution without EDTA (C0205, Beyotime) was used to dissociate PANc-1 cells. After collection of 1 \( \times \) \( 10^6 \) PANc-1, the cells were washed twice with cold PBS and suspended with 400ul Annexin V. Then 5ul Annexin V PE staining solution was added to the suspension. After the cells were incubated for 5–10 min at 2–8 °C under dark conditions, 5-10ul 7-AAD staining solution was added and incubated for 1–3 min. At last, PANc-1 cells was analyzed through flow cytometry (BD Biosciences, USA).
2.13. 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. Cox-regression analysis was used in the univariate analysis for OS and DFS.

2.14. Ethical description

This study obtained approval from the Ethics Association of Guangdong Provincial People’s Hospital or the First Affiliated Hospital of Sun Yat-sen University and all enrolled patients provided written informed consent before participating in this research.

3. Results

3.1. 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.

3.2. Identification and functional enrichment of DEGs

The flowchart of the data analysis process was shown in Figure 1. Heatmap analysis showed that DEGs were differentially expressed between normal and cancer tissues (Figure 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 (Figure 2C and D). We only screened 30 DEGs from GSE62452, making it not conducive to the establishment and validation for diagnostic models. So GSE15471 and GSE62165 datasets were used to identify DEGs by comparing pancreatic carcinoma tissues with normal pancreatic tissues (Figure 2). Furthermore, we identified 103 co-regulated (82 co-upregulated and 21 co-downregulated) DEGs between the two datasets (Figure 2E). GO and KEGG analyses were performed to identify functions of the DEGs (Figure 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”.

3.3. 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 (Figure 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 (Figure 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 (Figure 4E). We performed 10 rounds of cross-validation to test the stability of the model (Figure 4F). The results showed that the sensitivity and specificity were above 0.750.

Then, the diagnostic 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 (Tables 4 and 5).

3.4. 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 (95 cases) and a test cohort (82 cases). The prognostic score was calculated in the training cohort with a median value of 9.1566. Cases with prognostic scores higher than 9.1566 were identified as belonging to the high-risk group. Cases with prognostic scores lower than 9.1566 were identified as belonging to the low-risk group. The high-risk group (48 cases) showed a significantly poorer survival than the low-risk group (47 cases) in the training cohort (p = 0.027, Figure 5A). The AUC of the ROC is 0.633 (Figure 5D). Then, the risk model was validated in the testing cohort (p = 0.029, Figure 5B) and the entire TCGA cohort (p = 0.001, Figure 5C), both of which showed inferior survival outcome in the high-risk groups with AUC higher than 0.600 of the ROC (Figure 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 (72 cases) and a test cohort (65 cases). The prognostic score was calculated in the training cohort with a median value of 10.5966. Forty-five cases with prognostic scores higher than 10.5966 were identified as belonging to the high-risk group; 27 cases with prognostic 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 = 0.002, Figure 6A). The AUC of the ROC is 0.576 (Figure 6D). Then, the risk

Table 1. Characteristics of datasets used in the study.

| Dataset | Tumor | Normal | OS | DFS |
|---------|-------|--------|----|-----|
| GSE15471 | 36    | 36     | -  | -   |
| 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.
model was validated in the testing cohort \( (p = 0.445, \text{Figure 6B}) \) and the entire TCGA cohort \( (p = 0.004, \text{Figure 6C}) \), both of which showed inferior survival outcomes in the high-risk groups with AUC of the ROC higher than 0.600 (Figure 6E and F). Although the difference of testing cohort is not significant, the training cohort and entire TCGA cohort support the significant difference of high and low. Besides, univariate analyses were performed and it’s showed that prognostic score, age, T stage, N stage, grade and AJCC stage were associated with OS and DFS of patients (Table 6).

3.5. Validation of candidate genes in PC tissues

The expression of 4 candidate genes in the TCGA cohort and GTEx samples is shown in Figure 7A and Figure S2. All were expressed at higher levels in PC tissues than in normal pancreatic tissues. The ROC curves indicated that not all the AUC results of these 4 genes were satisfactory in the TCGA cohort (Figure S3 A), however, we found that all of the AUC results were up to 0.9 in the TCGA + GTEx cohort (Figure S3 B). Next, to evaluate the prognostic value of four genes in PDAC, we searched for GEPIA dataset and found that higher expression of LCN2 \( (p = 0.000272) \) and SLC6A14 \( (p = 0.0216) \) were associated with more advanced tumor pathological stage (Figure S4). Furthermore, the high expression of SLC6A14 and SPOCK1 is associated with inferior OS (Figure 7B). Therefore, SLC6A14 and SPOCK1 were chosen to be 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 (Figure 7C-D and Figure S5 A-B). 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 (Figure 7E). Overall, these data suggest that SLC6A14 and SPOCK1 were constantly overexpressed in PC.

