Bioinformatics analysis reveals TSPAN1 as a candidate biomarker of progression and prognosis in pancreatic cancer

Chenhui Ma[1,2], ZeLong Cui[3], YiChao Wang[2], Lei Zhang[1], JunYe Wen[1], HuaiBin Guo[1], Na Li[1], WanXing Zhang[1]*

ABSTRACT

Pancreatic cancer (PCC) is a common malignant tumor of the digestive system that is resistant to traditional treatments and has an overall 5-year survival rate of <7%. Transcriptomics research provides reliable biomarkers for diagnosis, prognosis, and clinical precision treatment, as well as the identification of molecular targets for the development of drugs to improve patient survival. We sought to identify new biomarkers for PCC by combining transcriptomics and clinical data with current knowledge regarding molecular mechanisms. Consequently, we employed weighted gene co-expression network analysis and differentially expressed gene analysis to evaluate genes co-expressed in tumor versus normal tissues using pancreatic adenocarcinoma data from The Cancer Genome Atlas and dataset GSE16515 from the Gene Expression Omnibus. Twenty-one overlapping genes were identified, with enrichment of key Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathways, including epidermal growth factor receptor signaling, cadherin, cell adhesion, ubiquinone, and glycosphingolipid biosynthesis pathways, and retinol metabolism. Protein-protein interaction analysis highlighted 10 hub genes, according to Maximal Clique Centrality. Univariate and multivariate COX analyses indicated that TSPAN1 serves as an independent prognostic factor for PCC patients. Survival analysis distinguished TSPAN1 as an independent prognostic factor among hub genes in PCC. Finally, immunohistochemical staining results suggested that the TSPAN1 protein levels in the Human Protein Atlas were significantly higher in tumor tissue than in normal tissue. Therefore, TSPAN1 may be involved in PCC development and act as a critical biomarker for diagnosing and predicting PCC patient survival.

KEYWORDS: Pancreatic cancer; TSPAN1; TCGA; GEO dataset; diagnosis and prognosis

INTRODUCTION

Pancreatic cancer (PCC) is a highly immunosuppressive and malignant tumor of the digestive system, with insidious onset, rapid disease progression, and no breakthroughs in long-term efficacy or prognosis. In a recent report, PCC ranked 10th among cancers in the United States for its incidence but 4th for its mortality [1]. Furthermore, it is estimated to reach the second largest cause of cancer-related death by 2030 [2]. In China, PCC does not rank among the top 5 in mortality. However, the proportion of death caused by PCC has increased by 9% over the past 10 years, primarily accounted for by changes in lifestyle and diet and the acceleration of population aging [3]. The specific pathogenesis of PCC remains unclear, but a large number of clinical and epidemiological findings have revealed that smoking and obesity prolong pancreatitis, and diabetes has been identified as a significant independent risk factor for the development of PCC [4]. PCC progresses rapidly, and early detection and diagnosis are crucial for the prognosis of PCC patients [4]. Nonetheless, imaging examination, serological markers, and other diagnostic methods have limitations, especially for early PCC diagnosis, which compromises clinical care and prognosis [4,5]. The major strategies of PCC management include surgery, chemotherapy, radiotherapy, molecular guided therapy, and immunotherapy [4]. However, due to PCC's pathological and clinical characteristics, chemotherapy and radiotherapy have little benefit for patients with PCC, and currently, the most successful therapeutic choice for PCC is surgical resection [6]. Through the rapid growth and comprehensive implementation of gene detection technology, molecularly targeted drugs have been increasingly used in clinical practice. Nevertheless, molecular targeted therapy has not been successfully applied to PCC due to the poor understanding of its molecular pathological mechanism [7].

For a thorough and rigorous understanding of PCC, a clinical risk assessment needs to be combined with clinical

[1]Department of Hepatobiliary, Hebei General Hospital, Shijiazhuang, China
[2]Graduate School of North China University of Science and Technology, Tangshan, China
[3]Department of Hematology, Qilu Hospital of Shandong University, Jinan, China

*Corresponding author: Dr. Wanxing Zhang, Department of Hepatobiliary, Hebei General Hospital, Xinhua District (Heping West Road, No. 348), Shijiazhuang, Hebei. Phone: +86 13931881946. E-mail: zhangwx@hebmu.edu.cn

DOI: https://dx.doi.org/10.17305/bjbms.2020.5096

Submitted: 29 August 2020/ Accepted: 8 November 2020

Conflicts of interest statement: The authors declare no conflicts of interest.

characteristics. For example, the cancer stage, diagnostic grade, and cancer laterality in PCC are correlated with the patient's diagnostic age, overall survival (OS), and secondary malignancies [8]. The rapid development of microarray and sequencing technology provides a useful method and forum for the research of cancer and other diseases [9]. New biomarkers for diagnosis, treatment, and prognosis can be obtained by combining clinical data with molecular mechanisms [10]. Weighted gene co-expression network analysis (WGCNA) provides an approach for performing weighted network analysis in the R package. It can be used in research to characterize multiple sample and cluster-specific gene expression patterns and to detect highly related gene expression modules correlated with clinical characteristics [11]. WGCNA is also useful for identifying core genes as well as the function of co-expressed genes of tumors and other diseases [12,13]. For instance, Liu et al. successfully identified five lncRNAs associated with survival in hepatocellular carcinoma by co-expression analysis [14]. Differential gene expression analysis based on transcriptomics offers substantial insight into the molecular mechanisms of genome-regulated diseases, the transcriptional behavior of biological systems, and potential biomarkers for specific diseases [15].

We used WGCNA to explore and evaluate the etiology and molecular characteristics of PCC in a systematic manner, to measure transcriptional expression levels, and identified differentially expressed genes (DEGs) in PCC from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases. Moreover, we combined the DEG results with functional enhancement and protein-protein interaction (PPI) analysis, survival analysis, and Cox regression, and identified DEGs closely related to the prognosis of PCC. These studies constitute a basis for drug development and a potential reference for the clinical diagnosis and treatment of PCC.

MATERIALS AND METHODS

Study design

A detailed workflow of our study design is shown in Figure 1. We analyzed microarray data from GEO (GSE16515) and RNA-seq and clinical data from TCGA. The sets of DEGs identified by the limma R package, and the most highly co-expressed modules, which were identified by WGCNA, were evaluated for overlap. The 21 overlapping genes were subjected to functional analysis and PPI analysis. The correlation of the overlapping genes and clinical parameters, including OS, disease-free survival (DFS), and other prognostic factors, was evaluated, and immunohistochemistry (IHC) data were assessed to validate the expression of survival-related genes in the Human Protein Atlas (HPA).

Data collection and preprocessing

The GSE16515 dataset was accessed from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). We selected this dataset because it included stringent screening criteria and has gene expression profiles of 52 PCC samples that were obtained using the GPL570-55999 ([HG-U133_Plus_2] Affymetrix Human Genome U133 plus 2.0 Array), which is a well-established platform [16]. According to the manufacturer's Annotation document, the probes were first assigned corresponding genetic symbols, and then the median of all associated findings was calculated to exclude detection overlap for the same gene. Consequently, 21654 genes were evaluated. An additional set of RNA-sequencing data of 182 PCC samples was downloaded from the TCGA database (https://genome-cancer.ucsc.edu/) from the pancreatic adenocarcinoma (PAAD) cohort, along with full expression profiles and data on clinically relevant characteristics. The TCGA data were annotated using a Human hg38 gene track reference transcript array. As indicated by the edgeR package tutorial [17], genes with low read counts were not useful for further study. Therefore, in this study, the genes with CPM (count per million) at or above 1 were analyzed. In subsequent analysis, a total of 15035 genes with RPKM values were analyzed, and the RPKM function was filtered by the edgeR software package, which distinguishes the number of genes according to the gene length.

DEG analysis

The limma R package [18] provides an efficient approach for differential expression analysis of microarray and RNA-sequencing data. Therefore, it was utilized in this study to screen DEGs between non-malignant pancreatic samples and PCC tissues. DEGs were defined as genes with the cutoff criteria of $|\log FC| \geq 1.0$ and adj. $p < 0.05$. In the ggplot2 package in R, the DEGs of the TCGA-PAAD and GSE16515 datasets were presented through volcano plots [19].

Construction and identification of a gene co-expression network by WGCNA

Quality assessment of the data was performed, and a gene co-expression network was established using the WGCNA package in R for DEGs [20]. WGCNA reveals heavily clustered gene modules between specimens and connects the modules to outer template characteristics. Before building the network, the number of genes with different thresholds of expression was estimated, and the pickSoftThreshold function was used to construct a scale-free network. Next, the formula $aij = |Sij|\beta$ (aij: matrix of adjacency between gene i and gene j, Sij: Matrix of similarity made by Pearson correlation of all gene pairs) was used to construct an adjacency matrix [21].

