Original Article
Running Title: *COL10A1* and *TMPRSS4* Genes as Potential Diagnostic Biomarkers in Pancreatic Ductal Adenocarcinoma
Received: May 14, 2024; Accepted: November 19, 2024

Feature Selection Algorithms Combined with Experimental Validation Reveal *COL10A1* and *TMPRSS4* Genes as Potential Diagnostic Biomarkers in Pancreatic Ductal Adenocarcinoma

Neda Shajari[*,**], PhD candidate, Amin Ramezani[**], PhD, Ahmad Tahmasebi[***], PhD, Mohammad Hossein Anbardar[****], MD, Zahra Roshanizadeh[*,**], PhD, Abbas Ghaderi[*,**♦], PhD

[*] *Department of Immunology, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran*
[**] *Shiraz Institute for Cancer Research, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran*
[***] *Institute of Biotechnology, Shiraz University, Shiraz, Iran*
[****] *Department of Pathology, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran*

♦Corresponding Author
Abbas Ghaderi, PhD
Shiraz Institute for Cancer Research,
School of Medicine, Shiraz University of Medical Sciences,
Shiraz, Iran
Tel: +98-71-32303687
Email: Ghaderia@sums.ac.ir

Abstract

Background: Pancreatic adenocarcinoma (PAAD) is often diagnosed at a late stage, preventing curative surgery. Early detection is crucial for improving patient outcomes. This study aims to discover potential biomarkers for identifying asymptomatic PAAD tumors.

Methods: In this case-control study, two gene expression datasets of PAAD and normal samples were collected from GEO and TCGA databases. Independent analyses of these datasets were conducted, leading to the identification of genes common to both datasets. Gene ontology and pathway enrichment analyses for the feature genes were conducted. Following our strict criteria, three feature genes for experimental validation were selected. The reliability of the selected feature genes was determined through quantitative real-time polymerase chain reaction (qRT-PCR). Data were analyzed using GraphPad Prism 8 software, employing the Mann-Whitney test and unpaired t-test. A *P*-value of <0.05 was considered statistically significant.

Results: A total number of 33 genes common to both GEO and TCGA datasets were identified. Gene ontology and pathway enrichment methods revealed that the selected genes were primarily associated with proteolysis and extracellular matrix organization. Based on our criteria, three feature genes (*COL10A1*, *CTHRC1*, and *TMPRSS4*) were selected for experimental validation. The results of qRT-PCR on independent patient samples demonstrated that the expression levels of *COL10A1* and *TMPRSS4* were significantly upregulated in PAAD tissues as compared with normal pancreatic tissues. In contrast, *CTHRC1* expression levels did not change significantly in PAAD in comparison with normal samples.

Conclusions: Our findings suggest that *COL10A1* and *TMPRSS4* can be attractive biomarkers for the mRNA-based diagnosis of PAAD.

Keywords: Pancreatic neoplasms, Gene expression profiling, Biomarkers, Diagnosis

Introduction

Pancreatic cancer (PC), a highly fatal malignancy with a poor prognosis, is highlighted by the close parallel between disease incidence (496,000) and mortality (466,000). Based on GLOBOCAN 2020, PC is ranked the 12th most common cancer and the 7th leading cause of cancer-related death worldwide. Despite medical oncology advances, PC has the lowest survival rate of all major organ cancers. The five-year survival rate for patients with PC is almost 10%.[1] Pancreatic tumors can take different forms, but they can generally be classified into two main categories: exocrine tumors, which include pancreatic adenocarcinoma (PAAD), also called pancreatic ductal adenocarcinoma (PDAC), and neuroendocrine tumors.[2]

PAAD occurs in the lining of the ducts in the pancreas. This malignancy is the most prevalent and aggressive form of PC, comprising more than 90% of all pancreatic tumors.[2,3] A diagnosis is typically made in the later stages of the disease when a dense, desmoplastic stroma has developed and the cancer has spread to other organs, making potentially curative surgery impossible. In the case of these patients, the disease tends to progress rapidly, and the majority do not survive beyond one year following their diagnosis. The prognosis is more promising for patients diagnosed earlier and eligible for potentially curative surgery.[4,5]

The carbohydrate antigen CA 19–9 is widely used for diagnosing PC, but its sensitivity and specificity range only around 80%. Therefore, identifying new disease-specific biomarkers to aid in the early detection and prognosis evaluation of patients with PC is necessary to increase patient survival rates.[6] The lack of reliable biomarker tests for PC necessitates the development of novel strategies to identify and characterize effective biomarkers. In recent years, the use of high-throughput sequencing technologies to investigate the genetic aspects of various diseases, including cancer, has become increasingly popular. These techniques monitor gene expression levels across the genome, and are especially useful for the identification of differentially expressed genes (DEGs) and fundamental mechanisms underlying cancer pathogenesis.[7] Machine learning techniques are also increasingly used to detect gene features from complex expression datasets. Several supervised and unsupervised learning algorithms have been widely applied to identify DEGs, prognostic-related genes, and therapeutic targets. These techniques provide an attractive approach to gaining new insights into biological processes. Specifically, feature selection methodologies are capable of integrating transcriptome studies, resulting in the identification of significant features and potential biomarkers.[8]

In this study, we applied feature selection algorithms on large-scale gene expression data from PAAD to identify potential diagnostic biomarkers. Furthermore, the reliability of these potential biomarkers was validated through quantitative real-time polymerase chain reaction (qRT-PCR).

Materials and Methods

Expression data collection and preprocessing

In this case-control study, two gene expression datasets of PAAD and normal samples were collected from GEO (microarray) and TCGA (RNA-seq). For the first dataset, the raw microarray experiments data were retrieved from Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) database. The "PC, PDAC, PAAD", and their combinations were used as keywords to search in the GEO database. The datasets were filtered by organism (Homo sapiens), and only studies that included normal participants and patients were selected. Finally, seven microarray data sets of PC comprising a total of 388 samples were collected. Raw expression data for Affymetrix datasets were preprocessed through quantile normalization and background correction with Robust Multi array Average (RMA) algorithm[9] in the Expression Console software (Affymetrix, Santa Clara, CA, USA). Moreover, the Agilent dataset was preprocessed based on quantile normalization and log2 transformation in the limma R package.[10] After preprocessing, probe IDs from different platforms were matched to their gene symbol. The probes that did not match the gene symbol were removed. We selected the probe with the greatest interquartile range (IQR) for the cases where multiple probe IDs were matched to the same gene symbol. The batch effects among the datasets were removed using empirical Bayes algorithm (ComBat) in the SVA package.[11]

The second dataset, consisting of RNA sequencing data, was obtained from The Cancer Genome Atlas (TCGA, www.portal.gdc.cancer.gov/) database using Bioconductor R package TCGAbiolinks.[12] The dataset contained 915 samples, including 178 PAAD samples and 737 normal samples of different tissues. The normalized expression data were retrieved based on the fragments per kilobase of transcript per million mapped reads (UQ-FPKMs) values. Genes with zero value in 50% or more of the samples were excluded, and the dataset was then transferred into the log2 scale.

Identification of the important genes

We used several feature selection methods as part of machine learning approaches to identify the significant feature genes. We employed nine distinct attribute weighting algorithms, namely support vector machine (SVM), chi square, information gain, information gain ratio, deviation, Gini index, Uncertainty, Relief, and principal component analysis (PCA) for the two data sets separately. All the algorithms were implemented in RapidMiner software.[13] Finally, genes selected by at least one of the algorithms were considered feature genes.

Pathway and functional analyses

The functional enrichment and pathway enrichment analysis feature genes were conducted using the g:Profiler tool.[14] The adjusted *P*-value threshold of <0.05 was deemed significant for biological process, molecular function, and cellular component terms. The immune-related genes were also acquired from the Immunology Database and Analysis Portal (ImmPort) InnateDB.[15]

Selection of candidate genes

Three genes were prioritized from the list of feature genes for further analysis and qRT-PCR based on the following criteria: (i) a higher expression level in pancreatic tumor tissues than in non-cancerous tissues, (ii) located on the cell surface or in the extracellular region, and (iii) involved in the creation of dense desmoplastic stroma.

Subcellular locations and external validation of the expression of selected feature genes

We first used the GeneCards (https://www.genecards.org/) and Human Protein Atlas (https://www.proteinatlas.org/) databases to explore the subcellular location of genes. The Gene Expression Profiling Interactive Analysis (GEPIA) tool[16] was used to conduct external validation of gene expression patterns in both healthy and malignant tissues. Furthermore, the GEPIA platform was employed to examine the correlation between expression levels of selected genes and pathological stages in PAAD tissues.