3.6. Downregulation of SLC6A14 or SPOCK1 inhibited PDAC cells proliferation, migration and promoted apoptosis

To further explore the function of SLC6A14 and SPOCK1 in PC, three different siRNAs targeting SLC6A14 and SPOCK1 were respectively
transfected into PANC-1 cells. RT-qPCR and Western blot analysis indicated that the level of SLC6A14 and SPOCK1 in the cells respectively transfected with SLC6A14-siRNA2 and SPOCK1-siRNA1 was significantly lower than that in the NC group (Figure 8A, B). Knocking down SLC6A14 (A) and SPOCK1 (B) with siRNA reduced their expression in PANC-1 cells (Figure S1A, B). With the silencing of SLC6A14 or SPOCK1, the cells proliferation was repressed, as showed in the CCK-8 assays results (Figure 8C). Furthermore, migratory assays showed that down-regulation of SLC6A14 and SPOCK1 remarkably reduced the numbers of migrated cells in vitro (Figure 8D). Moreover, flow cytometry results indicated that depletion of SLC6A14 or SPOCK1 increased the percentage of apoptotic PANC-1 cells (Figure 8E). To sum up, these findings suggested that SLC6A14 and SPOCK1 are crucial to proliferation, migration and apoptosis in PDAC.

We applied AI-based methods to develop a BP model, which evaluated the expression and effects of LCN2, SLC6A14, SPOCK1 and VCAN in PDAC. Subsequently, SLC6A14 and SPOCK1 were chosen to conduct a series of in vitro studies to verify their function in PC. However, our study had some limitations. For example, though our results showed that SLC6A14 and SPOCK1 may serve as a prognostic biomarker for PC, further prospective studies and in vivo experiments are needed to verify our results and further reveal the potential molecular mechanisms.
4. Discussion

Despite consistent progress in the diagnosis and management of PC, few breakthroughs for effective biomarkers and treatment strategies have emerged [22]. Identification of the biological and molecular mechanisms as well as discovery of timely diagnostic, prognostic, and therapeutic biomarkers for PC is therefore necessary [23]. 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 [24]. 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 [25].

Besides, Sima Kalantari et al used bioinformatics tools to recognize 10 key genes associated with SARS-CoV-2 infection in Caco-2 cell was significantly overexpressed in colon cancer [26]. 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, Limma Package and GA-ANN were used to construct a model. Follow, four DEGs (LCN2, SLC6A14, SPOCK1, and VCAN) were identified, and a four-gene signature was developed by our data processing system. Subsequently, SLC6A14 and SPOCK1 were selected to carry out a series of in vitro studies to verify their potential function in PC. 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 [27, 28, 29]; the increased expression of LCN2 has also been observed in various mouse models of PC [30]. However, LCN2 depletion was also found in poorly

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         | 1.2–2.5 | 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 Training | OS Test | OS Entire | DFS Training | DFS Test | DFS Entire |
|--------------------|-------------|---------|-----------|--------------|----------|-----------|
| Risk score         | 0.6296     | 0.6158  | 0.6237    | 0.5933       | 0.6099   | 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    |

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 0.633 in the training cohort (D), 0.664 in the testing cohort (E), and 0.645 in the entire TCGA cohort (F). Log-rank test was performed for Kaplan-Meier analyses.
differentiated PC (mesenchymal-like) and considered to be essential for invasion and metastasis [31]. 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) [32]. LCN2 is also reported to inhibit angiogenesis and cause hypovascular conditions in tumor microenvironment [33]. 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 \( \text{Na}^+ \)- and \( \text{Cl}^-/\text{C}^0^- \)-dependent solute transport molecule that activates the transport of neutral and basic amino acids [34]. Previous studies have revealed that the expression of SLC6A14 is increased in cervical cancer, colorectal cancer, breast cancer, as well as PC [35]. In our study, we also confirmed its increased expression in PC tissue, its correlation with poor survival of PC patients and its function in promoting PC progression. It has been shown that blockade of SLC6A14 with either \( \alpha \)-methyl-\( L \)-tryptophan (\( \alpha \)-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 [34]. Thus, \( \alpha \)-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) [36,37]. 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 \( \text{Ca}^{2+} \)-binding proteoglycan family members, was shown to be highly expressed in several cancer types [38]. Studies have indicated that SPOCK1-mediated EMT regulates proliferation and invasion in various malignancies [38]. A recent study has shown that SPOCK1