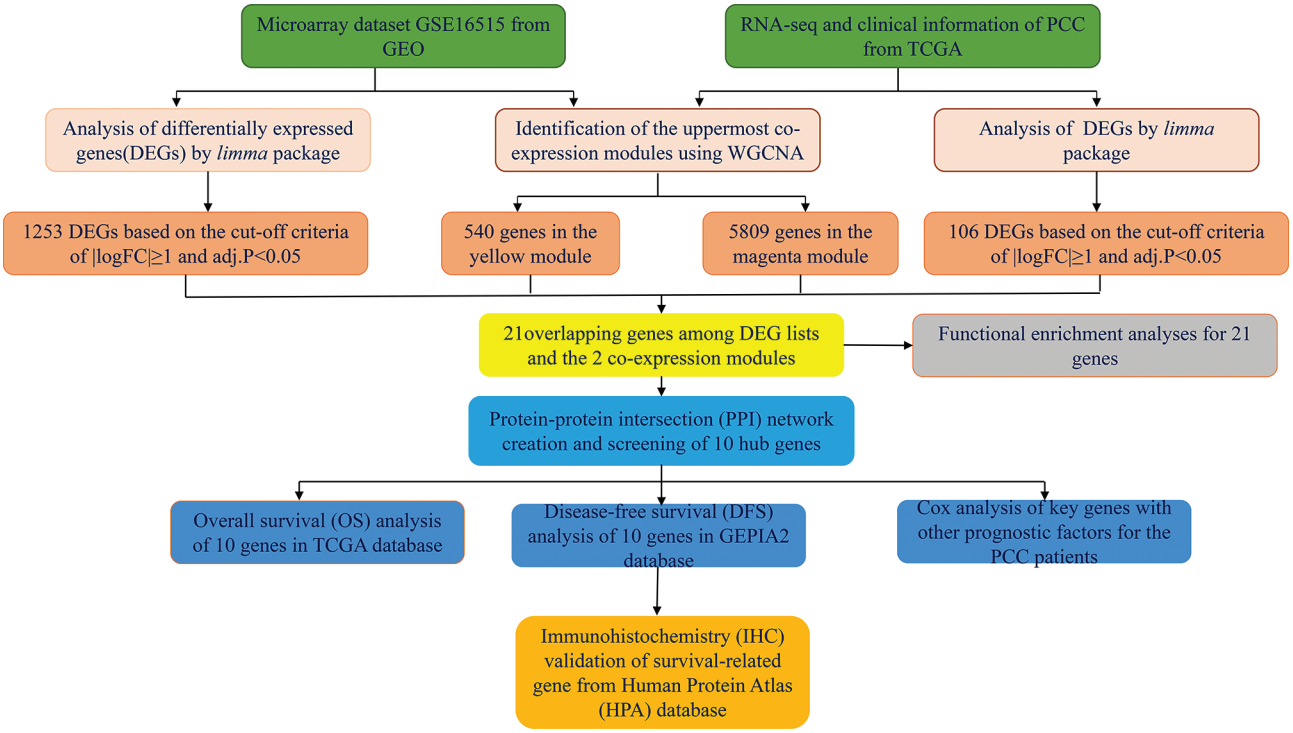

FIGURE 1. Overview of the study design. DEGs and co-expression modules in PCC were identified from microarray data from GEO (GSE16515); and RNA-seq and clinical data from TCGA-PAAD. Twenty-one overlapping genes were further evaluated by functional enrichment and protein-protein intersection analysis, and 10 Hub genes were identified. The correlations of these Hub genes with OS, DFS, and prognostic factors were assessed. IHC data from HPA were evaluated for further validation.

The adjacency was represented in a topological overlap matrix (TOM), and dissimilarity was represented in a corresponding dissimilarity matrix (1 – TOM). The topological overlap provides a measure of biological gene similarity based on the association between pairwise gene co-expression. To classify genes with similar expression characteristics in gene co-expression modules, a hierarchical clustering dendrogram of the 1 – TOM matrix was built.

Identification of clinically significant modules

The difference between the module-specific eigengenes (MEs) was calculated. A cutoff was selected for module dendrograms, and some modules were merged for further analysis [10]. Furthermore, the correlation between MEs and clinical trait information was evaluated to identify key modules that are significantly associated with PCC [21]. Next, the correlations of individual genes with clinical results were quantified by calculating the gene significance (GS) value [21]. Module significance (MS) was defined as the average GS for all genes in a module. In general, modules with the MS ranking of first or second were considered to be candidates for association with clinical characteristics. Overlapping genes between the DEGs and module genes were extracted from the co-expression network, and the genes closely related to the clinical phenotype of PCC were used to classify possible prognostic genes in a Venn diagram by the R-package [22].

Gene ontology (GO) and pathway enrichment analysis for genes of interest

To gain deeper insight into the role of the overlapping genes identified as described above, GO enrichment analysis was performed with classification according to biological process, cellular component, and molecular function designations; and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was carried out using an R package cluster profile [23]. Functional categories and pathways were enriched using a cutoff of $p < 0.05$, and the top 10 GO categories were selected.

Construction of a PPI network and screening of hub genes

The online tool STRING (Search Tool for the Retrieval of Interacting Genes), designed to predict functional interactions between proteins, has been used to create PPI networks for selected genes [24]. Genes with a score of ≥0.4 were selected using the STRING database to create a Cytoscape (v3.7.2) for visualization of the network model [25]. Maximal Clique Centrality (MCC) has been identified as a powerful index for detecting center nodes inside a network of co-expression [26]. Therefore, CytoHubba, a plugin at Cytoscape [26], was employed to measure each node's MCC. The 10 genes with the highest MCC values in this analysis were identified as core genes.

Survival analysis and prognostic values of hub genes

We calculated OS as an endpoint using the R packages survival and survminer. Survival curves were developed using the Kaplan-Meier method in R. In addition, the online platform GEPIA2 was used to calculate the correlation of DFS and hub genes expressed in patients with PCC [27]. Our survival analysis only included patients who had completed all follow-up examinations. The median expression values of hub genes were compared, and the samples were grouped into high-expression and low-expression groups. Survival-related hub genes with log-rank *p*-value significantly <0.05 were identified. Next, key survival genes and other prognostic predictors (age, gender, stage, and grade) were analyzed by univariate and multivariate COX analysis to assess the robustness of these genes compared with other prognostic indicators and to determine whether the key hub genes could be used as independent prognostic factors for PCC.

Validation of the HPA database for protein expressions in survival-related hub genes

The HPA database (https:/www.proteinatlas.org/) is a comprehensive resource that enables researchers to access a wide range of transcriptional and proteomic data from different tissues and cells [28]. Moreover, protein expression patterns based on immunohistochemistry (IHC) have become a universal immunostaining application for the determination of the relative protein position and abundance [29]. Therefore, HPA was used to determine the abundance of proteins encoded by survival genes in PCC and control tissues.

Ethics statement

This study protocol was reviewed and approved by the Hebei General Hospital Ethics Committee (No:202041). All patient data included in this study were de-identified.

RESULTS

Construction of PCC co-expression modules

To identify sets of genes that are co-expressed in PCC, we used WGCNA to sort genes from TCGA and GEO into modules. When the soft threshold values of $\beta = 2$ and 9 were selected, the connectivity between genes conformed to the distribution of the scale-free network (Figure 2A-D). We employed hierarchical clustering and dynamic branch cutting to recognize different co-expression modules of PCC and represented them by different colors. Ten modules of data from TCGA-PAAD (Figure 3A) and 12 modules from GSE16515 (Figure 3B) were detected after the fusion of related modules. Table S1 lists the number of genes present in the co-expression modules. To evaluate the correlation between each module and two clinical characteristics (cancer and normality), we plotted a heat map of module-trait relationships. The TCGA-PAAD magenta module and the GSE16515 yellow module represented the greatest association with the normal tissue (magenta module: $r = -0.21$, $p = 0.005$; yellow module: $r = -0.78$, $p = 1e{-}11$) (Figure 3C-D).

Identification of genes between the DEG lists and co-expression modules

To further evaluate the differential expression pattern in PCC, we identified sets of DEGs. After data preprocessing and quality assessment through the limma package, 106 DEGs in TCGA-PAAD (Figure 4A) and 1253 DEGs in GSE16515 (Figure 4B) were identified to be dysregulated in tumor tissues. We further analyzed these DEGs according to their distribution in the co-expression modules. As shown in Figure 4C, the TCGA-PAAD magenta module had 5809 DEGs, and the GSE16515 yellow module had 540 DEGs. A total of 21 overlapping genes were identified as candidates for validation (Figure 4C).