FFPE tissue samples preparation

The formalin-fixed, paraffin-embedded (FFPE) tissue sections were collected between 2018 and 2020 in Abu Ali Sina Hospital, Shiraz, Iran. The Ethics Committee of Shiraz University of Medical Sciences granted ethical approval for this study (Approval ID: IR.SUMS.REC.1400.193). Informed consent was obtained from all patients or their legal guardian(s). The inclusion criteria for the study encompassed patients diagnosed with PAAD based on clinical and pathological findings who had not received any therapeutic interventions, such as chemotherapy or radiotherapy, before surgery. The exclusion criteria were: receiving prior therapies such as chemotherapy and radiotherapy prior to surgery, bacterial and viral infection, diabetes, cardiovascular diseases, autoimmune diseases, or other cancers. The clinicopathological characteristics of the patients are summarized in table 1. In accordance with our experimental protocol, we excised five sections from each sample, each measuring ten μm in thickness. These sections were subsequently deposited onto a 2 ml sterile microcentrifuge tube. A fresh sterile microtome blade was employed for each paraffin block to prevent any potential cross-contamination among the samples. To detect the border around the target tumor cell areas, tissue sections from each sample were stained with hematoxylin and eosin (H&E). A proficient pathologist evaluated tumor cells under a microscope.

RNA extraction and cDNA synthesis

The RNeasy FFPE kit (Qiagen) was used to extract total RNA from FFPE tissue sections following the manufacturer's instructions. DNase I treatment was performed to eliminate potential genomic DNA contamination of the samples. The RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific) was used along with oligo-dT and random hexamers as primers to perform cDNA synthesis.

qRT-PCR

The TB Green Premix Ex Taq II (TAKARA) and specific primers were used to run qRT-PCR on an ABI StepOne system (Applied Biosystems). The *β-actin* gene was used as an internal control gene for data normalization. AlleleID software (version 7.5) was used to design primer sets. Before the experiment, all primer sequences were blasted using the primer-blast tool available on the NCBI website (http://www. nchi.nlm.nih.gov). Table 2 presents the primer sequences. Amplification was carried out in 48-well microtitre plates under the following conditions: initial denaturation at 95°C for 30 seconds, followed by 40 cycles of denaturation at 95°C for 5 seconds and annealing at the specific primer annealing temperature for 30 seconds. All amplifications have been performed twice to ensure repeatability. Amplification efficiencies were calculated using reference curves with serial dilutions of PCR product, which were subsequently used for data normalization. Melt curve analysis was used to verify the specificity of the qRT-PCR. Data normalization was performed using the CtNorm algorithm available at http://www.ctnorm.sums.ac.ir.[17]

Statistical analysis

The results of the study were reported as the mean ± standard deviation (SD). Data distribution was assessed using the D'Agostino-Pearson test. The Mann-Whitney test was used to compare *COL10A1* and *TMPRSS4* expression levels, while the unpaired t-test was used to compare *CTHRC1* gene expression levels. Statistical analysis was conducted using GraphPad Prism version 8.0, which GraphPad Software developed. A *P*-value less than 0.05 was considered statistically significant.

Results

Identification of feature genes in PAAD

We implemented feature selection methods on two datasets to determine and prioritize the feature genes specific to PAAD. The first data set includes seven independent studies that were retrieved from the GEO database (Table 3). After data pre-processing and the batch effect correction, we obtained an expression dataset, including 388 samples that divided to tumor and normal groups. The results showed that 230 genes were selected by at least one of the algorithms to discriminate cancer from normal samples (Table 4). In order to distinguish target genes with greater specificity to cancerous tissues, we procured a secondary data set from the TCGA database. This data set consisted of 178 samples of PAAD and 737 normal samples obtained from different tissues. Our analytical focus in this data set centered on genes that exhibited elevated levels of expression in cancerous tissue compared with normal tissues. These genes could potentially be used as diagnostic or therapeutic targets for PAAD patients. Overall, 250 genes were screened by at least one of the algorithms to classify tumor and normal groups (Table 5). The Venn diagram indicated that 33 genes overlapped between these independent analyses (Figure 1 and Table 6).

Functional and pathway analyses

The Gene ontology (GO) enrichment analysis for common genes indicated that biological terms, including proteolysis and extracellular matrix organization, were highly enriched. Moreover, the result of molecular function terms revealed that many genes were annotated with serine-type endopeptidase activity and peptidase activity. The pathway analysis of the common genes also revealed that Protein digestion and absorption and Pancreatic secretion pathways were significantly enriched (Figure 2 and Table 7). Among the common genes, two genes, including *OLFM4* and *GP2* were overlapped with a list of immune-related genes derived from InnateDB. Then, we investigated the common genes identified via feature selection techniques, with emphasis on the evaluation criteria of subcellular location and high expression levels in pancreatic tumor tissues. Three genes, *COL10A1*, *CTHRC1*, and *TMPRSS4*, were prioritized from the list of feature genes for further analysis.

Subcellular locations

We explored the subcellular location of selected genes using the GeneCards and Human Protein Atlas databases. The findings indicated that COL10A1 and CTHRC1 proteins were predominantly found in the extracellular space, whereas TMPRSS4 was mainly expressed in the extracellular space and plasma membrane.

The mRNA expression level of selected feature genes in PAAD

To compare the mRNA expression of selected genes between PAAD and normal samples, we used the GEPIA dataset. The results demonstrated that the expression levels of *COL10A1*, *CTHRC1*, and *TMPRSS4* were higher in tumors than in normal samples (Figure 3). Further analysis also indicated significant correlations between *COL10A1*, *CTHRC1*, and *TMPRSS4* expression levels and the pathological stages of PAAD (Figure 4).

Validation of expression of selected genes in a separate patient cohort

We conducted qRT-PCR using independent patient samples to confirm the findings mentioned earlier. We collected 23 PAAD samples and 19 adjacent normal tissue samples to validate the expression of the selected genes in patients with PAAD. The results of our qRT-PCR analysis demonstrated that *COL10A1* and *TMPRSS4* mRNA exhibited significantly higher expression levels in PAAD tissues compared with normal pancreatic tissues (*P* = 0.0002 and *P* = 0.0019, respectively). Despite the upregulated expression in PAAD samples, no statistically significant difference was observed in the expression of *CTHRC1* when compared with that in normal samples (fold change = 1.27) (Figure 5). Furthermore, in the receiver operating characteristic (ROC) analysis, *COL10A1* and *TMPRSS4* had an area under the ROC Curve (AUC) values of 94.44% and 86.31%, respectively. This suggests a relatively better overall diagnostic accuracy for them compared with *CTHRC1*, with an AUC of 64.81%.

Discussion

We used feature selection algorithms combined with experimental validation to suggest potential diagnostic biomarkers for PAAD. By implementing feature selection techniques, we identified 33 genes. The further analysis prioritized *COL10A1*, *CTHRC1*, and *TMPRSS4* as potential diagnostic biomarkers. Our qRT-PCR results showed that in PAAD samples, the expression levels of *COL10A1* and *TMPRSS4* were higher than in normal pancreatic tissues. In contrast, there was no significant change in *CTHRC1* expression levels compared with normal samples. Our results suggest that *COL10A1* and *TMPRSS4* could open up opportunities for diagnosing PC in an earlier stage.

Improving the diagnostic accuracy and clinical outcomes of patients with PC is pivotal. This is because a delay in diagnosis, systemic spreading at the time of diagnosis, and inadequate effective treatment can negatively impact the patient's prognosis. Therefore, developing diagnostic and prognostic biomarkers and therapeutic targets is necessary. To this end, we applied nine feature selection models on gene expression datasets of PAAD and normal samples, and 33 feature genes were identified overall. Several studies have used gene expression microarray technology to identify tumor-associated genes. However, the results were inconsistent, possibly caused by dissimilarity in dataset selection and statistical procedures applied.[18] Because microarray gene expression data are highly dimensional, feature selection techniques, which form a subset of machine learning, can offer valuable assistance in differentiating genes possessing specific biological capabilities. Our GO and pathway investigation demonstrated that the feature genes were predominantly enriched with proteolysis and extracellular matrix organization. The result of molecular function terms revealed that many genes were annotated with serine-type endopeptidase and peptidase activity. The pathway analysis of the common genes revealed that protein digestion and absorption and pancreatic secretion pathways were significantly enriched. These pathways predominantly influence biological processes of energy metabolism and substance absorption. It is necessary to mention that most PCs originate from exocrine cells, which produce digestive juices, and abnormal metabolism is a key hallmark of PCs. We investigated the common genes identified via feature selection techniques, with a focus on the evaluation criteria of subcellular location, function, and high expression levels in pancreatic tumor tissues. Of the list of feature genes, *COL10A1*, *CTHRC1*, and *TMPRSS4* were the three genes selected for further analysis. In the next step, increased expression of three selected feature genes in PAAD tissues was confirmed by the GEPIA online tool. Furthermore, the correlation between selected genes and pathological stages in PAAD patients demonstrated that the expression of *COL10A1*, *CTHRC1*, and *TMPRSS4* are negatively associated with disease progression, supporting their prognostic significance. Using the qRT-PCR method, we also verified our bioinformatics findings in an independent patient cohort. Our results showed that in comparison with normal pancreatic tissues, *COL10A1*, *TMPRSS4*, and *CTHRC1* were highly expressed in PAAD samples. *CTHRC1* expression did not increase statistically significantly, which may be due to the limited sample size.

