

Screening of genes characteristic of pancreatic cancer by LASSO regression combined with support vector machine and recursive feature elimination, and immune correlation analysis

Longhui Zeng  and Zheng Chen

Abstract

Background: Pancreatic cancer is a malignant tumor of the digestive tract that shows increased mortality, recurrence, and morbidity year on year.

Methods: Differentially expressed genes between pancreatic cancer and healthy tissues were first analyzed from four datasets within the Gene Expression Omnibus (GEO). Gene ontology, disease ontology, and gene set enrichment analysis of differentially expressed genes were performed, and genes identified as characteristic of pancreatic cancer were screened using LASSO regression combined with support vector machine and recursive feature elimination (SVM–RFE). Differential analysis and receiver operating characteristic curve analysis were performed on the identified eigengenes, and validation was carried out using another dataset from the GEO database. Differences and correlations between characteristic pancreatic cancer genes and immune cells were analyzed.

Results: A total of 90 differentially expressed genes were identified by screening, and six genes characteristic of pancreatic cancer were obtained by taking the intersection of two characteristic genes identified by machine learning. Immunoassays yielded multiple immune cells associated with pancreatic cancer signature genes.

Conclusion: The six characteristic genes screened by a combination of LASSO regression and SVM–RFE are potential new biomarkers for the early diagnosis and prognosis of pancreatic cancer, and could be a novel therapeutic target.

Corresponding author:

Zheng Chen, Organ Transplant Center, Second Affiliated Hospital of Guangzhou Medical University, no. 63 Asian Games South Road, Panyu District, Guangzhou City, Guangdong 511447, China.
Email: dochenzheng@163.com

Organ Transplant Center, Second Affiliated Hospital of
Guangzhou Medical University



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Keywords

Pancreatic cancer, immunity, biomarker, machine learning, differential expression, LASSO regression

Date received: 29 September 2023; accepted: 29 January 2024

Introduction

Pancreatic cancer is a malignant tumor of the digestive tract. Although good progress has been made in its treatment during the past century, disease mortality, recurrence, and morbidity are still increasing year on year.¹ Indeed, the global incidence of pancreatic cancer is expected to increase to 18.6 cases per 100,000 by 2050, with an average annual growth rate of approximately 1.1%.² Additionally, pancreatic cancer is now the third leading cause of cancer-related death in China, with a 5-year survival rate of less than 10%.³

Because of the lack of typical clinical manifestations in the early stages, many patients with pancreatic cancer are often in the advanced stage or metastatic stage of disease when they are diagnosed in the clinic, but early diagnosis and treatment are crucial to maximize survival. Antigen 19-9 (CA19-9) is often used as a biomarker for the early diagnosis of pancreatic cancer, but the lack of absolute sensitivity and specificity of CA19-9 limits its clinical application.⁴ Specific biomarkers of pancreatic cancer have yet to be identified, so this is a key focus of current pancreatic cancer research.⁵⁻⁷

In recent years, disease risk prediction based on bioinformatics and machine learning algorithms has become a research hotspot. The rapid development of bioinformatics has also enabled the use of public data to screen genes that are characteristic for the occurrence and development of diseases, while the in-depth application of machine learning methods in the field

of medical big data has greatly improved the ability to predict disease risk. Identifying diagnostic biomarkers of pancreatic cancer through machine learning algorithms has important clinical relevance for the early prevention, diagnosis, and treatment of pancreatic cancer, as well as improving disease survival rate. However, few studies have simultaneously used the least absolute shrinkage and selection operator (LASSO) and support vector machine (SVM) to identify such biomarkers, so this was undertaken in the present study.