Functional enrichment analyses for 21 overlapping genes

To provide additional insight into the functional roles of the 21 dysregulated genes that overlapped in the DEG lists and co-expression modules, we performed KEGG and GO enrichment analyses using the clusterProfiler package. Several GO-enriched gene sets were observed (Figure 5A). Genes in the biological process category were primarily concentrated in O–glycan processing, regulation of the epidermal growth factor receptor signaling pathway, digestive system process, regulation of the ERBB signaling pathway, and protein localization to the cell periphery. In the cellular component category, enriched components included desmosomes, cell cortex part, cornified envelope, cortical cytoskeleton, and microvillus membrane. Moreover, in the molecular function category, cadherin binding, cell adhesion molecule binding, and epidermal growth factor receptor binding were the top functions of the 21 genes. In KEGG enrichment analysis, the genes were enriched in the biosynthesis of ubiquinone and other terpenoid-quinone, glycosphingolipid biosynthesis – lacto and neolacto series, and retinol metabolism (Figure 5B).

PPI network construction and hub gene identification

For a more comprehensive understanding of the functions of the 21 overlapping genes, we constructed a PPI network. The network contained 19 nodes and 46 edges (Figure 6A).

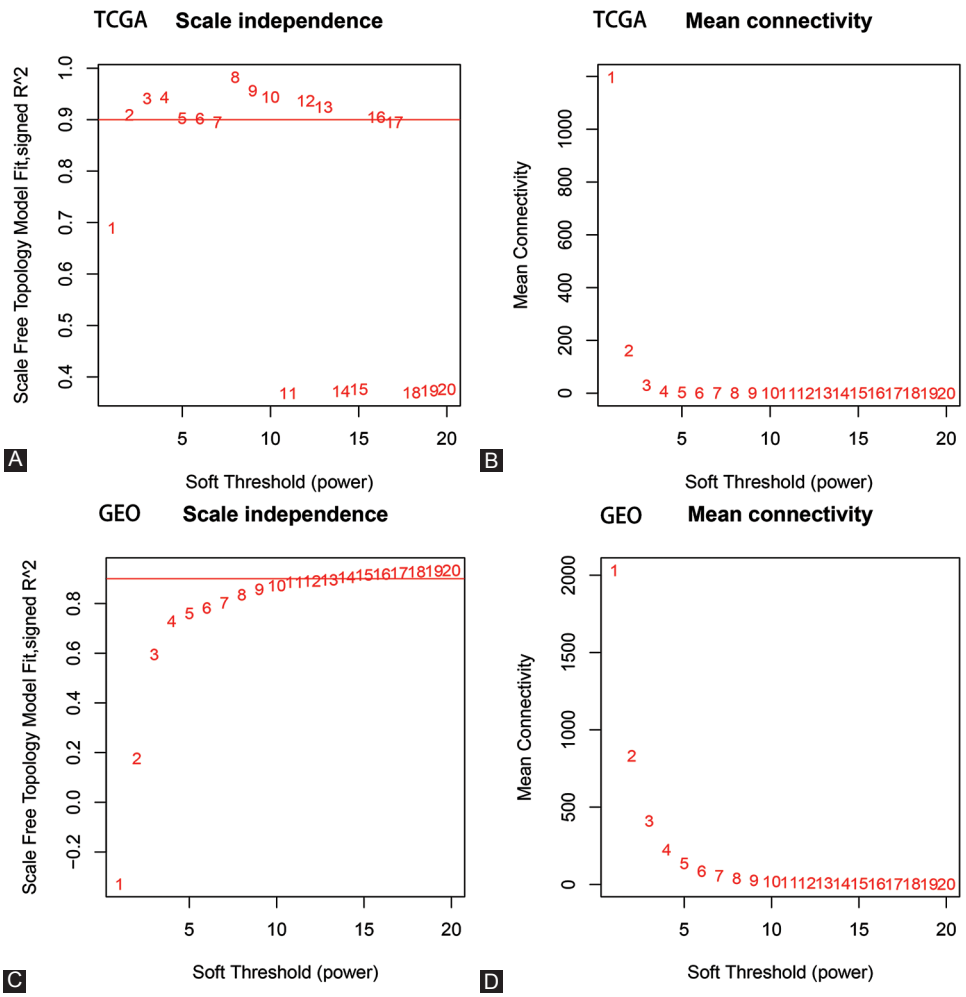

FIGURE 2. Identification of soft threshold weights by WGCNA. (A-D) Scale-free fitting index and average connectivity analysis of different soft-threshold weights of data from TCGA-PAAD (panels A and B) and GSE16515 (panels C and D).

Further analysis identified core genes within the PPI network (Figure 6B). The 10 genes with the highest MCC scores, including Tetraspanin-1 (TSPAN1), E3 ubiquitin-protein ligase CBL-C (CBLC), transmembrane protein 45B (TMEM45B), mitotic interactor and substrate of PLK1 (MISP), FXYD domain-containing ion transport regulator 3 (FXYD3), beta-1,3-N-acetylglucosaminyltransferase-3 (B3GNT3), anterior gradient protein 2 (AGR2), Plakophilin-3 (PKP3), S100P, and mucin-13 (MUC13), were identified as Hub genes.

Survival analysis and prognostic values of hub genes

To evaluate the clinical value of Hub gene expression, we determined whether they may be associated with survival or prognosis of PCC patients (Figure 7). Notably, high expression levels of TSPAN1 were significantly linked to the poor OS in PCC ($p < 0.05$) (Figure 7J). Although no substantial difference was found in the TSPAN1 expression level for DFS in PCC patients ($p > 0.05$) (Figure S1J), univariate and multivariate COX analysis outcomes indicated that TSPAN1 could serve as an independent prognostic factor for PCC patients (Table S2, Figure S2 and 8). These results suggest that TSPAN1 may represent a novel biomarker for PCC.

Validation of the HPA database for core survival genes

To verify the enhanced expression of TSPAN1 in PCC, we accessed IHC data from the HPA database. The TSPAN1 protein levels were considerably higher in tumor tissues relative to healthy tissues (Figure 9). Therefore, these findings confirm that elevated TSPAN1 expression, as determined at both the mRNA and protein levels, is aligned with worse prognosis and lower OS in PCC patients.

DISCUSSION

PCC, a common digestive tumor with a high degree of malignancy, tends to be aggressive and easily exacerbated by local nervous and vascular invasion. PCC tumors can form distal metastases in the early stages of cancer and often become resistant to traditional treatments such as chemotherapy and radiotherapy, making PCC prognosis extremely poor (5-year