Collagen, a key element of the tumor microenvironment, serves as a scaffold for cell growth and induces epithelial cell proliferation, differentiation, and migration. A collagen-rich fibrotic extracellular matrix is an important hallmark of PAAD. Increased fibrosis due to collagen deposition can promote tumor development and invasion.[19] As a member of the collagen family, type X collagen alpha 1 chain (*COL10A1*) has been reported to be a tumor-associated gene. The expression of this gene is negligible in most normal adult tissues and is elevated in various solid tumor tissues. As a result of integrated bioinformatics analysis, Li et al. identified *COL10A1* as a potential prognostic marker in esophageal squamous cell carcinoma.[20] In colorectal cancer, Huang et al. found that overexpression of COL10A1 protein levels inhibits proliferation, suppresses EMT, and reduces the migration and invasion ability of colorectal cancer cells. Additionally, they reported that *COL10A1* overexpression could independently predict prognosis and overall survival (OS) in colorectal cancer patients.[21] Liang et al. provided clues for the contribution of *COL10A1* in the proliferation and metastasis of lung adenocarcinoma (LUAD) cells through the COL10A1/DDR2/FAK axis. The study also revealed that patients diagnosed with LUAD who had elevated *COL10A1* expression exhibited unfavorable OS and RFS outcomes.[22] Furthermore, a study conducted by Andriani et al. confirmed that the levels of the circulating protein COL10A1 are much higher in the plasma of individuals with lung cancer compared with lung cancer cells. This suggests that COL10A1 could be a promising candidate for diagnosing the disease.[23] Epigenetic modifications are inheritable and reversible mechanisms that alter gene expression and chromatin structure without modifying the DNA sequence. They are generally recognized to impact all aspects of tumor progression.[24] A recent study has shown that the presence of m6A (N6-methyladenosine) modification, facilitated by enhanced METTL3 (methyltransferase like 3) in CAFs (cancer-associated fibroblasts), leads to an increase in the expression of *COL10A1*. This, in turn, promotes cell proliferation and inhibits apoptosis, ultimately speeding tumor growth in a lung cancer model, both in vitro and in vivo.[25] In line with our findings, Wen et al. observed significant upregulation of the *COL10A1* gene in both PC cells and tissues. This increased expression of *COL10A1* was associated with an unfavorable prognosis. Moreover, the researchers uncovered that the COL10A1-DDR2 axis prompts the activation of the MEK/ERK signaling pathway which, in turn, induces epithelial-mesenchymal transition (EMT) and facilitates the progress of PC.[19] Also, using bioinformatics analysis, Xu et al. recognized *COL10A1* as a gene involved in PAAD tumorigenesis. They discovered that the expression of *COL10A1* was abnormally elevated in PAAD and linked with poor prognosis. Additionally, it was revealed that *COL10A1* knockdown reduced PC cell proliferation, migration, and invasion in vitro. They identified CD276 as a downstream target of *COL10A1* and demonstrated that *COL10A1* promotes tumorigenesis in PAAD by regulating CD276.[26] CD276, also known as B7-H3, is an inhibitory member of the B7 family that is highly expressed in tumors and participates in tumor cell immunosuppression by inhibiting NK cell and T-cell activity.[27]

Collagen triple helix repeat containing-1 (*CTHRC1*) plays a pivotal role in several physiological and pathological processes. Several studies have suggested that overexpression of *CTHRC1* promotes tumor growth, invasion, and metastasis through various signaling pathways.[28,29] Our findings are in line with prior studies demonstrating that high levels of *CTHRC1* are associated with the progression and metastasis of PC. Liu et al. discovered that *CTHRC1* expression was markedly elevated in PDAC tissues compared with normal tissues, as observed at both the mRNA and protein levels. According to their findings, increased *CTHRC1* expression is a negative indicator of prognosis in PDAC. Patients with higher *CTHRC1* expression had significantly shorter OS. Elevated levels of CTHRC1 protein expression were strongly linked to the occurrence of invasion and metastasis.[30] Furthermore, *CTHRC1* can regulate immune cells to mediate PC development and progression. Lee et al. found that *CTHRC1* regulates the expression of angiopoietin-2 (Ang-2), a ligand for the Tie2 receptor. Moreover, they have demonstrated that *CTHRC1* facilitates the process of angiogenesis in pancreatic tumors by recruiting monocytes that express Tie2 to the tumor microenvironment.[31]

The involvement of proteases in almost all biological processes highlights their immense importance in both healthy and pathological conditions. Notably, dysregulation of proteases is a pivotal event in cancer development. There has been evidence that proteases play an important role in various stages of cancer. Their impact is twofold, directly exerting influence through their proteolytic activity, as well as indirectly by regulating cellular functions and signaling. Transmembrane serine protease 4 (*TMPRSS4*), a protease belonging to the Type II transmembrane serine protease (TTSP) family, is expressed at high levels in various cancer types and directly correlates with poor outcomes.[32] In lung cancer, *TMPRSS4* has been shown to promote tumor growth and affect chemotherapy treatment by imparting drug resistance.[33] In gastric cancer, *TMPRSS4* enhanced the invasion of malignant cells by activating the NF-kB/MMP-9 signaling pathway.[34] Wang et al. showed a significant association between *TMPRSS4* expression and hepatocellular carcinoma (HCC) progression. They found that *TMPRSS4* acts as a positive regulator of the Raf/MEK/ERK1/2 pathway, leading to EMT and promoting invasion, migration, and metastasis of HCC. Additionally, they demonstrated that *TMPRSS4* induces angiogenesis by suppressing the expression of RECK.[35] The results of Bhasin et al. agree with our findings that TMPRSS4 is an oncogenic protein overexpressed in PAAD tissues. Using a meta-analysis approach based on PAAD datasets, they identified and validated a five-gene classifier signature. This signature, which included *TMPRSS4*, *AHNAK2*, *POSTN*, *ECT2*, and *SERPINB5*, exhibited 95% sensitivity and 89% specificity in distinguishing PAAD from non-tumor samples.[36] Studies conducted by Gu et al. in PAAD have shown that *TMPRSS4* is significantly overexpressed in tumoral tissues compared with non-tumor tissues, which has been linked to poor prognosis. They revealed that *TMPRSS4* has a proto-oncogene function in PC by promoting proliferation, inhibiting apoptosis, and increasing cell invasion. TMPRSS4 accomplishes its oncogenic functions by stimulating the ERK1/2 signaling pathway in PC cells.[37]

Overall, our results suggest that *COL10A1* and *TMPRSS4* could open up opportunities for diagnosing PC in an earlier stage. However, further validation testing is needed to move these biomarkers from research to clinical application. To do so, repeating this study with larger-scale studies with diverse patient populations, comprehensive evaluation of these markers' sensitivity and specificity in distinguishing PAAD from normal pancreatic tissue and other pancreatic conditions, and finding a proper cut-off point for distinguishing between normal and pathological levels of these markers is essential.

One notable limitation of this study was the small sample size, which can affect the reliability and generalizability of the results and make it difficult to detect significant differences. Further studies with larger sample sizes are required to validate our findings. Another issue was the limitations in accessing comprehensive patient data. Metastasis is an important factor in PC staging. In this study, we had access only to the patient's pathology reports, and we did not have access to patient history and other necessary reports, such as treatments, whole-body scans, or PET scan reports, which are essential for determining metastasis status. Due to these limitations, a definite stage evaluation was impossible. Therefore, we could not demonstrate a correlation between disease stage and gene expression levels in our independent patient cohort. Collaborating with clinical partners for more detailed patient records may help overcome this obstacle in future research. Another limitation of this study was the lack of in vitro examination of the molecular mechanisms and associated pathways in which selected genes were involved. Understanding the molecular mechanisms and associated pathways could provide deeper insights into the diagnostic, prognostic, and therapeutic potential of these genes in PAAD. Our future research will involve conducting experiments to confirm these findings using other laboratory approaches.

Conclusion

In this study, we applied feature selection algorithms on large-scale gene expression data of PAAD to identify potential diagnostic biomarkers. Furthermore, the reliability of these potential biomarkers was validated through the use of qRT-PCR. Overall, the findings suggest that *COL10A1* and *TMPRSS4* could open up opportunities for diagnosing PC in an earlier stage.