Methods

Data sources

Microarray datasets GSE15471, GSE16515, GSE32676, GSE55643, and GSE71729 of pancreatic cancer-related healthy tissues and tumor tissues were downloaded from the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>). After annotating the data, the “limma” package and “sva” package in R studio software were used to merge data from GSE15471, GSE16515, GSE32676, and GSE55643 as the training group. Healthy tissue samples from the training group were designated the control group, and pancreatic cancer tissue samples were designated the experimental group. GSE71729 data were used as the test group to validate pancreatic cancer genes screened by machine learning. This study followed relevant Equator guidelines, and all patient details were de-identified.

Differentially expressed genes

Genes showing differential expression between pancreatic cancer tissue and healthy tissue were identified using the “limma” and “pheatmap” packages in R studio software, with conditions set to: $\log_{FC} \text{filter} = 2$, $\text{adj.P.Val.Filter} = 0.05$. Simultaneously, a volcano plot was drawn of the differentially expressed genes obtained. Gene ontology (GO) enrichment analysis was performed using the “clusterProfiler”, “org.Hs.eg.db”, “enrichplot”, and “ggplot2” packages in R studio software, with the conditions set as: $\text{pvalueFilter} = 0.05$, $\text{qvalueFilter} = 0.05$. Disease ontology (DO) enrichment analysis was performed to identify disease names enriched by differentially expressed genes, with the conditions set as: $\text{pvalueFilter} = 0.05$, $\text{qvalueFilter} = 0.05$. Finally, gene set enrichment analysis (GSEA) of the differentially expressed genes was performed to determine active functions and pathways in the control and experimental groups.

Signature genes

First, genes characteristic of pancreatic cancer were screened by LASSO regression, and the “glmnet” package in R studio software was used to construct a LASSO regression model. Next, screening for characteristic genes was carried out by SVM-RFE, using “e1071”, “kernlab”, and “caret” packages in R studio software. The characteristic genes identified both by LASSO regression and SVM-RFE were analyzed by intersection analysis using the “venn” package in R studio software to identify intersection genes.

Validation and immune correlation analysis

“limma” and “ggpubr” packages in R studio software were used to draw a differential boxplot of eigengenes identified from the training group to determine expression differences between control and

experimental groups. Validation group GSE71729 data were then used to validate the signature genes by determining expression differences between control and experimental groups. Next, receiver operating characteristic (ROC) curve analysis was performed in the training and test groups using the “pROC” package in R studio software to determine the accuracy of characteristic genes screened by machine learning as diagnostic genes of pancreatic cancer. Then, single-sample GSEA analysis was performed to obtain immune cell scores, using “reshape2”, “ggpubr”, “limma”, “GSEABase”, and “GSVA” packages in R studio software to draw heatmaps visualizing immune cells seen in control and experimental groups. Differential analysis was performed using the “vioplot” package in R studio software to obtain immune cells showing significant differences in gene expression between the experimental group and control group. Finally, “limma”, “reshape2”, “tidyverse”, and “ggplot2” packages in R studio software were used to perform correlation analysis between immune cells and the genes characteristic of pancreatic cancer identified in this study.

Results

Differentially expressed genes and analysis

A total of 90 genes showing differential expression between pancreatic cancer tissue and healthy tissue were screened from GSE15471, GSE16515, GSE32676, and GSE55643 datasets (Figure 1a). GO enrichment analysis revealed that these genes were mainly enriched in extracellular matrix organization, extracellular structure organization, and external encapsulating structure organization (Figure 1b). DO disease enrichment analysis showed that the genes were mainly enriched in female reproductive organ cancer, ovarian cancer, and

malignant ovarian surface epithelial–stromal neoplasm (Figure 1c). GSEA enrichment analysis identified active functions in the experimental group (Figure 1d) and control group (Figure 1e), and active pathways in the experimental group (Figure 1f) and control group (Figure 1g).