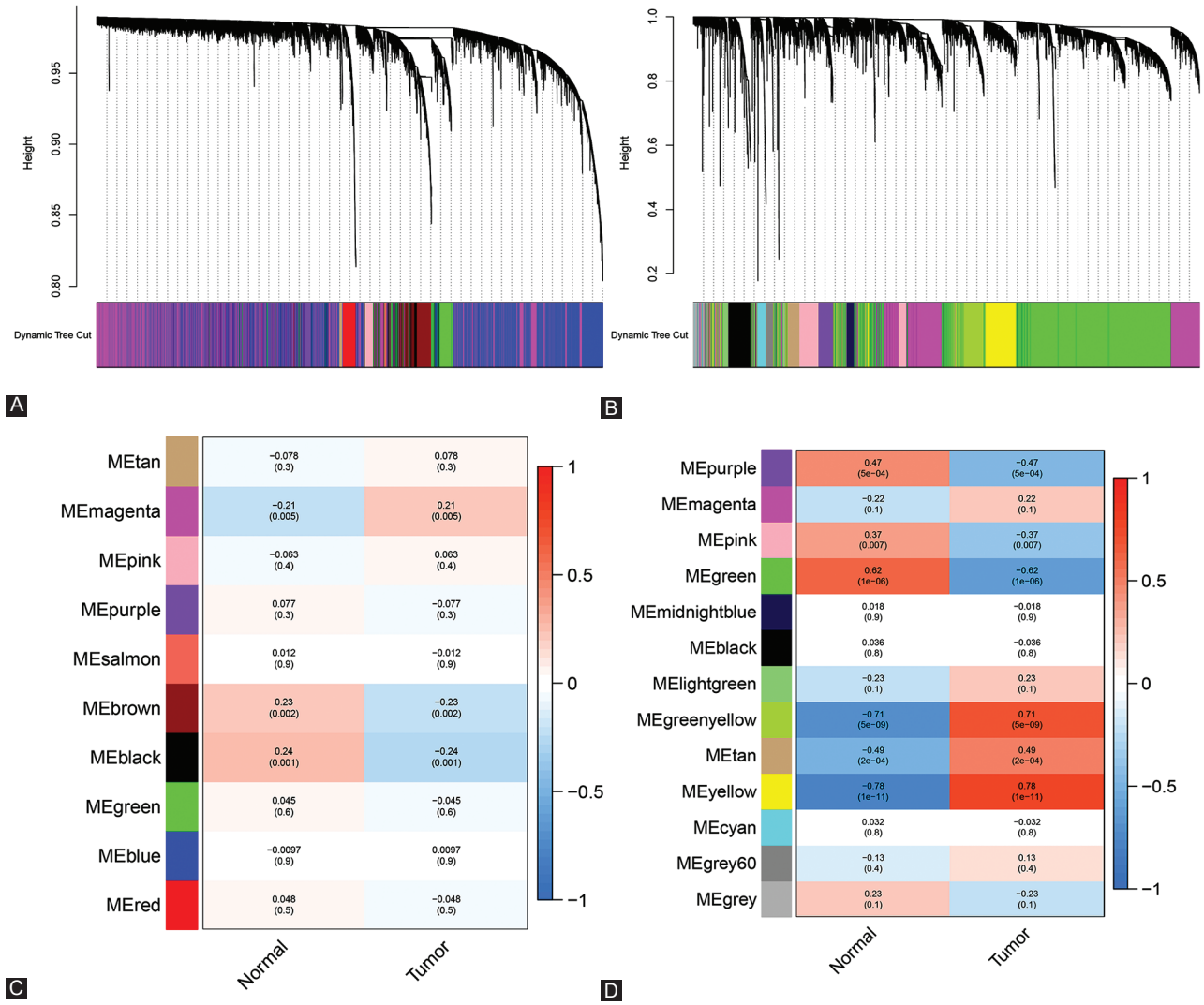

FIGURE 3. Relationships between PCC co-expression modules from TCGA-PAAD and GEO-PCC datasets. (A) Hierarchical gene clustering of TCGA co-expression modules according to the 1 − TOM matrix. Each module is color-coded. (B) Module-trait diagram for TCGA co-expression modules. Columns are colors and rows represent clinical features. (C) Hierarchical gene clustering of GEO co-expression modules according to the 1 − TOM matrix. Each module is color-coded. (D) Module-trait diagram for GEO co-expression modules. Columns are colors and rows represent clinical features.

OS rate of <7%) [30]. Although many breakthroughs have been made for diverse cancers, the development of a PCC drug remains challenging [31]. With the advent and advancement of targeted therapies and precise medication, traditional histopathological evaluation and diagnosis are becoming increasingly outdated. The most common type of mutation in PCC patients is the KRAS mutation. The KRAS gene mutation plays a major role in the occurrence and development of PCC, with a rate as high as 90% [7]. KRAS pathway-targeted therapies for PCC have been explored, but current limitations involving drug resistance and safety considerations hinder their applicability. Therefore, to promote the development of precision medicine, including individualization and standardization of targeted drugs, we need to continue to explore new clinical survival targets for diagnosis, prediction, and treatment of PCC patients. Research involving genomics and transcriptomics has the potential to provide reliable and detailed information for clinical precision therapy, to extend patient survival, and to act as a guide for new drug development, including target selection for therapeutic trials and population screening. Matching different molecular subtypes to clinical drugs and treatment regimens have the potential to advance PCC therapy. Therefore, we used advanced bioinformatics methods and transcriptional and clinical data from curated databases to identify new potential molecular targets.

In this study, we identified 21 genes with consistent expression patterns using integrated WGCNA and DEG analysis. As indicated by GO functional enrichment results, the function of these 21 genes includes regulation of epidermal growth factor receptor (EGFR) signaling, cadherin binding, and cell adhesion molecule binding. Notably, EGFR (also known as ERBB1 and HER1) is overexpressed in 90% of PCC cells [32]. EGFR is a transmembrane receptor tyrosine kinase [33] that belongs to the ERBB family of cell surface

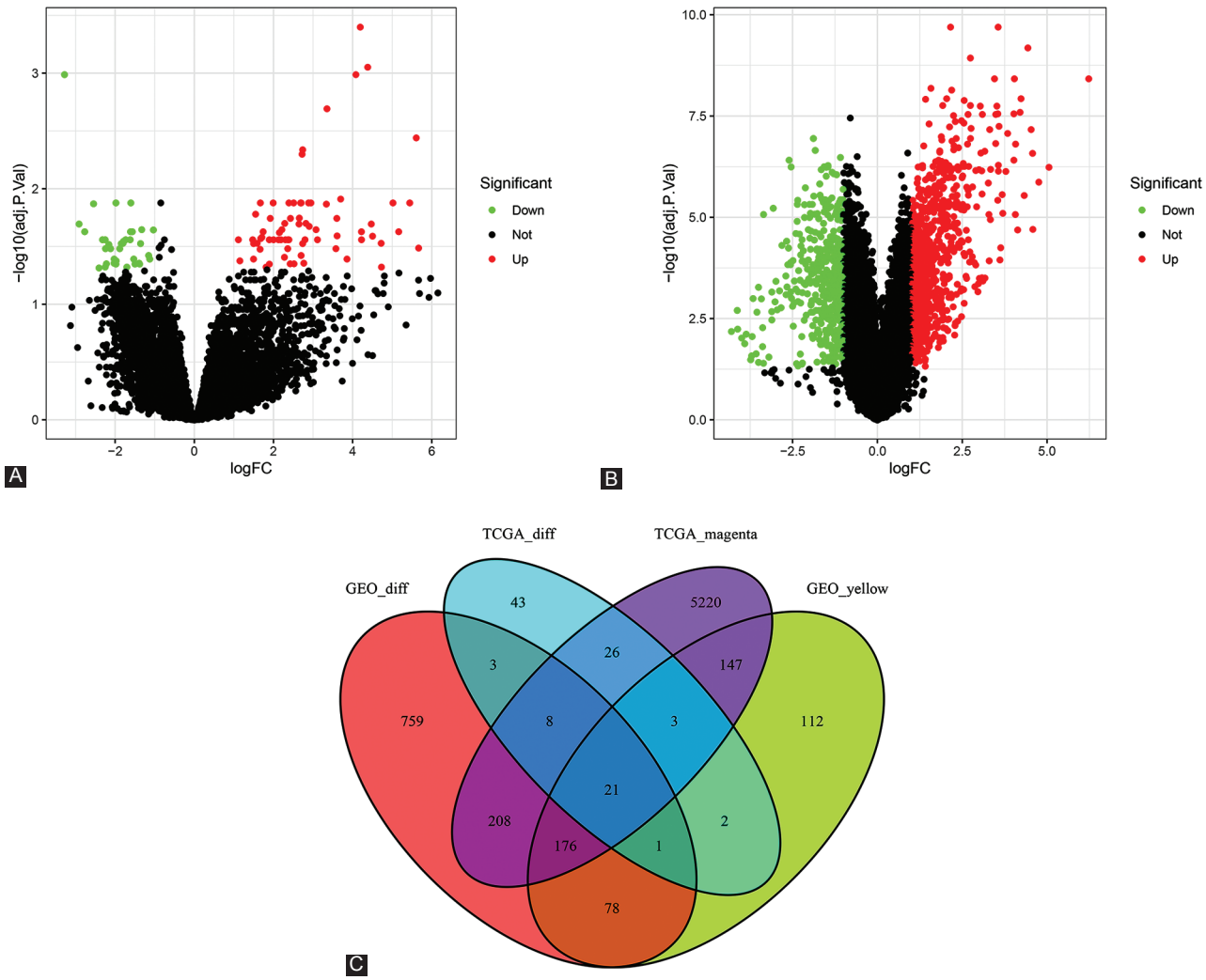

FIGURE 4. Detection of differentially expressed genes (DEGs) in PCC from the GSE16515 and TCGA datasets. (A-B) Volcano plots representing DEGs from the TCGA dataset (panel A) and the GSE16515 dataset (panel B). (C) The Venn diagram showing the intersection of genes between co-expression modules and DEG lists.