Acknowledgments

This paper is part of Neda Shajari's PhD thesis and received financial support from Shiraz University of Medical Sciences, Shiraz, Iran (Grant No. 22288) and the Shiraz Institute for Cancer Research (ICR-100-109). We sincerely appreciate the help we received from Shiraz Institute for Cancer Research (ICR) and Abu Ali Sina Hospital staff.

Funding

The project received financial support from Shiraz University of Medical Sciences, Shiraz, Iran (Grant No. 22288) and Shiraz Institute for Cancer Research (ICR-100-109).

Authors’ Contribution

A.Gh: Study design, data interpretation, reviewing the manuscript; A.R: Study design, data interpretation, reviewing the manuscript; N.Sh: Data analysis and interpretation, drafting and critical reviewing of the manuscript; A.T: Data analysis and interpretation, reviewing the manuscript; M.H.A: Data collection, drafting; Z.R: Data collection, drafting. All authors have read and approved the final manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflict of Interest

None declared.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-49. doi: 10.3322/caac.21660. PMID: 33538338.

2. Wood LD, Canto MI, Jaffee EM, Simeone DM. Pancreatic cancer: Pathogenesis, screening, diagnosis, and treatment. *Gastroenterology*. 2022;163(2):386-402.e1. doi: 10.1053/j.gastro.2022.03.056. PMID: 35398344; PMCID: PMC9516440.

3. Klein AP. Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. *Nat Rev Gastroenterol Hepatol*. 2021;18(7):493-502. doi: 10.1038/s41575-021-00457-x. PMID: 34002083; PMCID: PMC9265847.

4. Amaral MJ, Oliveira RC, Donato P, Tralhão JG. Pancreatic cancer biomarkers: Oncogenic mutations, tissue and liquid biopsies, and radiomics-A review. *Dig Dis Sci*. 2023;68(7):2811-23. doi: 10.1007/s10620-023-07904-6. PMID: 36988759; PMCID: PMC10293428.

5. Karandish F, Mallik S. Biomarkers and targeted therapy in pancreatic cancer. *Biomark Cancer*. 2016;8(Suppl 1):27-35. doi: 10.4137/BiC.s34414. PMID: 27147897; PMCID: PMC4847554.

6. Zhang Y, Jiang L, Song L. Meta-analysis of diagnostic value of serum Carbohydrate antigen 199 in pancreatic cancer. *Minerva Med*. 2016;107(1):62-9. PMID: 26824636.

7. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res*. 2019;47(D1):D711-d5. doi: 10.1093/nar/gky964. PMID: 30357387; PMCID: PMC6323929.

8. van IJzendoorn DGP, Szuhai K, Briaire-de Bruijn IH, Kostine M, Kuijjer ML, Bovée JVMG. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS Comput Biol*. 2019;15(2):e1006826. doi: 10.1371/journal.pcbi.1006826. PMID: 30785874; PMCID: PMC6398862.

9. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249-64. doi: 10.1093/biostatistics/4.2.249. PMID: 12925520.

10. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi: 10.1093/nar/gkv007. PMID: 25605792; PMCID: PMC4402510.

11. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882-3. doi: 10.1093/bioinformatics/bts034. PMID: 22257669; PMCID: PMC3307112.

12. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44(8):e71. doi: 10.1093/nar/gkv1507. PMID: 26704973; PMCID: PMC4856967.

13. Tahmasebi A, Niazi A, Akrami S. Integration of meta-analysis, machine learning and systems biology approach for investigating the transcriptomic response to drought stress in Populus species. *Sci Rep*. 2023;13(1):847. doi: 10.1038/s41598-023-27746-6. PMID: 36646724; PMCID: PMC9842770.

14. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;47(W1):W191-w8. doi: 10.1093/nar/gkz369. PMID: 31066453; PMCID: PMC6602461.

15. Charitou T, Bryan K, Lynn DJ. Using biological networks to integrate, visualize and analyze genomics data. *Genet Sel Evol*. 2016;48:27. doi: 10.1186/s12711-016-0205-1. PMID: 27036106; PMCID: PMC4818439.

16. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*. 2017;45(W1):W98-w102. doi: 10.1093/nar/gkx247. PMID: 28407145; PMCID: PMC5570223.

17. Ramezani A. CtNorm: Real time PCR cycle of threshold (Ct) normalization algorithm. *J Microbiol Methods*. 2021;187:106267. doi: 10.1016/j.mimet.2021.106267. PMID: 34116107.

18. Islam S, Kitagawa T, Baron B, Abiko Y, Chiba I, Kuramitsu Y. ITGA2, LAMB3, and LAMC2 may be the potential therapeutic targets in pancreatic ductal adenocarcinoma: an integrated bioinformatics analysis. *Sci Rep*. 2021;11(1):10563. doi: 10.1038/s41598-021-90077-x. PMID: 34007003; PMCID: PMC8131351.

19. Wen Z, Sun J, Luo J, Fu Y, Qiu Y, Li Y, et al. COL10A1-DDR2 axis promotes the progression of pancreatic cancer by regulating MEK/ERK signal transduction. *Front Oncol*. 2022;12:1049345. doi: 10.3389/fonc.2022.1049345. PMID: 36530986; PMCID: PMC9750160.

20. Li J, Wang X, Zheng K, Liu Y, Li J, Wang S, et al. The clinical significance of collagen family gene expression in esophageal squamous cell carcinoma. *PeerJ*. 2019;7:e7705. doi: 10.7717/peerj.7705. PMID: 31598423; PMCID: PMC6779144.

21. Huang H, Li T, Ye G, Zhao L, Zhang Z, Mo D, et al. High expression of COL10A1 is associated with poor prognosis in colorectal cancer. *Onco Targets Ther*. 2018;11:1571-81. doi: 10.2147/ott.S160196. PMID: 29593423; PMCID: PMC5865565.

22. Liang Y, Xia W, Zhang T, Chen B, Wang H, Song X, et al. Upregulated collagen COL10A1 remodels the extracellular matrix and promotes malignant progression in lung adenocarcinoma. *Front Oncol*. 2020;10:573534. doi: 10.3389/fonc.2020.573534. PMID: 33324550; PMCID: PMC7726267.

23. Andriani F, Landoni E, Mensah M, Facchinetti F, Miceli R, Tagliabue E, et al. Diagnostic role of circulating extracellular matrix-related proteins in non-small cell lung cancer. *BMC Cancer*. 2018;18(1):899. doi: 10.1186/s12885-018-4772-0. PMID: 30227835; PMCID: PMC6145327.

24. Xie Z, Zhou Z, Yang S, Zhang S, Shao B. Epigenetic regulation and therapeutic targets in the tumor microenvironment. *Mol Biomed*. 2023;4(1):17. doi: 10.1186/s43556-023-00126-2. PMID: 37273004; PMCID: PMC10241773.

25. Li Y, Li X, Deng M, Ye C, Peng Y, Lu Y. Cancer-associated fibroblasts hinder lung squamous cell carcinoma oxidative stress-induced apoptosis via METTL3 mediated m(6)A methylation of COL10A1. *Oxid Med Cell Longev*. 2022;2022:4320809. doi: 10.1155/2022/4320809. PMID: 36246404; PMCID: PMC9560815.

26. Xu Q, Zheng J, Su Z, Chen B, Gu S. COL10A1 promotes tumorigenesis by modulating CD276 in pancreatic adenocarcinoma. *BMC Gastroenterol*. 2023;23(1):397. doi: 10.1186/s12876-023-03045-2. PMID: 37974070; PMCID: PMC10652574.

27. Azuma M. Co-signal molecules in T-cell activation : Historical overview and perspective. *Adv Exp Med Biol*. 2019;1189:3-23. doi: 10.1007/978-981-32-9717-3_1. PMID: 31758529.

28. Mei D, Zhu Y, Zhang L, Wei W. The role of CTHRC1 in regulation of multiple signaling and tumor progression and metastasis. *Mediators Inflamm*. 2020;2020:9578701. doi: 10.1155/2020/9578701. PMID: 32848510; PMCID: PMC7441421.

29. Jiang N, Cui Y, Liu J, Zhu X, Wu H, Yang Z, et al. Multidimensional roles of collagen triple helix repeat containing 1 (CTHRC1) in malignant cancers. *J Cancer*. 2016;7(15):2213-20. doi: 10.7150/jca.16539. PMID: 27994657; PMCID: PMC5166530.

30. Liu W, Fu XL, Yang JY, Yang MW, Tao LY, Liu DJ, et al. Elevated expression of CTHRC1 predicts unfavorable prognosis in patients with pancreatic ductal adenocarcinoma. *Am J Cancer Res*. 2016;6(8):1820-7. PMID: 27648368; PMCID: PMC5004082.

31. Lee J, Song J, Kwon ES, Jo S, Kang MK, Kim YJ, et al. CTHRC1 promotes angiogenesis by recruiting Tie2-expressing monocytes to pancreatic tumors. *Exp Mol Med*. 2016;48(9):e261. doi: 10.1038/emm.2016.87. PMID: 27686285; PMCID: PMC5050301.