Pancreatic cancer signature genes

LASSO regression screening identified a total of 13 genes characteristic of pancreatic cancer (Figure 2a), while SVM–RFE identified a total of 19 genes characteristic of pancreatic cancer (Figure 2b). Intersecting the genes identified by both methods of

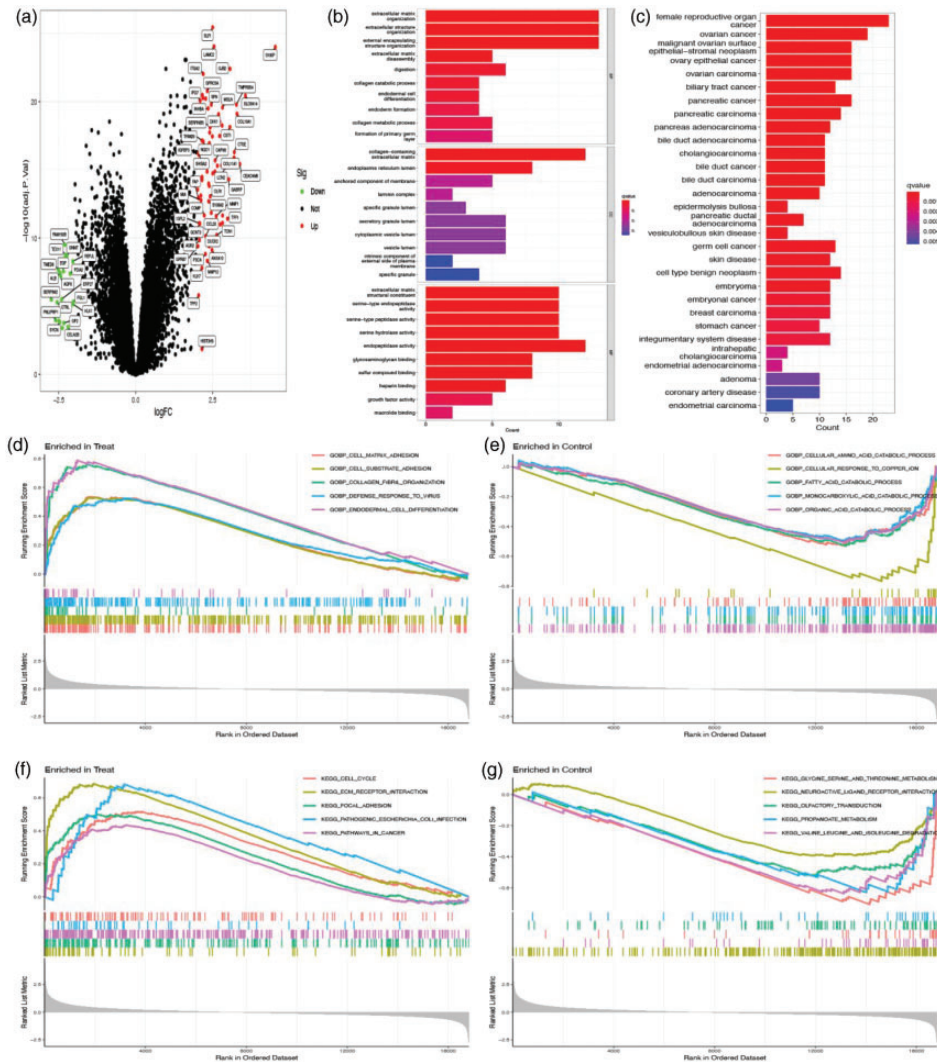


Figure 1. (a) Volcano plot of differentially expressed genes. Red represents up-regulated genes and green represents down-regulated genes. (b) Gene ontology enrichment analysis. (c) Disease ontology enrichment analysis. (d) Functional gene set enrichment analysis of the experimental group. (e) Functional GSEA enrichment analysis of the control group. (f) Pathway GSEA enrichment analysis of the experimental group and (g) Pathway GSEA enrichment analysis of the control group. GSEA, gene set enrichment analysis.