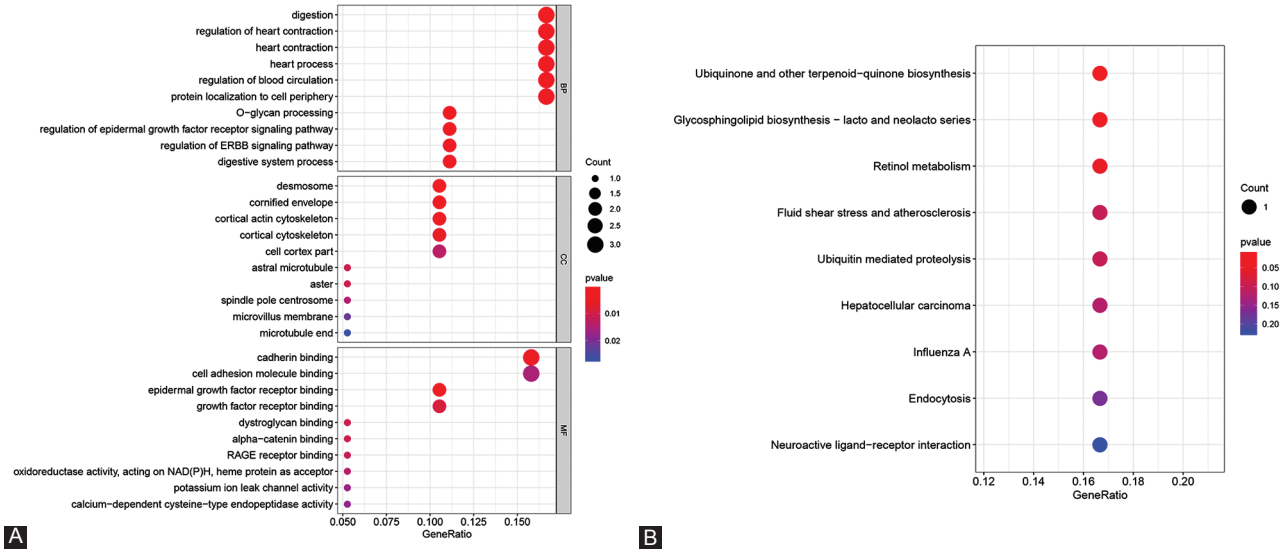

FIGURE 5. Functional annotation of pathways that are differentially activated in pancreatic cancer (PCC). (A) GO analysis of pathways modulated in PCC. (B) KEGG pathways enriched in PCC. The color represents the modified *p*-values. The scale of the spots indicates the number of genes involved.

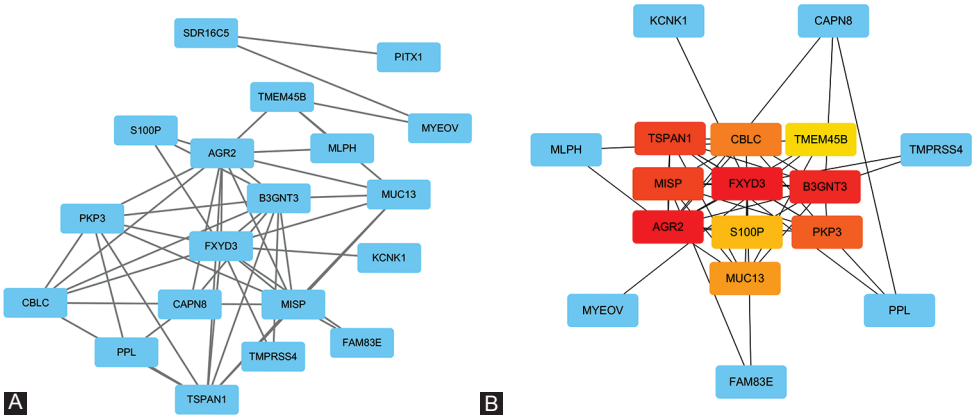

FIGURE 6. Construction of a PPI network and selection of the candidate core genes. (A) PPI network of intersecting genes from the Venn diagram. The genes are represented by the blue nodes. Edges indicate associations of interactions among nodes. (B) PPI network hub gene recognition using MCC. Genes with the highest MCC scores are red nodes; genes with the lowest MCC scores are yellow nodes.

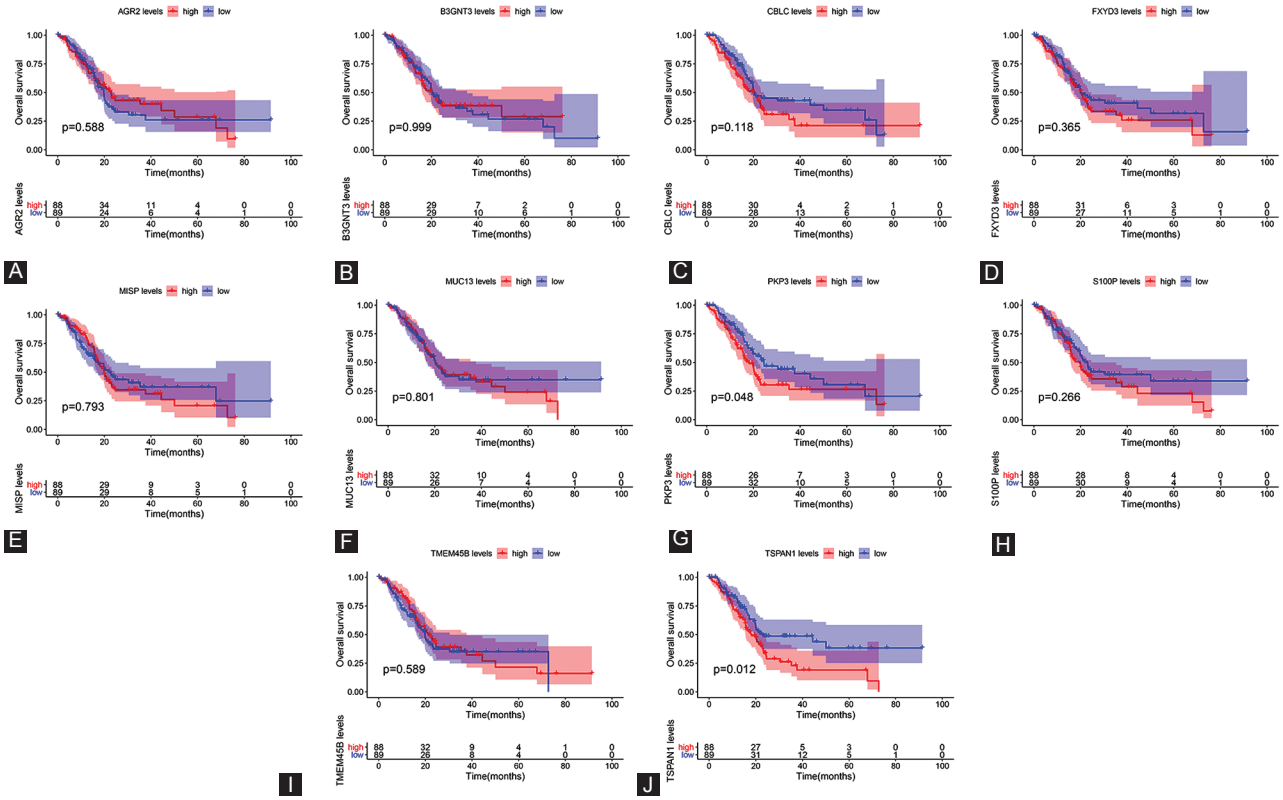

FIGURE 7. Correlation between the expression of the top 10 core genes and OS in PCC patients based on the GEPIA2 database. (A-J) ARG2, B3GNT3, CBLC, FXYD3, MISP, MUC13, PKP3, S100P, TMEM45B, and TSPAN1 in PCC survival analysis. $p < 0.05$ was defined as a statistically significant difference.

receptor tyrosine kinases [34]. Binding of EGF to EGFR induces binding to other ERBB homologous or heterologous dimers. EGFR then induces receptor phosphorylation and activation of downstream effect molecules, including, for example, the RAS–RAF–MEK–ERK MAPK and PI3K–AKT mTOR pathways, and finally leading to cell proliferation. Furthermore, classical cadherin, which is a cell surface glycoprotein, mediates calcium-dependent cell adhesion in a homotypic manner [35,36]. The adhesion regulation function of cadherins requires interaction between beta-catenin and the actin cytoskeleton. During invasion and metastasis of tumor cells, the conversion of the cadherin isoform from E-type cadherin to N-type cadherin is related to epithelial-to-mesenchymal transition [37,38]. In particular, changes in the expression of E-cadherin in the pancreas contribute to the development of human intraepithelial pancreatic neoplasia [39].

Our KEGG enrichment data further suggest that the 21 overlapping genes perform biological roles in ubiquinone and another terpenoid-quinone biosynthesis, glycosphingolipid

biosynthesis-lacto, and neolacto series, and retinol metabolism. Ubiquinone, also known as coenzyme Q, plays a central role in the mitochondrial electron transport chain, is involved in the production of mitochondrial oxidative phosphorylation and reactive oxygen species, and acts as a pivotal mediator of the pathogenesis of tumors [40]. Relevant studies have shown that ubiquinone can exert anti-tumor activity by promoting tumor cell proliferation and apoptosis [41], migration and invasion [42], and aerobic glycolysis [40]. Furthermore, work by Gehrmann et al. has demonstrated that the glycosphingolipid Gb3 facilitates tumor-specific Hsp70 plasma membrane localization [43]. Levels of HSP70 (a major stress-inducing member of the HSP70 family) on the plasma membrane have been considered as a prognostic indicator of OS in leukemia, lower rectal, and non-small cell lung carcinomas; however, it is unclear why tumors, but not healthy cells, present HSP70 on their cell surface, and the effect of the HSP70 membrane on cancer incidence remains to be clarified. Nevertheless, these results support a potential role for Gb3 in PCC prognosis.