32. Kim S. TMPRSS4, a type II transmembrane serine protease, as a potential therapeutic target in cancer. *Exp Mol Med*. 2023;55(4):716-24. doi: 10.1038/s12276-023-00975-5. PMID: 37009799; PMCID: PMC10167312.

33. Exposito F, Villalba M, Redrado M, de Aberasturi AL, Cirauqui C, Redin E, et al. Targeting of TMPRSS4 sensitizes lung cancer cells to chemotherapy by impairing the proliferation machinery. *Cancer Lett*. 2019;453:21-33. doi: 10.1016/j.canlet.2019.03.013. PMID: 30905815.

34. Jin J, Shen X, Chen L, Bao LW, Zhu LM. TMPRSS4 promotes invasiveness of human gastric cancer cells through activation of NF-κB/MMP-9 signaling. *Biomed Pharmacother*. 2016;77:30-6. doi: 10.1016/j.biopha.2015.11.002. PMID: 26796262.

35. Wang CH, Guo ZY, Chen ZT, Zhi XT, Li DK, Dong ZR, et al. TMPRSS4 facilitates epithelial-mesenchymal transition of hepatocellular carcinoma and is a predictive marker for poor prognosis of patients after curative resection. *Sci Rep*. 2015;5:12366. doi: 10.1038/srep12366. PMID: 26190376; PMCID: PMC4507176.

36. Bhasin MK, Ndebele K, Bucur O, Yee EU, Otu HH, Plati J, et al. Meta-analysis of transcriptome data identifies a novel 5-gene pancreatic adenocarcinoma classifier. *Oncotarget*. 2016;7(17):23263-81. doi: 10.18632/oncotarget.8139. PMID: 26993610; PMCID: PMC5029625.

37. Gu J, Huang W, Zhang J, Wang X, Tao T, Yang L, et al. TMPRSS4 Promotes Cell Proliferation and Inhibits Apoptosis in Pancreatic Ductal Adenocarcinoma by Activating ERK1/2 Signaling Pathway. *Front Oncol*. 2021;11:628353. doi: 10.3389/fonc.2021.628353. PMID: 33816264; PMCID: PMC8012900.

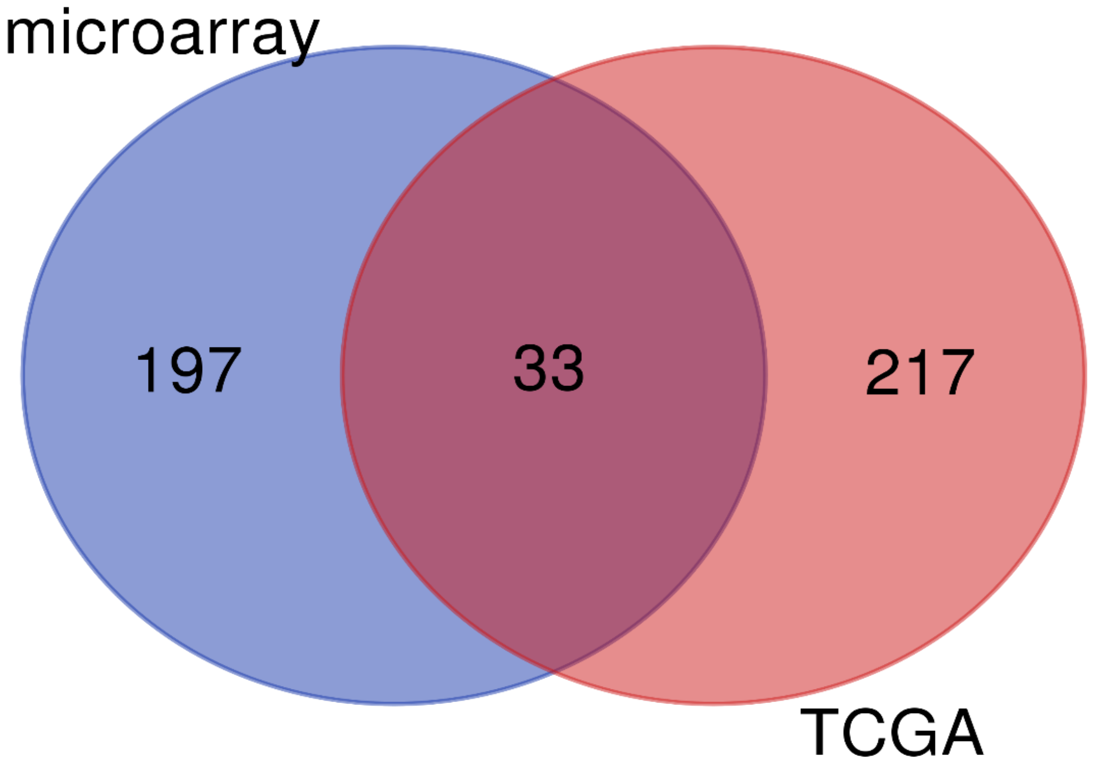

Figure 1. This figure shows the Venn diagram of 33 genes overlapped between two independent analyses.

TCGA: The cancer genome atlas

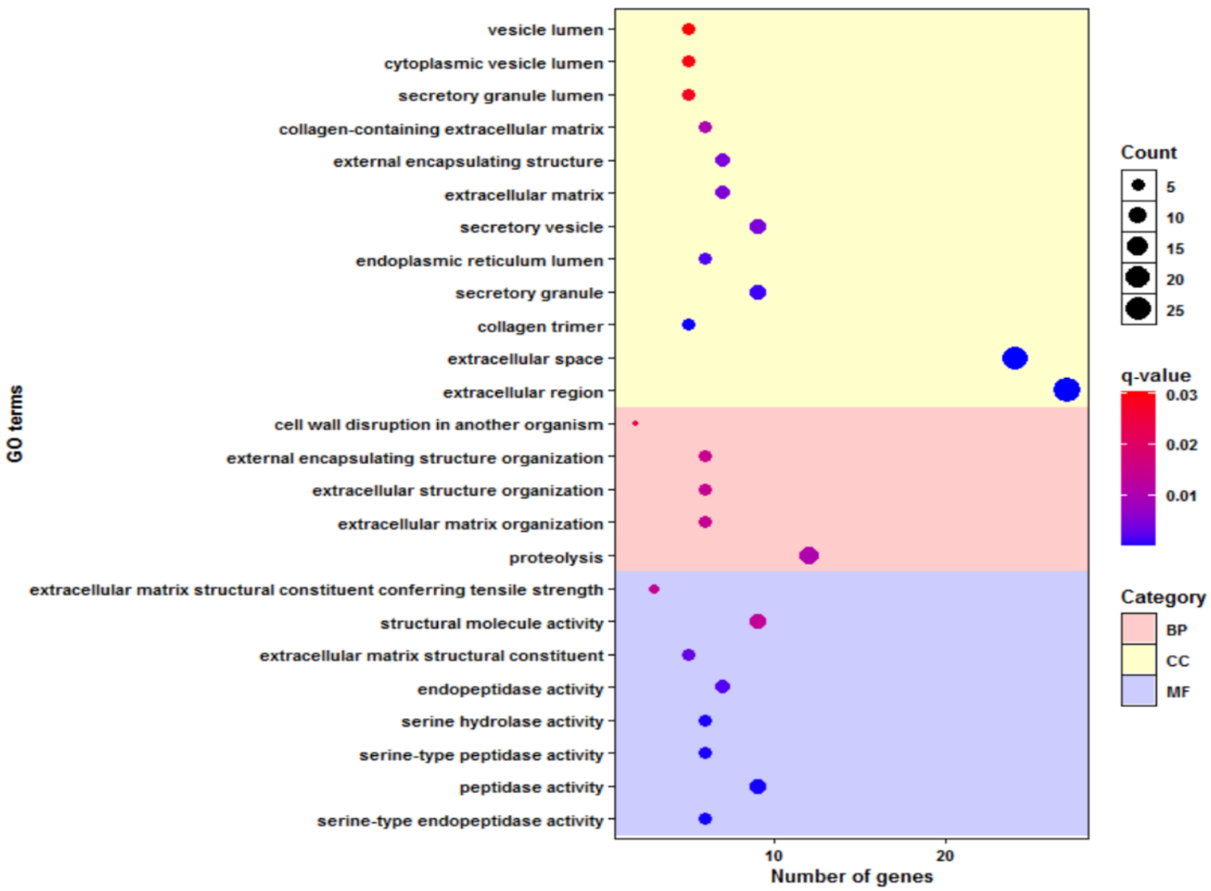

Figure 2. This figure shows the GO enrichment analysis for common genes.