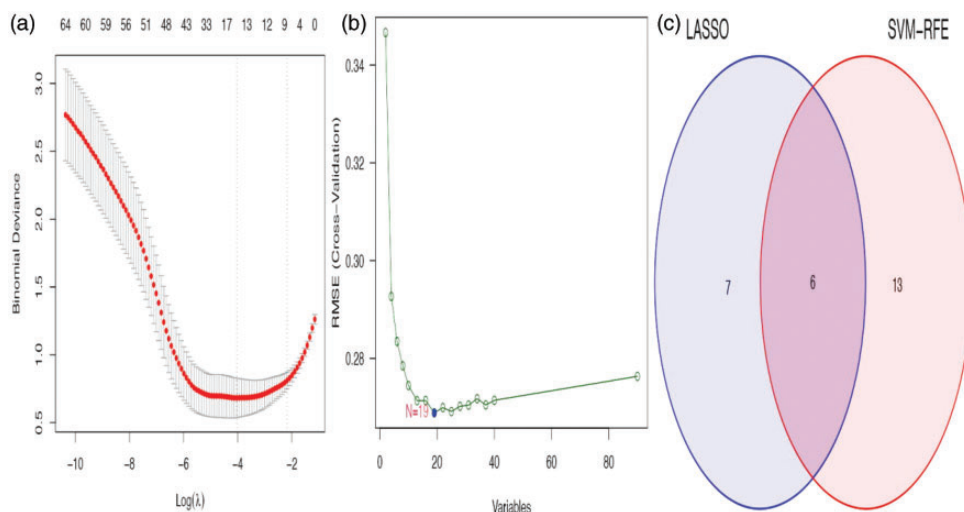


Figure 2. (a) LASSO regression screening for genes characteristic of pancreatic cancer. The point with the smallest cross validation error is 13, representing a total of 13 genes. (b) SVM-RFE screening for genes characteristic of pancreatic cancer. The point with the smallest cross-validation error is 19, representing a total of 19 genes and (c) Venn diagram of the intersection of the characteristic genes of pancreatic cancer screened by two machine learning methods. LASSO, least absolute shrinkage and selection operator; SVM-RFE, support vector machine and recursive feature elimination.

machine learning obtained six genes characteristic of pancreatic cancer (Figure 2c).

Validation of signature genes

In the training group, boxplot analysis showed that the expression of six genes characteristic of pancreatic cancer (*SLPI*, *S100P*, *IFI27*, *MSLN*, *DKK1*, and *SERPINB5*) was significantly different between experimental and control groups (Figure 3a). These six genes also showed significant differences in expression in the validation test group (Figure 3b). ROC curve analysis showed that the characteristic genes identified have a high level of accuracy as pancreatic cancer diagnostic genes, with values all exceeding 0.8 (Figure 3c).

Immune correlation analysis

The distribution of immune cells in experimental and control groups was visualized as a heat map (Figure 4a). Nine immune cell

types were identified as having significant differences in gene expression between control and experimental groups in immune cell difference analysis: activated CD4⁺ T cells, activated dendritic cells, CD56^{bright} natural killer cells, macrophages, natural killer T cells, natural killer cells, T helper type 2 cells, central memory CD8⁺ T cells, and effector memory CD8⁺ T cells (Figure 4b, $P < 0.001$). Immune cell correlation analysis identified a significant correlation between the expression of genes characteristic of pancreatic cancer and the following nine immune cell types: T helper type 2 cells, T helper type 17 cells, plasmacytoid dendritic cells, monocytes, eosinophils, effector memory CD8⁺ T cells, effector memory CD4⁺ T cells, central memory CD8⁺ T cells, and central memory CD4⁺ T cells (Figure 4c, $P < 0.001$).