![Figure 6](image)

**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 0.576 in the training cohort (D), 0.679 in the testing cohort (E), and 0.625 in the entire TCGA cohort (F).

**Table 6.** Univariate 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 (0.3–3.5)        | 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    |

The difference is analyzed by Cox-regression analysis.
induces EMT to promote PC metastasis and inactivates the PI3K/Akt signaling pathway to attenuate PC cell apoptosis, in vitro and in vivo [39]. Our work showed that the expression level of \textit{SPOCK1} was up-regulated and associated with a shorter overall survival in PC. In vitro studies also demonstrated knockdown of \textit{SPOCK1} would inhibit PDAC cell proliferation, migration and promote apoptosis. For \textit{VCAN}, an ECM macromolecule, induces several biological activities such as apoptosis and is known to accumulate in several types of cancers [40]. It has been reported that \textit{VCAN} interacts with numerous ECM components including hyaluronan, fibronectin, thrombospondin 1, and fibrillin to create an

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 \textit{SLC6A14} (C) and \textit{SPOCK1} (D) in 92 clinical PC tissues and paired normal pancreatic tissues using immunohistochemistry. (E) Kaplan-Meier analyses for \textit{SLC6A14} and \textit{SPOCK1} of (OS) in the clinical cohort. Log-rank test was performed for Kaplan-Meier analyses. Scale bars: 250 μm. Chi-square tests was used to compare categorical data.
active biopolymer that affects cell morphosis, adhesion, proliferation, and migration \[41, 42\]. 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. Primarily, PC is a heterogeneous tumor at the genetic and molecular level; therefore, the established model needs to be further validated in other clinical studies.
In addition, this study used publicly available datasets so that the outcomes may be influenced by the quality of these data to some extent. Besides, the relationship between LCN2, SLCE6A14, SPOCK1, and VCAN in immune regulation in PC is still unclear. To better understand the potential roles of these four genes in PC, further experiments and in vivo studies are necessary to validate our results and explore the underlying molecular mechanisms.

In conclusion, the application of Limma Package 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.

5. Statements

A preprint of this article has previously been published in Research Square [39]. The following link: https://www.researchsquare.com/article/rs-151851/v1 (DOI: https://doi.org/10.21203/rs.3.rs-151851/v1).

Declaration

Author contribution statement

Zhencong Li, Zuoyi Ma, Qi Zhou, and Shujie Wang: Performed the experiments; Wrote the paper.
Qian Yan, Hongkai Zhuang, and Zixuan Zhou: Contributed reagents, materials, analysis tools or data.
Chunsheng Liu, Zhongshi Wu and Jinglin Zhao: Analyzed and interpreted the data.
Baohua Hou, Chuanzhao Zhang and Shanzhou Huang: Conceived and designed the experiments.

Funding statement

Chuanzhao Zhang was supported by National Natural Science Foundation of China [82102961, 82173149], Basic and Applied Basic Research Foundation of Guangdong Province [2020A1515110536], Guangdong Provincial People's Hospital [2020B09, 8200100290].
Baohua Hou was supported by National Natural Science Foundation of China [82072635 and 82072637], Science and Technology Program of Guangzhou [202102020030 and 202102020107], and Special Events Supported by Heyuan People's Hospital [YNKT202202].
Shanzhou Huang was supported by Guangdong Provincial People's Hospital [KY012021164], Basic and Applied Basic Research Foundation of Guangdong Province [2021A1515011473, 2021A1515012441].
Qi Zhou was supported by Construction of Key Specialty in Huizhou [Qi Zhou].

Data availability statement

Data associated with this study has been deposited at TCGA (https://portal.gdc.cancer.gov/repository) and GEO databases (GEO, http://www.ncbi.nlm.nih.gov/geo/).