GO: Gene ontology; BP: biological process; CC: cellular component; MF: molecular function

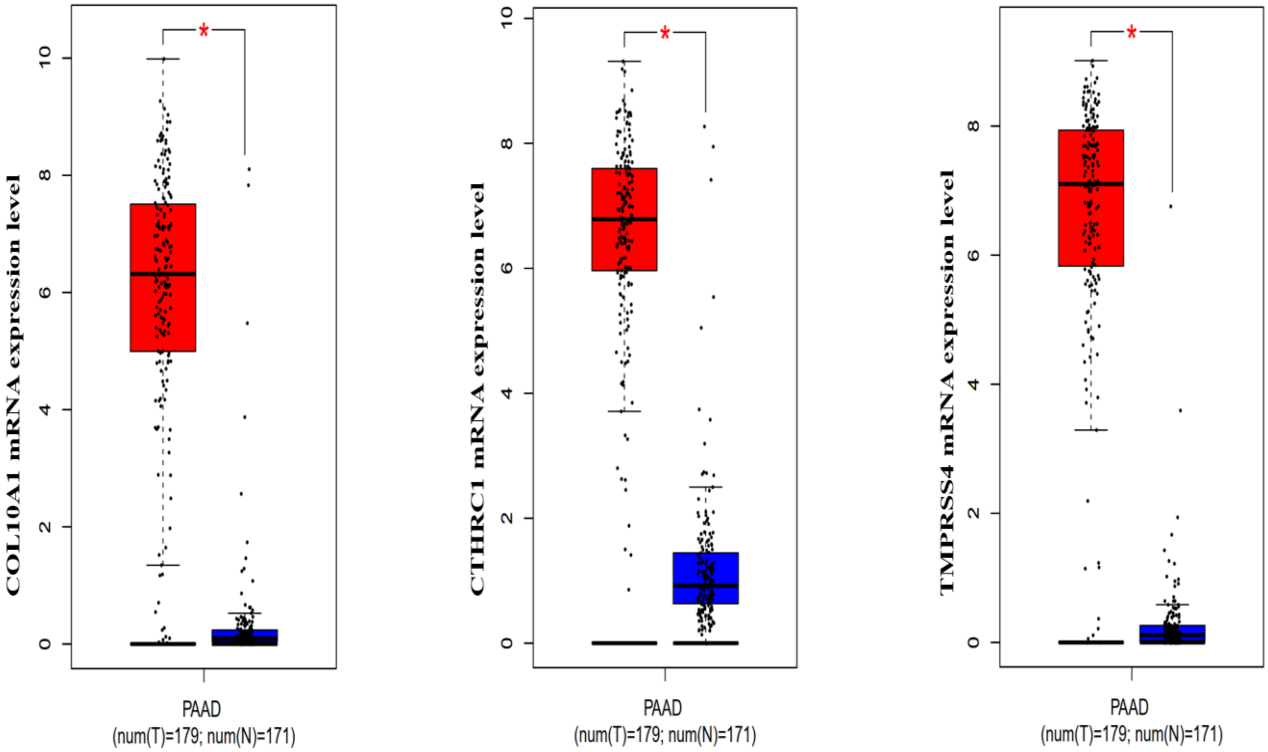

Figure 3. Validation of selected genes expression using RNA-Seq data in PAAD tissues: The GEPIA website was the source of the downloaded boxplots. The blue boxes denote the expression levels in normal tissues, the red boxes denote the expression levels in pancreatic adenocarcinoma tissues, and the asterisk means statistically significant. Statistical significance was defined as $P < 0.05$.

GEPIA: Gene expression profiling interactive analysis; PAAD: Pancreatic adenocarcinoma; num(T): sample size of tumor data; num(N): sample size of normal data

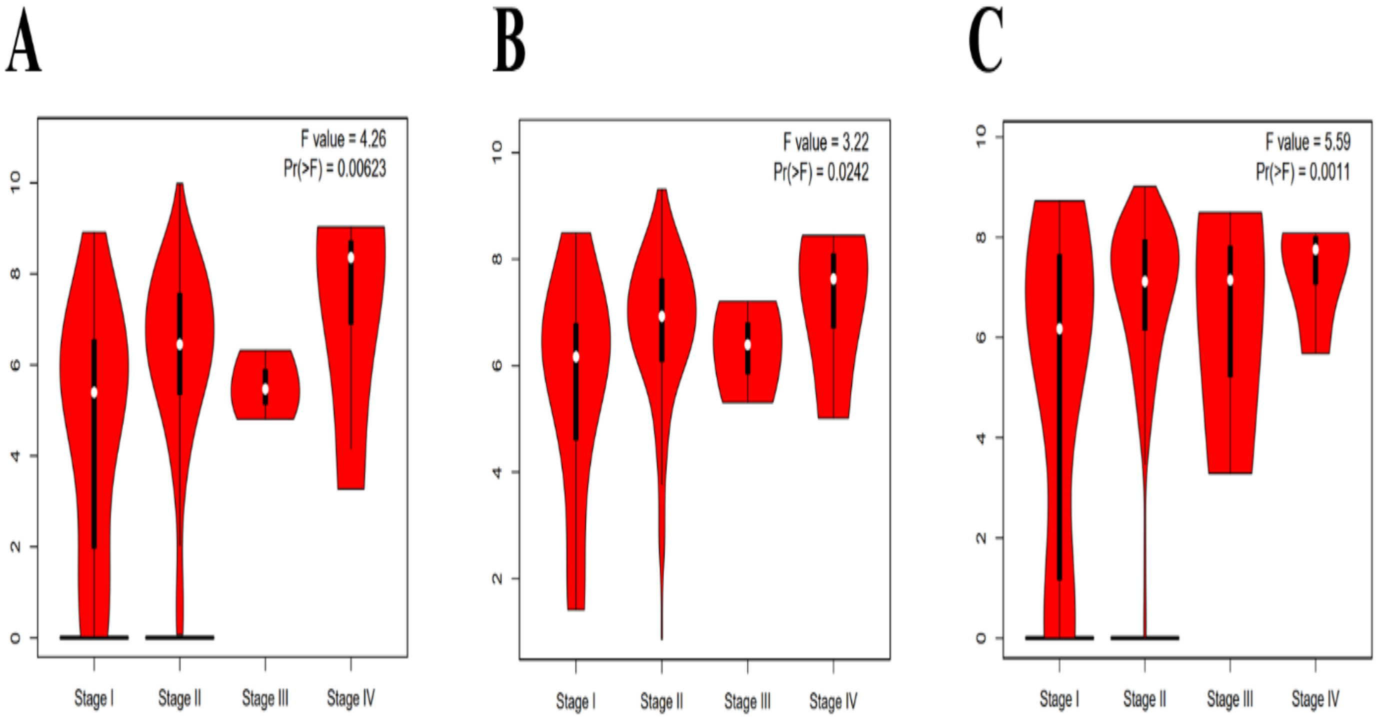

Figure 4. Correlation among the mRNA expression levels of selected genes and the pathologic stages of PAAD: Violin plots were generated based on TCGA data in GEPIA. F-value represents the statistical value of the F test; Pr (> F) represents the *P*-value. Statistical significance was defined as *P* < 0.05. (A) *COL10A1*; (B) *CTHRC1*; (C) *TMPRSS4*.

PAAD: Pancreatic adenocarcinoma; TCGA: The cancer genome atlas; GEPIA: Gene expression profiling interactive analysis