Discussion

In this study, a total of 90 genes differentially expressed between pancreatic cancer

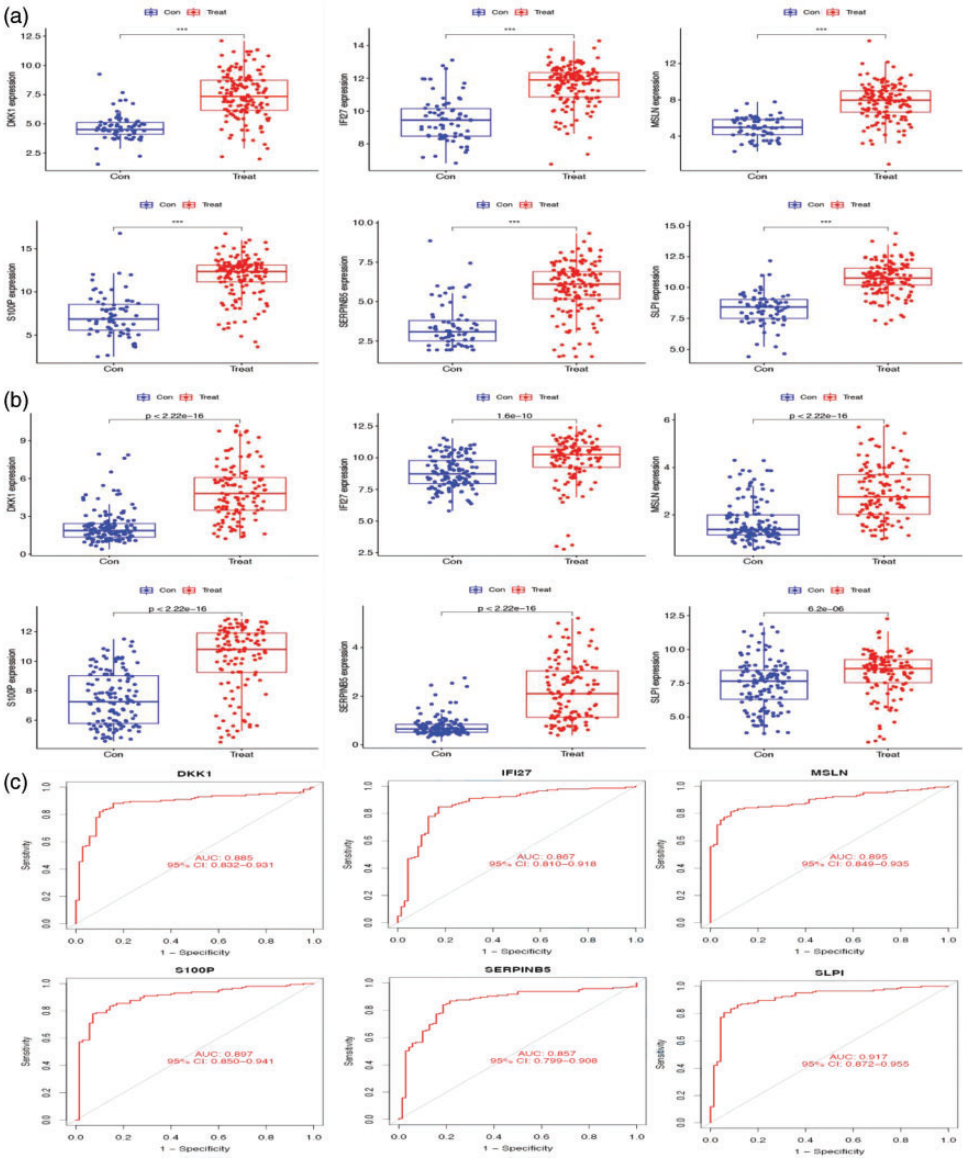


Figure 3. (a) Difference boxplot for the training group. (b) Difference boxplot for the test group and (c) ROC curve. ROC, receiver operating characteristic.