Declaration of interest’s statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e11321.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA A Cancer J. Clin. 68 (2018) 394–424.
[2] M. Hidalgo, S. Cascini, J. Kleeff, et al., Addressing the challenges of pancreatic cancer: future directions for improving outcomes, Pancreatology 15 (2015) S8–18.
[3] S. McGuire, World Cancer Report 2014, in: Adv Nutr 2016, vol. 7, World Health Organization, International Agency for Research on Cancer, WHO Press, Geneva, Switzerland, 2015, pp. 418–419.
[4] S. Mohammed, G.N. Van Buren, W.E. Fisher, Pancreatic cancer: advances in treatment, World J. Gastroenterol. 20 (2014) 9354–9365.
[5] M.S. De La Cruz, A.P. Young, M.T. Ruffin, Diagnosis and management of pancreatic cancer, Am. Fam. Physician 89 (2014) 626–632.
[6] D.V. Averinov, J. Bjornsson, Malignant neoplasms: discordance between clinical diagnoses and autopsy findings in 3,118 cases, Apmis 109 (2001) 774–780.
[7] M.A. Sens, X. Zhou, T. Weiland, et al., Unexpected neoplasia in autopsies: potential implications for tissue and organ safety, Arch. Pathol. Lab Med. 133 (2009) 1923–1931.
[8] M. Klaun, M. Schobinger, I. Wolf, et al., Value of three-dimensional reconstructions in pancreatic carcinoma using multidetector CT: initial results, World J. Gastroenterol. 15 (2009) 5827–5832.
[9] J.C. Wongs, D.S. Lu, Staging of pancreatic adenocarcinoma by imaging studies, Clin. Gastroenterol. Hepatol. 6 (2008) 1301–1308.
[10] A. Vincent, J. Herman, R. Schlick, et al., Pancreatic cancer, Lancet 378 (2011) 607–620.
[11] G.C. Harewood, M.J. Wierisma, Endosonography-guided fine needle aspiration biopsy in the evaluation of pancreatic masses, Am. J. Gastroenterol. 97 (2002) 1386–1391.
[12] A. Nazeri, S. Brahnam, A. Lumini, Combining multiple approaches for gene microarray classification, Bioinformatics 28 (2012) 1151–1157.
[13] S. Bengio, Y. Bengio, Taking on the curse of dimensionality in joint distributions using neural networks, IEEE Trans. Neural. Network. 11 (2000) 550–557.
[14] C.P. Del, A. Jankevics, R. Eisinga, et al., RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets, Bioinformatics 33 (2017) 2774–2775.
[15] S. Peri, K. Devarajan, D.H. Yang, et al., Meta-analysis identifies NF-kappab as a therapeutic target in renal cancer, PLoS One 8 (2013), e76746.
[16] Q. Hou, Z.T. Bing, C. Hu, 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, ElbioMedicine 32 (2018) 234–244.
[17] S. Huang, C. Zhang, C. Sun, et al., Obs-like ATPase 1 (OLA1) overexpression predicts poor prognosis and promotes tumor progression by regulating P21/CDK2 in hepatocellular carcinoma, Aging (Albany NY) 12 (2020) 3025–3041.
[18] A. Sadeghi, R. Roudi, A. Mirzaee, et al., CD44 epithelial isoform inversely associates with invasive characteristics of colorectal cancer, Biomarkers Med. 13 (2019) 419–426.
[19] S. Sedaghat, E. Gheyetchian, M. Asgari, et al., Expression of cancer stem cell markers OCT4 and CD133 in transitional cell carcinoma, Appl. Immunohistochem. Mol. Morphol. 25 (2017) 196–202.
[20] E. Kalantari, M. Asgari, S. Nikpanah, et al., Co-expression of putative stem cell markers CD44 and CD133 in prostate carcinomas, Pathol. Oncol. Res. 23 (2017) 793–802.
[21] Z. Zhou, Z. Ma, Z. Li, et al., CMTM3 overexpression predicts poor survival and promotes proliferation and migration in pancreatic cancer, J. Cancer 12 (2021) 5797–5806.
[22] T. Kamiński, L.D. Wood, T. Ito, et al., Pancreatic cancer, Lancet 388 (2017) 73–95.
[23] A.D. Singh, E.J. Koay, S.T. Chari, et al., Early detection of pancreatic cancer: opportunities and challenges, Gastroenterology 156 (2019) 2024–2040.
[24] Y. Cheng, K. Wang, L. Geng, et al., Identification of candidate diagnostic and prognostic biomarkers for pancreatic carcinoma, ElbioMedicine 40 (2019) 382–393.
[25] M. Wu, X. Li, T. Zhang, et al., Identification of a nine-gene signature and establishment of a prognostic nomogram predicting overall survival of pancreatic cancer, Front. Oncol. 9 (2019) 996.
[26] S. Ghani, S. Kalanta, S.A. Mirmotalebisohi, et al., Specific regulatory motifs network in SARS-CoV-infected Caco-2 cell line, as a model of gastrointestinal infections, Cell. Reprogram. 24 (2022) 26–37.
[27] D.K. Bartosh, N. Gerecke, K. Strauch, et al., The combination of MRNA-1366, LCN2, and TIMP1 is a potential set of circulating biomarkers for screening individuals at risk for Familial pancreatic cancer, J. Clin. Med. 7 (2018).
[28] E.P. Slater, V. Fendrich, K. Strauch, et al., LCN2 and TIMP1 as potential serum markers for the early detection of Familial pancreatic cancer, Transl. Oncol. 6 (2013) 99–103.
[29] S. Kaur, N. Sharma, S.R. Krishn, et al., MUC4-mediated regulation of acute phase protein lipocalin 2 through HER2/AKT/NF-kB signaling in pancreatic cancer, Clin. Cancer Res. 20 (2014) 688–700.
[30] S.B. Goméz-Chou, A.K. Swidnicka-Siergiejko, N. Badi, et al., Lipocalin-2 promotes pancreatic ductal adenocarcinoma by regulating inflammation in the tumor microenvironment, Cancer Res. 77 (2017) 2647–2660.
[31] J. Hanai, T. Mammoto, P. Seth, et al., Lipocalin 2 diminishes invasiveness and metastasis of Ras-transformed cells, J. Biol. Chem. 280 (2005) 13641–13647.
[32] Z. Tong, S. Chakraborty, B. Sung, 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. 117 (2011) 2408–2418.
[33] S. Venkatesha, J. Hanai, P. Seth, et al., Lipocalin 2 antagonizes the proangiogenic action of ras in transformed cells, Mol. Cancer Res. 4 (2006) 821–829.