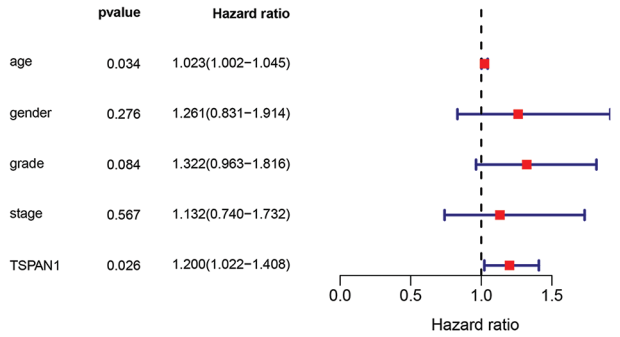

FIGURE 8. Multivariate COX regression analysis of TSPAN1 with other factors (age, gender, grade, and stage).

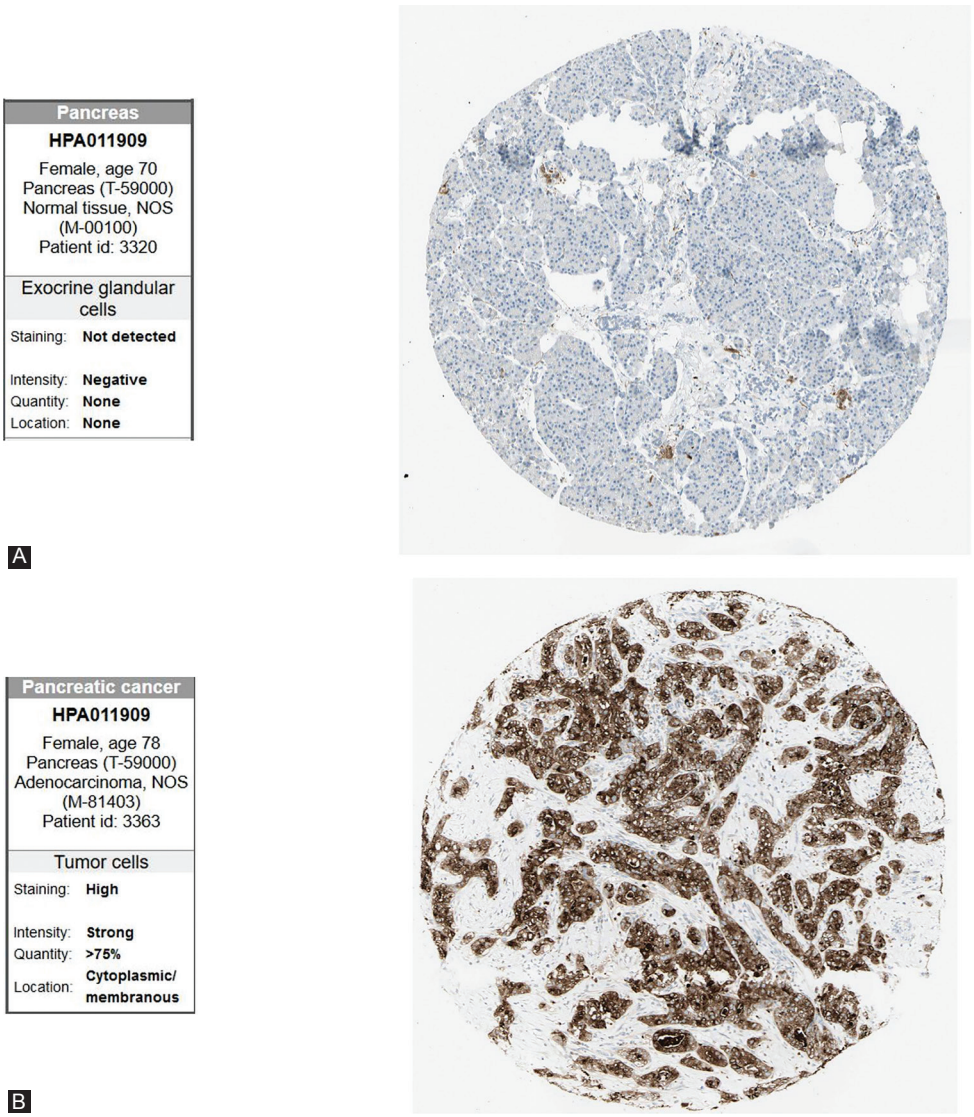

FIGURE 9. IHC of TSPAN1 in tumor tissues from the HPA database. (A) Protein levels of TSPAN1 in normal exocrine glandular cells tissues (antibody HPA011909; staining: Not detected; intensity: Negative; quantity: None). (B) Levels of TSPAN1 protein in PCC tissues (antibody HPA011909; staining: High; intensity: Strong; quantity: >75%).

We further identified 10 core genes (AGR2, FXYD3, B3GNT3, MISP, TSPAN1, PKP3, CBLC, MUC13, S100P, and TMEM45B) that were up-regulated in PCC tissues relative to healthy controls based on MCC evaluation results. Among them, increased expression of TSPAN1 was significantly associated with the poor OS rate of PCC. According to our COX analysis results, TSPAN1 also serves as an independent prognostic factor. TSPAN1 is a membrane glycoprotein and a member of a superfamily of transmembrane proteins (TM4SF) that have 4 members. While TM4SF has been studied only recently, its function in tumor invasion and metastasis has begun to be recognized. TSPAN1 has been shown to cause cancer cell proliferation and angiogenesis by switching cell division signals and inducing differentiation or dedifferentiation of cells [44]. According to prior research, TSPAN1 is widely expressed in gastric, lung, liver, and esophageal cancers [45-47]. As discussed above, beta-catenin is likely to have biological functions in the growth of PCC, and consistently, silencing of TSPAN1 has been shown to facilitate Smad2/3 phosphorylation and stabilize beta-catenin [48]. In addition, Hou et al. demonstrated that positive immunostaining of TSPAN1 is substantially associated with metastasis of the lymph node, TNM stage, and poor prognosis in PCC [49]. Of note, retinol metabolism and dissemination also play significant roles in the formation and evolution of tumors [50]. Our findings raise the possibility that TSPAN1 can interfere with the pathogenesis of PCC through the retinol metabolism pathway. Tian et al. demonstrated that the production of TSPAN1 in tumor tissues of PCC is dramatically higher than that of healthy tissues and that silencing of TSPAN1 reduces cell migration and invasion [51]. Furthermore, Wang et al. validated the oncogenic role of TSPAN1 in PCC, showing that TSPAN1 contributes to cell proliferation, migration, invasion, and tumorigenesis [52]. Zhang et al. also demonstrated that TSPAN1 up-regulates MMP2 through PLCγ to promote PCC cell migration and invasion [53]. Therefore, our results demonstrating that TSPAN1 is overexpressed in tumor tissues but not healthy tissues reveal a clear link with survival in PCC patients that are compatible with previous findings.

CONCLUSIONS

In summary, we identified gene co-expression modules and Hub genes linked to the progression and poor prognosis of PCC to guide further research, with potential benefits in the development of novel therapeutics. The present study does, however, have some limitations that should be acknowledged. First, although we did perform a thorough bioinformatics review to classify potential genes for diagnosis of PCC, the data may not be reliable for patients of every PCC subtype. Second, our research was constrained by the availability of experimental data. Confirmation using large-scale studies with subtype analysis may help to bring further insight into the role of TSPAN1 and other genes in PCC.

ACKNOWLEDGMENTS

The authors would like to thank Professor Zhang from Hebei General Hospital for their biostatistics expertise.

REFERENCES

[1] Duffy MJ, Sturgeon C, Lamerz R, Haglund C, Holubec VL, Klapdor R, et al. Tumor markers in pancreatic cancer: A European Group on Tumor Markers (EGTM) status report. Ann Oncol 2010;21(3):441-7. https://doi.org/10.1093/annonc/mdp332.

[2] Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer Res 2014;74(11):2913-21. https://doi.org/10.1158/0008-5472.can-14-0155.

[3] Feng RM, Zong YN, Cao SM, Xu RH. Current cancer situation in China: Good or bad news from the 2018 global cancer statistics? Cancer Commun (Lond) 2019;39(1):22. https://doi.org/10.1158/0008-5472.can-14-0155.

[4] Vincent A, Herman J, Schulick R, Hruban RH, Goggins M. Pancreatic cancer. Lancet 2011;378(9791):607-20. https://doi.org/10.1016/s0140-6736(10)62307-0.

[5] Williams DB, Sahai AV, Aabakken L, Penman ID, van Velse A, Webb J, et al. Endoscopic ultrasound guided fine needle aspiration biopsy: A large single centre experience. Gut 1999;44(5):720-6. https://doi.org/10.1136/gut.44.5.720.