tissue and healthy tissue were screened from GSE15471, GSE16515, GSE32676, and GSE55643 datasets. GO enrichment analysis of these genes showed them to be mainly enriched in extracellular matrix organization, extracellular structure organization, and external encapsulating structure

organization. Extracellular matrix organization has previously been found to influence cellular phenotypes and biophysical properties. It also controls cell behavior and differentiation, and its dysregulation leads to the development and progression of diseases such as cancer.⁸⁻¹⁰

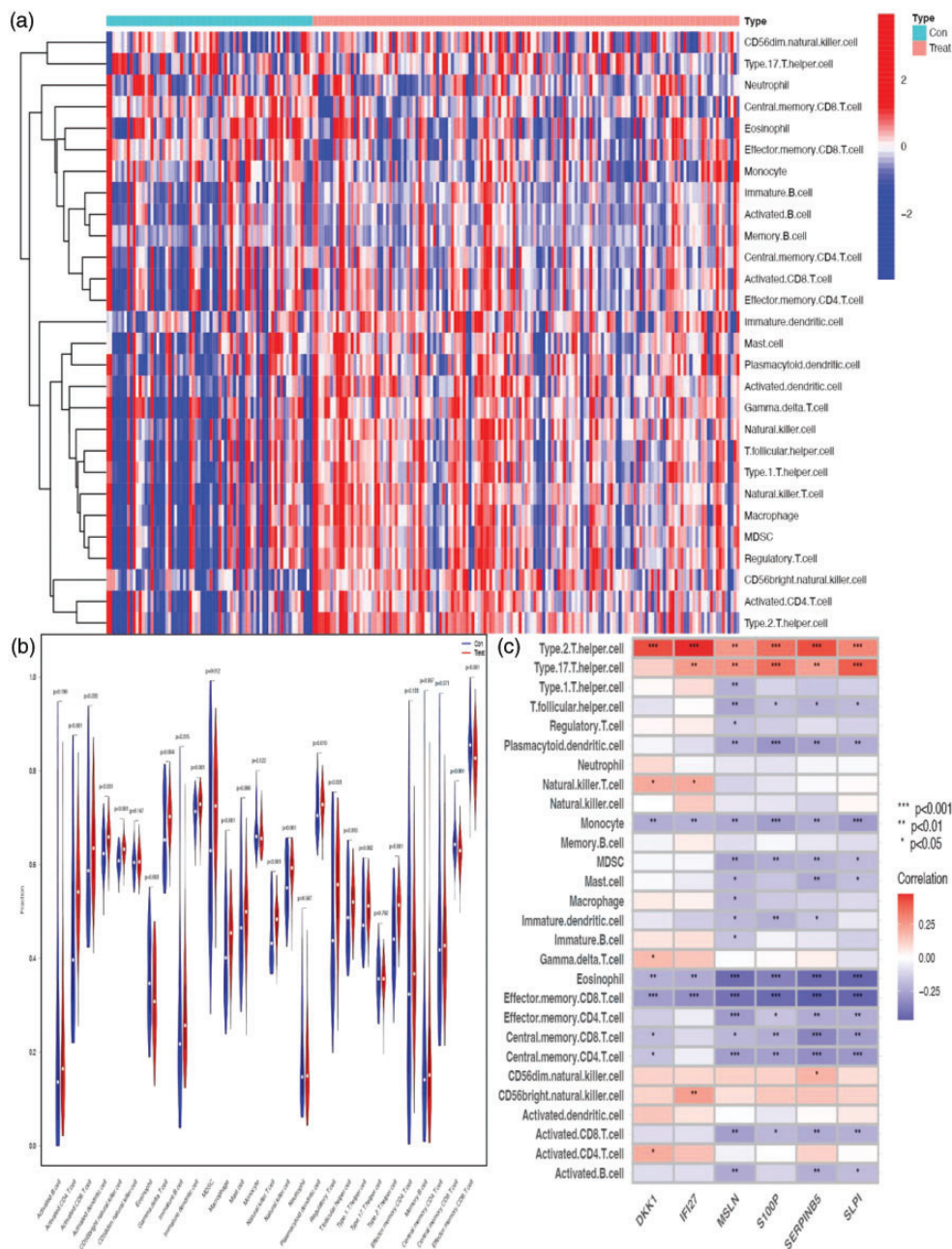


Figure 4. (a) Heatmap of immune cells in experimental and control groups. Red represents high expression and blue represents low expression. (b) Differential analysis of immune cells and (c) immune cell correlation analysis.