[34] V. Coothankandaswamy, S. Cao, Y. Xu, et al., Amino acid transporter SLO6A14 is a novel and effective drug target for pancreatic cancer, Br. J. Pharmacol. 173 (2016) 3292–3306.

[35] M. Sikder, S. Yang, V. Ganapathy, et al., The Na(+/)-Cl(-)-Coupled, broad-specific, amino acid transporter SLO6A14 (ATB(0,+)p): emerging roles in multiple diseases and therapeutic potential for treatment and diagnosis, AAPS J. 20 (2017) 12.

[36] Y.D. Bhutia, S.W. Hung, B. Patel, et al., CNT1 expression influences proliferation and chemosensitivity in drug-resistant pancreatic cancer cells, Cancer Res. 71 (2011) 1825–1835.

[37] R. Andersson, U. Abo, B.I. Nilsson, et al., Gemcitabine chemoresistance in pancreatic cancer: molecular mechanisms and potential solutions, Scand. J. Gastroenterol. 44 (2009) 782–786.

[38] L.R. Sun, S.Y. Li, Q.S. Guo, et al., SPOCK1 involvement in epithelial-to-mesenchymal transition: a new target in cancer therapy? Cancer Manag. Res. 12 (2020) 3561–3569.

[39] J. Li, J. Ke, J. Fang, et al., A potential prognostic marker and therapeutic target: SPOCK1 promotes the proliferation, metastasis, and apoptosis of pancreatic ductal adenocarcinoma cells, J. Cell. Biochem. 121 (2020) 743–754.

[40] T.N. Wight, M.G. Kinsella, S.P. Evanko, et al., Versican and the regulation of cell phenotype in disease, Biochim. Biophys. Acta 1840 (2014) 2441–2451.

[41] W.W. Du, W. Yang, A.J. Yee, Roles of versican in cancer biology—tumorigenesis, progression and metastasis, Histol. Histopathol. 28 (2013) 701–713.

[42] Y.J. Wu, D.P. La Pierre, J. Wu, et al., The interaction of versican with its binding partners, Cell Res. 15 (2005) 483–494.