Identification of a stemness-related prognostic model in Pancreatic ductal adenocarcinoma by Weighted gene co-expression network analysis

Xiao-Yan Huang¹; Wen-Tao Qin²; Qi-Sheng Su³; Cheng-Cheng Qiu¹, Ruo-Chuan Liu¹, Shan-Shan Xie¹, Yang Hu², Shang-Yong Zhu¹*

¹Department of Medical Ultrasound, First Affiliated Hospital of Guangxi Medical University, 6 Shuangyong Rd, 530021 Nanning, Guangxi, China.
²Department of Bone and Joint Surgery, First Affiliated Hospital of Guangxi Medical University, 6 Shuangyong Rd, 530021 Nanning, Guangxi, China.
³Department of Clinical Laboratory, First Affiliated Hospital of Guangxi Medical University, 6 Shuangyong Rd, 530021 Nanning, Guangxi, China.

*Correspondence author:
Shang-Yong Zhu
6 Shuangyong Rd, 530021 Nanning, Guangxi, China.
E-mail: zhushangyongsound@163.com

Abstract

Objective: Aim of this study was to identify the stemness-related genes in Pancreatic ductal adenocarcinoma (PDAC).

Methods: The RNA-seq data of PADC were downloaded from The Cancer Genome Atlas (TCGA) and the Genotype-Tissue Expression (GTex) databases. mRNA expression base-index (mRNAsi) and epigenetically regulated mRNAsi (EREG-mRNAsi) of PADC were evaluated. The mRNAsi-based gene sets in PADC were identified by Weighted gene co-expression network analysis (WGCNA). Functional Enrichment Analyses were performed with key genes. Kaplan–Meier survival analysis and the Cox proportional hazards model were used to evaluate the prognostic value of the key genes. Prognosis-associated hub genes were applied to establish nomograms. The receiver operating characteristic curves (ROC) and
concordance index (C-index) were utilized to assess the discrimination and accuracy of the nomogram. Finally, these results were validated in the Gene Expression Omnibus (GEO) database and immunohistochemistry (IHC).

**Results:** 36 key genes associate with mRNAsi were screened via WGCNA. Next a five-gene signature compromised TPX2, NCAPH, UBE2C, CCNB2, CEP55. Based on the expression of the signature, the PADC patients were classify patients into high- and low-risk groups. Cox regression analysis revealed that the high-risk group was significantly positive with overall survival (OS). Moreover, the nomogram has better sensitivity and specificity for predicting the OS. And the ROC, C-index indicated good performance of the prognostic signature in the TCGA and GEO dataset.

**Conclusion:** Prognostic model associated with cancer stem cells (CSCs) reliably predict OS in PADC. this might be beneficial for the treatment of PADC patients.

**Keywords:** Pancreatic ductal adenocarcinoma; mRNAsi; stem cell characteristics; WGCNA; prognostic model

**Introduction**

Pancreatic ductal adenocarcinoma (PDAC) is the most prevalent type of pancreatic neoplasm, accounting for more than 90% of the total number of pancreatic tumors. It has a 5-year survival rate of less than 10% [1,2]. Surgical resection represents the only chance for cure, and advancements in adjuvant chemotherapy have improved long-term outcomes in PDAC patients[3]. Lack of early diagnosis and effective interventions are the major factors in PADC. Badly, these treatments cause great trauma to the patient, thus reducing the patient's quality of life to some extent. Fortunately, molecular targeted therapy, which blocks cancer growth, progression and metastasis by specific molecular biomarkers, emerges as an individual treatment strategy because of less trauma and excellent efficacy[4]. However, identifying molecular biomarkers remains a challenge due to the insufficient knowledge of PDAC.

Cancer stem cells (CSCs) were found to promote tumor recurrence, metastasis and
drug resistance because of their self-renewal ability, proliferation and multi-lineage differentiation potential[5]. Besides, emerging evidence indicated that stemness of CSCs is associated with high intratumoral heterogeneity in most cancers, including PADC[6, 7]. Intratumoral heterogeneity reportedly drives disease progression and causes treatment failure. And it includes phenotypic diversity such as cell surface markers and (epi)genetic abnormality. This markers as well as gene mutation types are often considered as basis for the pathological classification and clinical treatment of tumor[8]. Pancreatic CSCs, first described in 2007[9], account for less than 1% of all pancreatic cancer cells[10], and are responsible for PDAC tumor development, metastasis, and chemoresistance. As previously reported, PADC involves the activation of biomarkers and signaling pathways related to CSCs, including CD133, CD24, CD44, MYC, WNT/β-catenin, Notch, etc. [11]. Although the use of these markers has increased our potential biological knowledge of Pancreatic CSCs, such as unlimited self-renewal capacity, unique tumorigenicity, and chemoresistance, more research is needed on the molecular mechanisms of Pancreatic CSCs. Our study focuses on uncovering heterogeneity of PADC from the perspective of stemness features of CSCs below.