DO enrichment analysis showed that the differentially expressed genes were mainly enriched in tumors affecting female reproductive organs including ovarian cancer, and malignant ovarian surface epithelial-stromal neoplasm. This may provide new ideas and directions for future research into gynecological tumors. The identification of enriched functions and pathways through GSEA, such as cell matrix adhesion and endodermal cell differentiation, provides a direction for future *in vitro* studies to explore the mechanism pancreatic cancer of occurrence and development.

The LASSO regression model involves the smallest number of predictors and the smallest prediction error, so is a preferred method of screening for pancreatic cancer signature genes.^{11,12} Previous studies have also reported the use of SVM-REF screening for gene analysis of tumors.^{13,14} In the present study, genes characteristic of pancreatic cancer screened by LASSO regression and SVM-REF were intersected to obtain the six pancreatic cancer signature genes *SLPI*, *S100P*, *IFI27*, *MSLN*, *DKK1*, and *SERPINB5*. Zhang et al.¹⁵ previously showed that *SLPI* is a multifunctional protein involved in regulating immune responses, inhibiting protease activity, and blocking the transcription of pro-inflammatory genes through the nuclear factor-kappa B pathway. It also regulates the tumor microenvironment in pancreatic cancer cells, so is a possible diagnostic and prognostic biomarker with therapeutic potential in several cancers.

Li et al.¹⁶ reported that many proteins highly expressed in pancreatic ductal adenocarcinoma (PDAC) are associated with low tumor survival rates and exhibit tumor suppressive properties in the extracellular environment. For example, extracellular prostatic stem cell antigen (PSCA) was shown to function as a tumor suppressor, and significantly reduced the viability

and transwell invasion of PDAC cells. Its anti-PDAC effect was partially mediated by mesothelin, another tumor-associated antigen highly expressed in PDAC. This anti-PDAC potential suggests a dual role for tumor proteins such as PSCA in PDAC inhibition. A similar study to our current one, by Huang¹⁷ et al., identified four pancreatic cancer prognostic genes, including *DKK1*, showing significant differences in expression between tumor and healthy tissue. ROC curve analysis found the genes to be highly accurate at diagnosing pancreatic cancer, with ROC values exceeding 0.8.

Immune cells are not only an important part of the tumor stroma, but their mutual interference with cancer cells can lead to the occurrence, development, and metastasis of tumors.¹⁸ An in-depth analysis of tumor-infiltrating immune cells will hopefully reveal the mechanism of cancer immune evasion, thereby providing new directions for the study of novel tumor immunotherapy.^{19,20} For example, the immune cells identified in the present study in cell difference and correlation analyses will help focus future immunotherapy research on pancreatic cancer. Our study has some limitations. First is that we did not perform *in vitro* or *in vivo* analysis to determine the functions of signature genes in pancreatic cancer tissue progression. Second, although we explored the biological process of differentially expressed genes in pancreatic cancer through enrichment analysis, the link between them needs to be verified by follow-up experiments and clinical trials.

Author contributions

Longhui Zeng proposed the study concept, performed data mining and genetic analysis, and wrote the manuscript. Zheng Chen proposed the study concept and contributed to revision of the manuscript. Both authors contributed to the article and approved the submitted version.

Consent for publication

Not applicable.

Data availability statement

Datasets GSE15471, GSE16515, GSE32676, GSE55643, and GSE71729 were downloaded from the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo>).