[6] Venkat R, Edil BH, Schulick RD, Lidor AO, Makary MA, Wolfgang CL. Laparoscopic distal pancreatectomy is associated with significantly less overall morbidity compared to the open technique: A systematic review and meta-analysis. Ann Surg 2012;255(6):1048-59. https://doi.org/10.1097/sla.0b013e318251ee09.

[7] Ansari D, Tingstedt B, Andersson B, Holmquist F, Sturesson C, Williamsson C, et al. Pancreatic cancer: Yesterday, today and tomorrow. Future Oncol 2016;12(16):1929-46. https://doi.org/10.2217/fon-2016-0010.

[8] Nipp R, Tramontano AC, Kong CY, Pandharipande P, Dowling EC, Schrag D, et al. Disparities in cancer outcomes across age, sex, and race/ethnicity among patients with pancreatic cancer. Cancer Med 2018;7(2):525-35. https://doi.org/10.1002/cam4.1277.

[9] Srivastava P, Mangal M, Agarwal SM. Understanding the transcriptional regulation of cervix cancer using microarray gene expression data and promoter sequence analysis of a curated gene set. Gene 2014;535(2):233-8. https://doi.org/10.1016/j.gene.2013.11.028.

[10] Zhou Z, Cheng Y, Jiang Y, Liu S, Zhang M, Liu J, et al. Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. Int J Biol Sci 2018;14(2):124-36. https://doi.org/10.7150/ijbs.22619.

[11] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4:Article17.

[12] Li J, Zhou D, Qiu W, Shi Y, Yang JJ, Chen S, et al. Application of weighted gene co-expression network analysis for data from paired design. Sci Rep 2018;8(1):622. https://doi.org/10.1038/s41598-017-18705-z.

[13] Ni Y, Zhang Z, Chen G, Long W, Tong L, Zeng J. Integrated analyses identify potential prognostic markers for uveal melanoma. Exp Eye

Res 2019;187:107780.
https://doi.org/10.1016/j.exer.2019.107780.
[14] Liu DH, Wang SL, Hua Y, Shi GD, Qiao JH, Wei H. Five lncRNAs associated with the survival of hepatocellular carcinoma: A comprehensive study based on WGCNA and competing endogenous RNA network. Eur Rev Med Pharmacol Sci 2020;24(14):7621-33.
[15] Loven J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, et al. Revisiting global gene expression analysis. Cell 2012;151(3):476-82.
https://doi.org/10.1016/j.cell.2012.10.012.
[16] Pei H, Li L, Fridley BL, Jenkins GD, Kalari KR, Lingle W, et al. FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. Cancer Cell 2009;16(3):259-66.
https://doi.org/10.1016/j.ccr.2009.07.016.
[17] Robinson MD, McCarthy DJ, Smyth GK. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26(1):139-40.
https://doi.org/10.1093/bioinformatics/btp616.
[18] 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.
https://doi.org/10.1093/nar/gkv007.
[19] Ito K, Murphy D. Application of ggplot2 to pharmacometric graphics. CPT Pharmacometrics Syst Pharmacol 2013;2:e79.
https://doi.org/10.1038/psp.2013.56.
[20] Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.
https://doi.org/10.1186/1471-2105-9-559.
[21] Wang CC, Li CY, Cai JH, Sheu PC, Tsai JJ, Wu MY, et al. Identification of prognostic candidate genes in breast cancer by integrated bioinformatic analysis. J Clin Med 2019;8(8):1160.
https://doi.org/10.3390/jcm8081160.
[22] Chen H, Boutros PC. VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics 2011;12:35.
https://doi.org/10.1186/1471-2105-12-35.
[23] Yu G, Wang LG, Han Y, He QY. clusterProfiler: An R package for comparing biological themes among gene clusters. OMICS 2012;16(5):284-7.
https://doi.org/10.1089/omi.2011.0118.
[24] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47(D1):D607-13.
https://doi.org/10.1093/nar/gky1131.
[25] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498-504.
https://doi.org/10.1101/gr.1239303.
[26] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: Identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 2014;8 Suppl 4:S11.
https://doi.org/10.1186/1752-0509-8-s4-s11.
[27] Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: An enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res 2019;47(W1):W556-60.
https://doi.org/10.1093/nar/gkz430.
[28] Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science 2015;347(6220):1260419.
[29] Maity B, Sheff D, Fisher RA. Immunostaining: Detection of signaling protein location in tissues, cells and subcellular compartments. Methods Cell Biol 2013;113:81-105.
[30] Kleeff J, Korc M, Apte M, La Vecchia C, Johnson CD, Biankin AV, et al. Pancreatic cancer. Nat Rev Dis Primers 2016;2:16022.
https://doi.org/10.1038/nrdp.2016.22.
[31] Henriksen A, Dyhl-Polk A, Chen I, Nielsen D. Checkpoint inhibitors in pancreatic cancer. Cancer Treat Rev 2019;78:17-30.
https://doi.org/10.1016/j.ctrv.2019.06.005.
[32] Citri A, Yarden Y. EGF-ERBB signalling: Towards the systems level. Nat Rev Mol Cell Biol 2006;7(7):505-16.
https://doi.org/10.1038/nrm1962.
[33] Yamaoka T, Ohba M, Ohmori T. Molecular-targeted therapies for epidermal growth factor receptor and its resistance mechanisms. Int J Mol Sci 2017;18(11):2420.
https://doi.org/10.3390/ijms18112420.
[34] Cataldo VD, Gibbons DL, Perez-Soler R, Quintas-Cardama A. Treatment of non-small-cell lung cancer with erlotinib or gefitinib. N Engl J Med 2011;364(10):947-55.
https://doi.org/10.1056/nejmct0807960.
[35] Sotomayor M, Gaudet R, Corey DP. Sorting out a promiscuous superfamily: Towards cadherin connectomics. Trends Cell Biol 2014;24(9):524-36.
https://doi.org/10.1016/j.tcb.2014.03.007.
[36] Hoffman BD, Yap AS. Towards a dynamic understanding of cadherin-based mechanobiology. Trends Cell Biol 2015;25(12):803-14.
https://doi.org/10.1016/j.tcb.2015.09.008.
[37] Hotz B, Arndt M, Dullat S, Bhargava S, Buhr HJ, Hotz HG. Epithelial to mesenchymal transition: Expression of the regulators snail, slug, and twist in pancreatic cancer. Clin Cancer Res 2007;13(16):4769-76.
https://doi.org/10.1158/1078-0432.ccr-06-2926.
[38] Wheelock MJ, Shintani Y, Maeda M, Fukumoto Y, Johnson KR. Cadherin switching. J Cell Sci 2008;121(Pt 6):727-35.
[39] Al-Aynati MM, Radulovich N, Riddell RH, Tsao MS. Epithelial-cadherin and beta-catenin expression changes in pancreatic intraepithelial neoplasia. Clin Cancer Res 2004;10(4):1235-40.
https://doi.org/10.1158/1078-0432.ccr-03-0087.
[40] Suhane S, Kanzaki H, Arumugaswami V, Murali R, Ramanujan VK. Mitochondrial NDUFS3 regulates the ROS-mediated onset of metabolic switch in transformed cells. Biol Open 2013;2(3):295-305.
https://doi.org/10.1242/bio.20133244.
[41] He X, Zhou A, Lu H, Chen Y, Huang G, Yue X, et al. Suppression of mitochondrial complex I influences cell metastatic properties. PLoS One 2013;8(4):e61677.
https://doi.org/10.1371/journal.pone.0061677.
[42] Suhane S, Berel D, Ramanujan VK. Biomarker signatures of mitochondrial NDUFS3 in invasive breast carcinoma. Biochem Biophys Res Commun 2011;412(4):590-5.
https://doi.org/10.1016/j.bbrc.2011.08.003.
[43] Gehrmann M, Liebisch G, Schmitz G, Anderson R, Steinem C, De Maio A, et al. Tumor-specific Hsp70 plasma membrane localization is enabled by the glycosphingolipid Gb3. PLoS One 2008;3(4):e1925.
https://doi.org/10.1371/journal.pone.0001925.
[44] Serru V, Dessen P, Boucheix C, Rubinstein E. Sequence and expression of seven new tetraspans. Biochim Biophys Acta 2000;1478(1):159-63.
https://doi.org/10.1016/s0167-4838(00)00022-4.
[45] Garcia-Espana A, Chung PJ, Sarkar IN, Stiner E, Sun TT, Desalle R. Appearance of new tetraspanin genes during vertebrate evolution. Genomics 2008;91(4):326-34.
https://doi.org/10.1016/j.ygeno.2007.12.005.
[46] Tarrant JM, Robb L, van Spriel AB, Wright MD. Tetraspanins: Molecular organisers of the leukocyte surface. Trends Immunol 2003;24(11):610-7.
https://doi.org/10.1016/j.it.2003.09.011.
[47] Haining EJ, Yang J, Bailey RL, Khan K, Collier R, Tsai S, et al. The TspanC8 subgroup of tetraspanins interacts with A disintegrin and metalloprotease 10 (ADAM10) and regulates its maturation and cell surface expression. J Biol Chem 2012;287(47):39753-65.
https://doi.org/10.1074/jbc.m112.416503.
[48] Zhao T, Jiang W, Wang X, Wang H, Zheng C, Li Y, et al. ESE3 Inhibits pancreatic cancer metastasis by up-regulating E-cadherin. Cancer Res 2017;77(4):874-85.
https://doi.org/10.1158/0008-5472.can-16-2170.
[49] Hou FQ, Lei XF, Yao JL, Wang YJ, Zhang W. Tetraspanin 1 is involved in survival, proliferation and carcinogenesis of pancreatic cancer. Oncol Rep 2015;34(6):3068-76.
https://doi.org/10.3892/or.2015.4272.
[50] Das BC, Thapa P, Karki R, Das S, Mahapatra S, Liu TC, et al. Retinoic acid signaling pathways in development and diseases. Bioorg Med