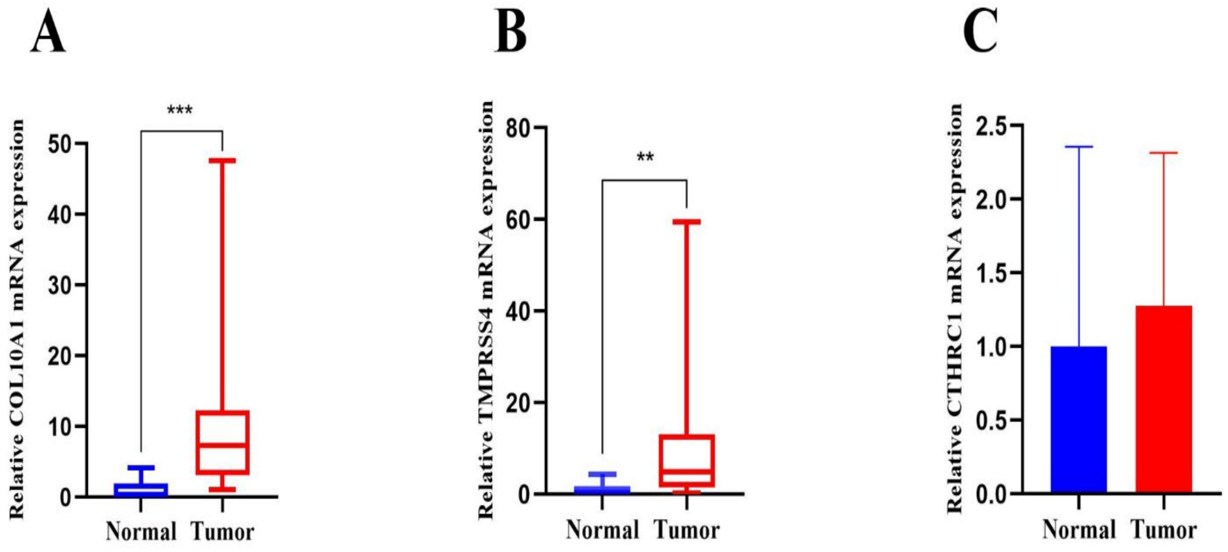

Figure 5. qRT-PCR compared mRNA expression levels of selected genes in tumor and adjacent normal tissues: Selected gene expression levels in PAAD tissues (n = 23) and matched normal pancreatic tissues (n = 19) were assessed in our independent patient sample using qRT-PCR. For *COL10A1* and *TMPRSS4*, Mann-Whitney was performed. For *CTHRC1*, an unpaired t-test was performed, and the value was shown as the mean ± SD of two separate experiments. The red boxes reflect the levels of expression in pancreatic adenocarcinoma tissues, whereas the blue boxes show the levels of expression in normal tissues. Statistical significance was defined as $P < 0.05$. (A) *COL10A1*; (B) *TMPRSS4*; (C) *CTHRC1*.

PAAD: Pancreatic adenocarcinoma; qRT-PCR: Quantitative real-time polymerase chain reaction; n: number

Table 1. Clinicopathological characteristics of PAAD patients

CPC	**Factors**	**Number (%)**
Gender	Male	15 (65.2)
	Female	8 (34.8)
Age (Year)	≤55	9 (39.1)
	>55	14 (60.9)
Histologic grade	Well-differentiated (grade 1)	6 (26.1)
	Moderately-differentiated (grade 2)	15 (65.2)
	Poorly-differentiated (grade 3)	2 (8.7)
	Undifferentiated (grade 4)	-
Lymph node status	Positive	17 (73.9)
	Negative	6 (26.1)
Tumor size	≤3	16 (69.5)
	>3	7 (30.5)

CPC: Clinicopathological characteristics; PAAD: Pancreatic adenocarcinoma

Table 2. Primers sequences

Target	**Sequence**	**Annealing temperature**	**PCR product lengths**
β-actin	F: 5' GCCTTTGCCGATCCGC 3' R: 5' GCCGTAGCCGTTGTCG 3'	59 ˚C	182 bp
COL10A1	F: 5' ACGCTGAACGATACCAAATGC 3' R: 5' TGCTCTCCTCTTACTGCTATACCT 3'	56 ˚C	115 bp
CTHRC1	F: 5' GGGAATGTCTGAGGGAAAG 3' R: 5' AATGTGAAATACCAACGCTGA 3'	56 ˚C	211 bp
TMPRSS4	F: 5' CTGGTTCTCTGCCTGTTTCG 3' R: 5' AAGTGGGTTTGCTGCTGTAG 3'	56 ˚C	86 bp

PCR: Polymerase chain reaction

Table 3. Description of GEO datasets used in the analysis

Accession	**Platform full name**	**Platform abbreviation**	**Samples (tumor/normal)**
GSE15471	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	GPL570	78 (36/42)
GSE28735	[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array	GPL6244	90 (45/45)
GSE62452	[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array	GPL6244	130 (69/61)
GSE55643	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	GPL6480	53 (45/8)
GSE22780	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	GPL570	10 (5/5)[a]
GSE71989	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	GPL570	21 (13/8)[b]
GSE46234	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	GPL570	6 (2/4)[c]

[a]: The GSM563307, GSM563308, GSM563315, GSM563316, GSM563317, and GSM563318 samples were removed from GSE22780 dataset; [b]: The GSM1849348 sample was removed from GSE71989 dataset because the patient is a sample of pancreatitis; [c]: The GSM1126853 and GSM1126854 samples were removed from GSE46234 dataset; GEO: Gene Expression Omnibus

Table 4. The most important feature genes selected by different attribute weighting algorithms for microarray dataset

Attribute [*]	Number of attribute weighting algorithms [**]
S100P	7
KCNN4	6
NHS	6
MLPH	6
CTSE	6
TRIM29	6
SLC6A14	6
SDR16C5	6
PLEK2	5
RHBDL2	5
MYOF	5
GALNT5	5
ECT2	5
CTHRC1	5
LY6E	5
CAPG	5
ASAP2	5
KRT19	5
TSPAN1	5
MST1R	5
MBOAT2	5
S100A6	5
STYK1	5
GJB2	5
S100A11	5
TMPRSS4	5
TRIP10	5
DNTTIP1	5
SULF1	5
CDH3	5
COL10A1	5
SLPI	5
IFI27	5
LAMB3	5
LAMC2	5
SFN	5
AMIGO2	5
NQO1	5
PYGB	5
PKM	4
CSTB	4

COL5A1	4
ANO1	4
MET	4
NTM	4
ZDHHC7	4
SLC16A3	4
COL11A1	4
SDC1	4
MARVELD1	4
ITGA2	4
GPRC5A	4
CEACAM6	4
PCDH7	3
S100A16	3
HK1	3
HK2	3
NOX4	3
DKK1	3
ITGA3	3
LPCAT4	3
XAF1	3
FERMT1	3
FBXO32	3
OSBPL3	2
APOL1	2
MYO1E	2
CLIC1	2
DLG5	2
ACSL5	2
PLPP4	2
FN1	2
CD109	2
FGD6	2
ENO2	2
PLAU	2
ACTN1	2
LRRC8A	2
IGFBP3	2
PTTG1IP	2
RRAS	2
KRT17	2
ANTXR1	2
RSAD2	2

VILL	2
ZNF185	2
MTMR11	2
COL17A1	2
PI3	2
PRSS1	1
ERP27	1
CTRB2	1
CLPS	1
AQP8	1
RPS4Y1	1
GP2	1
CTRC	1
PLA2G1B	1
PNLIP	1
CPB1	1
OLFM4	1
CPA1	1
SERPINI2	1
GEM	1
FAM3B	1
CELA3B	1
GCG	1
REG1B	1
ANO5	1
UBE2T	1
CPA2	1
CELA2B	1
PNLIPRP2	1
REG3A	1
PNLIPRP1	1
REG1A	1
ADAM9	1
NT5C2	1
CXCR4	1
SYCN	1
PLAUR	1
RAI14	1
TRIP4	1
ALB	1
SPARC	1
PPY	1
EPHX2	1

C1S	1
SULF2	1
IL1RAP	1
FHL2	1
RIC3	1
COL1A2	1
COL5A2	1
CORO2A	1
MMP11	1
CFH	1
PPP1R18	1
CST1	1
THBS2	1
RAB31	1
MX2	1
CTSK	1
FAP	1
EXOC1	1
TIMP1	1
FXYD5	1
COL8A1	1
SPOCK1	1
S100A10	1
CCL20	1
IAPP	1
ANKRD10	1
VCAN	1
IGFBP7	1
CD55	1
LEMD1	1
SLC44A1	1
PIAS3	1
NMU	1
NLRP14	1
ANGPTL7	1
KLK7	1
ABCA12	1
ZNF425	1
FSD1L	1
WFDC1	1
SKI	1
COL6A6	1
ZNF583	1

DUOXA1	1
SCN5A	1
AGR2	1
CST11	1
CACNB4	1
C8orf48	1
PEX14	1
MRO	1
VSIG1	1
GLT8D1	1
OR3A3	1
CDH22	1
ZNF93	1
LMNTD1	1
NPTX1	1
ZFP28	1
MOSPD3	1
MUC1	1
STAC	1
ZNF251	1
PITHD1	1
CCR8	1
TNMD	1
FZD2	1
NDNF	1
KCNMB4	1
SLC9A2	1
KBTBD12	1
HHIP	1
ADH7	1
SLC26A9	1
DRP2	1
IL31RA	1
PITX1	1
MMP28	1
GLUD2	1
KLK10	1
PPP1R14A	1
ATF3	1
WRAP53	1
SERPINB5	1
FCN3	1
KCNH1	1

DOK3	1
TBX15	1
KCTD4	1
PLXNB3	1
C6orf118	1
SSMEM1	1
ZNF257	1
HBD	1
BEST3	1
MUC17	1
GJB3	1
BATF	1
LRFN5	1
AFAP1L2	1
CAPNS2	1
POU1F1	1
CLDN18	1

[*] The official symbol for each human gene; [**]The number of algorithms that selected the attribute

Table 5. The most important feature genes selected by different attribute weighting algorithms for TCGA dataset