Declarations

Ethics approval was not needed for this study because research analysis was based on bioinformatics and did not involve patient data.

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Funding statement

There was no financial support for this study.

ORCID iD

Longhui Zeng  <https://orcid.org/0009-0004-2101-3625>

References

1. Torphy RJ, Fujiwara Y, and Fujiwara Y. Pancreatic cancer treatment: better, but a long way to go. *Surg Today* 2020; 50: 1117–1125.
2. Hu JX, Zhao CF, Chen WB, et al. Pancreatic cancer: A review of epidemiology, trend, and risk factors. *World J Gastroenterol* 2021; 27: 4298–4321.
3. Garg SK and Chari ST. Early detection of pancreatic cancer. *Curr Opin Gastroenterol* 2020; 36: 456–461.
4. Ghafouri-Fard S, Fathi M, Zhai T, et al. LncRNAs: novel biomarkers for pancreatic cancer. *Biomolecules* 2021; 11: 1665.
5. Yang J, Xu R, Wang C, et al. Early screening and diagnosis strategies of pancreatic cancer: a comprehensive review. *Cancer Commun (Lond)* 2021; 41: 1257–1274.
6. Cai J, Chen H, Lu M, et al. Advances in the epidemiology of pancreatic cancer: Trends, risk factors, screening, and prognosis. *Cancer Lett* 2021; 520: 1–11.
7. Luu TT. Review of immunohistochemistry biomarkers in pancreatic cancer diagnosis. *Front Oncol* 2021; 11: 799025.
8. Walker C, Mojares E and Del Rio Hernandez A. Role of extracellular matrix in development and cancer progression. *Int J Mol Sci* 2018; 19: 3028.
9. Gilkes DM, Semenza GL and Wirtz D. Hypoxia and the extracellular matrix: drivers of tumour metastasis. *Nat Rev Cancer* 2014; 14: 430–439.
10. Pickup MW, Mouw JK and Weaver VM. The extracellular matrix modulates the hallmarks of cancer. *EMBO Rep* 2014; 15: 1243–1253.
11. Aheto JMK, Duah HO, Agbadi P, et al. A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, Ridge and Elastic net regression approaches compare? *Prev Med Rep* 2021; 23: 101475.
12. Pripp AH and Stanisic M. Association between biomarkers and clinical characteristics in chronic subdural hematoma patients assessed with lasso regression. *PLoS One* 2017; 12: e0186838.
13. Xia J, Sun L, Xu S, et al. A model using support vector machines recursive feature elimination (SVM-RFE) algorithm to classify whether COPD patients have been continuously managed according to GOLD guidelines. *Int J Chron Obstruct Pulmon Dis* 2020; 15: 2779–2786.
14. Zhu Z, Lei Z, Qian J, et al. The Ion channel-related gene signatures correlated with diagnosis, prognosis, and individualized treatment in patients with clear cell renal cell carcinoma. *Front Pharmacol* 2022; 13: 889142.
15. Zhang X, Liu SS, Ma J, et al. Secretory leukocyte protease inhibitor (SLPI) in cancer. pathophysiology: mechanisms of action and clinical implications. *Pathol Res Pract* 2023; 248: 154633.
16. Li K, Huo Q, Minami K, et al. Exploring the tumor-suppressing potential of PSCA in pancreatic ductal adenocarcinoma. *Cancers (Basel)* 2023; 15: 4917.

17. Huang W, Li J, Zhou S, et al. Discovering a four-gene prognostic model based on single-cell data and gene expression data of pancreatic adenocarcinoma. *Front Endocrinol (Lausanne)* 2022; 13: 883548.
18. Hinshaw DC and Shevde LA. The tumor microenvironment innately modulates cancer progression. *Cancer Res* 2019; 79: 4557–4566.
19. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007; 147: 573–577.
20. Zhang Y and Zhang Z. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. *Cell Mol Immunol* 2020; 17: 807–821.