Chem 2014;22(2):673-83.
[51] Tian J, Zhang R, Piao H, Li X, Sheng W, Zhou J, et al. Silencing Tspan1 inhibits migration and invasion, and induces the apoptosis of human pancreatic cancer cells. Mol Med Rep 2018;18(3):3280-8. https://doi.org/10.3892/mmr.2018.9331.
[52] Wang S, Liu X, Khan AA, Li H, Tahir M, Yan X, et al. miR-216a-mediated upregulation of TSPAN1 contributes to pancreatic cancer progression via transcriptional regulation of ITGA2. Am J Cancer Res 2020;10(4):1115-29.
[53] Zhang X, Shi G, Gao F, Liu P, Wang H, Tan X. TSPAN1 up-regulates MMP2 to promote pancreatic cancer cell migration and invasion via PLCgamma. Oncol Rep 2019;41(4):2117-25. https://doi.org/10.3892/or.2019.6989.

SUPPLEMENTARY

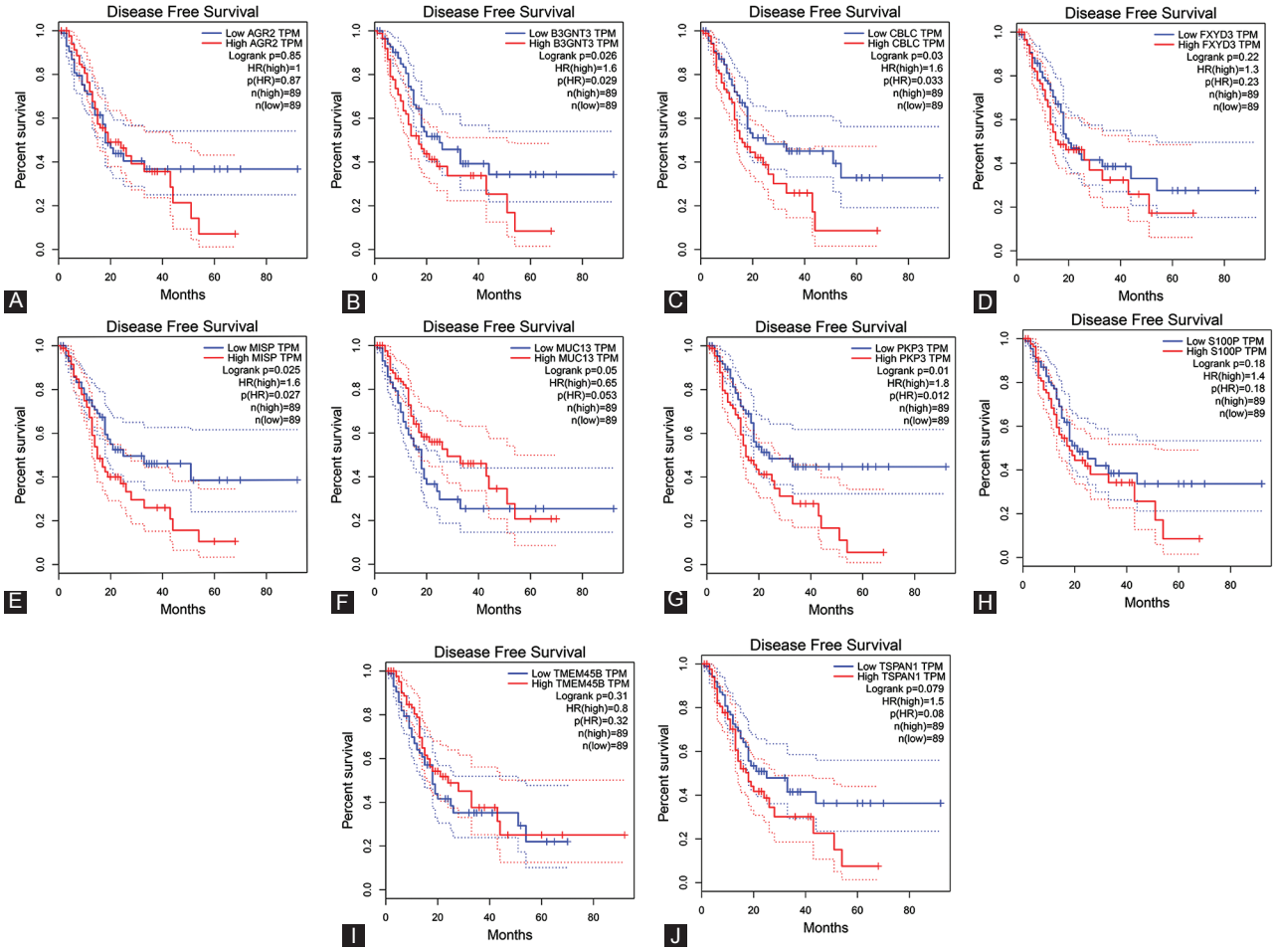

FIGURE S1. Correlation between the expression of the 10 core genes and DFS in PCC patients based on the GEPIA2 database. (A-J) Survival analysis of ARG2, B3GNT3, CBLC, FXYD3, MISP, MUC13, PKP3, S100P, TMEM45B, and TSPAN1 in PCC. Log-rank $p < 0.05$ was defined as a statistically significant difference.

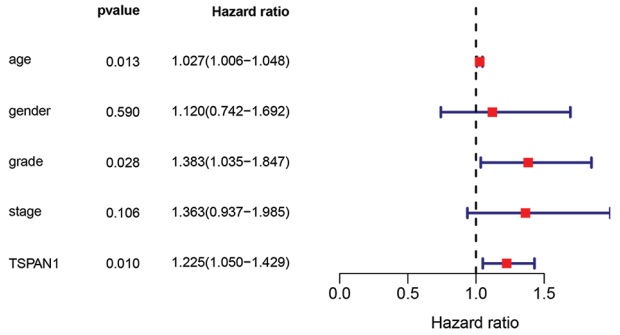

FIGURE S2. Univariate Cox regression analysis of TSPAN1 with other factors (age, gender, grade, and stage) in PCC patients.

SUPPLEMENTARY TABLE 1. The number of genes in the co-expression modules from various databases.

Module color	Frequency	Databases
Black	321	TCGA
Blue	5153	TCGA
Brown	1077	TCGA
Green	1368	TCGA
Red	367	TCGA
Magenta	5809	TCGA
Pink	224	TCGA
Purple	143	TCGA
Salmon	62	TCGA
Tan	82	TCGA
Black	334	GEO
Cyan	147	GEO
Green	2943	GEO
Green-yellow	555	GEO
Grey	269	GEO
Grey60	107	GEO
Light green	85	GEO
Magenta	1354	GEO
Midnight blue	144	GEO
Pink	440	GEO
Purple	223	GEO
Tan	178	GEO
Yellow	540	GEO

SUPPLEMENTARY TABLE 2. Univariate and multivariate COX analyses of TSPAN1 expression with other factors in PCC patients.

Parameter	Univariate analysis			Multivariate analysis		
	HR	95% CI	*p*	HR	95% CI	*p*
Age	1.03	1.01-1.05	0.013*	1.02	1.00-1.04	0.033*
Gender	1.12	0.74-1.69	0.590	1.26	0.83-1.91	0.276
Grade	1.38	1.04-1.85	0.028	1.32	0.96-1.81	0.084
Stage	1.36	0.94-1.98	0.106	1.13	0.74-1.73	0.567
TSPAN1	1.23	1.05-1.43	0.010*	1.20	1.02-1.41	0.025*

HR: Hazard ratio; CI: Confidence interval; *$p<0.05$