Attribute [*]	Number of attribute weighting algorithms [**]
INS	7
PNLIP	7
CTRB1	6
CTRB2	6
PRSS1	6
GAD2	6
PDX1	6
COL11A1	6
COL10A1	6
CGB5	5
COMP	5
G6PC2	5
PPY	5
ADAM8	4
FOXL1	4
CST2	4
CTHRC1	4
STX1A	4
CGB	4
MMP11	4
CTRC	4
C19orf30	4
TRY6	4
CST1	4
HTR1D	4
ONECUT3	3
CARD11	3
ADAMTS14	3
MMP14	3
ASPHD1	3
TIMP1	3
CELA3A	3
CGB8	3
CALHM3	3
ST8SIA3	3
TM4SF4	3
HSDL2	2
ENC1	2
C14orf105	2
GCG	2
FRMD5	2

C20orf103	2
CELSR3	2
ERN2	2
REG1A	2
EPS8L3	2
TCN1	2
ANXA10	2
DCBLD1	2
IGFL2	2
TFF2	2
IAPP	2
SYCN	2
PRSS3	2
CELA2A	2
CHST4	2
FGF19	2
ONECUT2	2
KNG1	1
SLC34A2	1
AGXT	1
HRG	1
KRT15	1
RHCG	1
FABP1	1
HP	1
CPB2	1
UMOD	1
XIST	1
KDM5D	1
PLG	1
FGG	1
HPD	1
CXCL17	1
ACSM2A	1
SLC26A3	1
EIF1AY	1
TF	1
C4BPA	1
AHSG	1
DDX3Y	1
SFTPA2	1
KRT4	1
TG	1

C20orf114	1
TAT	1
USP9Y	1
HMGCS2	1
PRODH2	1
KRT13	1
UGT3A1	1
UTY	1
DPYS	1
APOA2	1
GSTA2	1
CYP4A11	1
AQP2	1
APOC3	1
DSG3	1
SLC6A19	1
AGXT2	1
SLC12A1	1
RPS4Y1	1
KRT5	1
SFTPA1	1
PIP	1
G6PC	1
UGT1A9	1
COL17A1	1
MSMB	1
ADIPOQ	1
GLYATL1	1
KRT20	1
HAO2	1
SPRR3	1
ORM1	1
ACSM2B	1
FGA	1
SLC9A3	1
UGT2B7	1
MYBPC1	1
SCGB1A1	1
APOA1	1
NAT8	1
UGT1A10	1
KRT6B	1
CDH16	1

SFTPC	1
SFTA3	1
C20orf56	1
NKX2.1	1
ORM2	1
PSCA	1
KRT6A	1
ALB	1
KRT14	1
HNF1B	1
GLYAT	1
KRT17	1
PCK1	1
TMEM213	1
APOH	1
SCEL	1
APOB	1
FGB	1
UGT2A3	1
SFTPB	1
DMBT1	1
ABP1	1
PRAP1	1
CYP3A4	1
SLC2A2	1
AKR1B10	1
NAPSA	1
BHMT	1
F11	1
C19orf77	1
HABP2	1
KRT16	1
UGT2B15	1
SERPINA6	1
SERPINB5	1
TMPRSS4	1
SLC39A5	1
TM4SF5	1
CEACAM7	1
CEACAM5	1
HNF4A	1
PGC	1
C1QTNF6	1

S100P	1
PAH	1
CHGA	1
ALDOB	1
CDHR5	1
GP2	1
TFF1	1
MUC5B	1
CDHR2	1
SERPINA4	1
SLC3A1	1
CLCA4	1
AMBP	1
GC	1
SLC17A6	1
VIL1	1
CDX2	1
APCS	1
FGL1	1
LRP2	1
CPB1	1
CTSE	1
MSLN	1
LOC84740	1
CEACAM6	1
CRISP3	1
CLDN18	1
OLFM4	1
KLK6	1
C6orf222	1
MUC13	1
FXYD2	1
REG3A	1
SST	1
TTR	1
LGALS4	1
USH1C	1
MUC6	1
DPCR1	1
CDH17	1
CRP	1
NKX2.3	1
PPEF1	1

LEMD1	1
SPINK1	1
PTPRN	1
CLPS	1
MNX1	1
CPA1	1
REG4	1
CELA3B	1
CSMD2	1
GRIN2D	1
ANKH	1
EPYC	1
INHBA	1
SCNN1B	1
GOLGA8E	1
MBOAT4	1
ZNF679	1
SYT13	1
SCNN1G	1
LOC100127888	1
CUZD1	1
GJD2	1
SCG2	1
NKX2.6	1
OPN1LW	1
SALL4	1
ENPP3	1
AMY2B	1
LOC126536	1
GNAT3	1
CABP7	1
DIRC1	1
AMY2A	1
GDF6	1
NPC1L1	1
ASPN	1
POU6F2	1

* The official symbol for each human gene; **The number of algorithms that selected the attribute; TCGA: The cancer genome atlas

Table 6. List of 33 genes overlapped between two independent analyses (Genes were divided into two groups, upregulated differentially expressed genes and downregulated differentially expressed genes, based on the GEPIA online tool)

Official symbol [*]	Official full name
A. Upregulated differentially expressed genes	
COL11A1	collagen type XI alpha 1 chain
COL10A1	collagen type X alpha 1 chain
PPY	pancreatic polypeptide
CTHRC1	collagen triple helix repeat containing 1
MMP11	matrix metallopeptidase 11
CST1	cystatin SN
TIMP1	TIMP metallopeptidase inhibitor 1
GCG	glucagon
COL17A1	collagen type XVII alpha 1 chain
KRT17	keratin 17
SERPINB5	serpin family B member 5
TMPRSS4	transmembrane serine protease 4
S100P	S100 calcium binding protein P
CTSE	cathepsin E
CEACAM6	CEA cell adhesion molecule 6
CLDN18	claudin 18
OLFM4	olfactomedin 4
LEMD1	LEM domain containing 1
B. Downregulated differentially expressed genes	
PNLIP	pancreatic lipase
CTRB2	chymotrypsinogen B2
PRSS1	serine protease 1
CTRC	chymotrypsin C
REG1A	regenerating family member 1 alpha
IAPP	islet amyloid polypeptide
SYCN	syncollin
RPS4Y1	ribosomal protein S4 Y-linked 1
ALB	albumin
GP2	glycoprotein 2
CPB1	carboxypeptidase B1
REG3A	regenerating family member 3 alpha
CLPS	colipase
CPA1	carboxypeptidase A1
CELA3B	chymotrypsin like elastase 3B

[*] The official symbol for each human gene; GEPIA: Gene expression profiling interactive analysis

Table 7. GO enrichment analysis of common genes was retrieved using g:Profiler

Source	Term_id	Term name	Count	Adjusted_*P*_value
MF	GO:0004252	Serine-type endopeptidase activity	6	0.000126487
MF	GO:0008233	Peptidase activity	9	0.000146382
MF	GO:0008236	Serine-type peptidase activity	6	0.000222044
MF	GO:0017171	Serine hydrolase activity	6	0.000249781
MF	GO:0004175	Endopeptidase activity	7	0.001771483
MF	GO:0005201	Extracellular matrix structural constituent	5	0.002829998
MF	GO:0005198	Structural molecule activity	9	0.013651577
MF	GO:0030020	Extracellular matrix structural constituent conferring tensile strength	3	0.013960882
BP	GO:0006508	Proteolysis	12	0.010409277
BP	GO:0030198	Extracellular matrix organization	6	0.013745579
BP	GO:0043062	Extracellular structure organization	6	0.013991681
BP	GO:0045229	External encapsulating structure organization	6	0.014494631
BP	GO:0044278	Cell wall disruption in another organism	2	0.023847118
CC	GO:0005576	Extracellular region	27	7.72E-12
CC	GO:0005615	Extracellular space	24	8.09E-11
CC	GO:0005581	Collagen trimer	5	6.91074E-05
CC	GO:0030141	Secretory granule	9	0.001027823
CC	GO:0005788	Endoplasmic reticulum lumen	6	0.001660438
CC	GO:0099503	Secretory vesicle	9	0.004432751
CC	GO:0031012	Extracellular matrix	7	0.00457907
CC	GO:0030312	External encapsulating structure	7	0.004631483
CC	GO:0062023	Collagen-containing extracellular matrix	6	0.010027604
CC	GO:0034774	Secretory granule lumen	5	0.028209524
CC	GO:0060205	Cytoplasmic vesicle lumen	5	0.029469983
CC	GO:0031983	Vesicle lumen	5	0.030334047
KEGG	KEGG:04974	Protein digestion and absorption	8	9.42E-09
KEGG	kegg:04972	Pancreatic secretion	6	1.41923E-05

GO: Gene ontology; BP: biological process; CC: cellular component; MF: molecular function; KEGG: kyoto encyclopedia of genes and genomes