Currently, the stemness features of CSCs have been mainly characterized mRNA expression-based-index (mRNAsi) (eg. Epigenetic regulation based-index (EREG-mRNAsi))[5]. The high score of mRNAsi in CSCs helps to identify new targets and reflects the high degree of carcinogenic dedifferentiation[12]. Studies have reported that the scores of mRNAsi associated with CSCs stemness could be computed by the one-class logistic regression (OCLR) machine learning algorithm, and that there was significant correlation between mRNAsi obtained and prognosis of PADC[5]. However, genes related to CSCs stemness have not been identified throughoutly, or the biological function of genes related to CSCs stemness remains unknown. Analysis of differentially expressed genes (DEGs) has been reported to identify genes. However, such analysis was found to show limitations in elaborating the connections between genes. The gene network analysis can reflect the complex
biological processes during tumor neogenesis[8]. Weighted gene co-expression network analysis (WGCNA) was used to explore co-expression modules associate with clinical traits, which include mRNAsi[13, 14].

In current studies, WGCNA was used to identify stemness-related preserved modules for discovering the key genes. Next, the functions of these genes may be associated with cell cycle regulation during the PADC development by Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses and gene set enrichment analysis (GSEA). Subsequently, a novel seven-gene prognostic model (i.e. TPX2, NCAPH, UBE2C, CCNB2, CEP55) was constructed and verified and assessed in TCGA and GEO database. Finally, prognosis-associated genes were applied to established predicted nomograms. Overall, our new model and nomogram might provide a powerful tool to predict PADC prognosis.

**Material and methods**

**Datasets sources and processing**

In this work, the criteria for screening patients in the TCGA PADC cohort include the following: (1) histologically confirmed primary PADC (2) patients with RNA-seq data, (3) patients with detailed clinicopathological [age, gender, TNM stage, grade, OS, et.al.]. The corresponding RNA-seq data and clinical data of PADC were collected from TCGA data in the GDC database(https://portal.gdc.cancer.gov). expression of genes analyzed in normal tissues was collected using GTEx. And these data were all available on September 9, 2020. The RNA-seq data included 169 normal samples and 142 tumor samples, then were integrated into a matrix file using a merge script in Perl language (http://www.perl.org/).

**mRNAsi in PADC and its clinical significance**

The mRNAsi used to evaluate oncogenic dedifferentiation was from previous studies[5]. The mRNAsi score of PADC samples can be calculated from PADC datasets inside TCGA by OCLR. The gene expression-based stemness index ranges from low (zero) to high (one) stemness. Significant differences of mRNAsi in tumors
and non-tumors were analyzed by Wilcoxon-test. Next, to evaluate the prognostic value of mRNAsi scores, an overall survival analysis according to mRNAsi scores was performed using the "survival" package in R software v3.6.1 (https://CRAN.R-project.org/package=survival). The Kaplan-Meier (K-M) analysis for samples with the high and low mRNAsi set was carried out. Furthermore, we used the Wilcoxon-test to investigate the correlation between mRNAsi score and patient age.

**Differentially expressed genes (DEGs) in UCEC**

Wilcoxon-test was used to screen the DEGs between normal pancreases samples and tumor samples. The false discovery rate (FDR) <0.05 and |log2 fold change (FC)|≥1, were considered the cut-off criteria for screening.

**WGCNA construction and module preservation**

An R package "WGCNA" was utilized to discover the correlations among genes by constructing important modules. And the modules were obtained by clustering genes with similar expression patterns, reflecting external sample characteristics[14]. 1750 DEGs with high precision and accuracy were selected for the next step of WGCNA analysis. To obtain a high-scale independence and average connectivity, the soft-thresholding power value was calculated from a gradient test range from 1 to 20 and determined by pickSoftThreshold function. Then, we calculated the topological overlap matrix (TOM) according to the corresponding soft-thresholding power. The TOM was introduced to identify modules of highly co-expressed genes for making networks less sensitive to spurious connections. Genes with high absolute correlations were further clustered into the same modules. The modules were defined by cutting the clustering tree into branches by using a dynamic tree cutting algorithm and assigned to different colors for visualization. Through the analysis of hierarchical clustering based on TOM-based dissimilarity measurements, we constructed module dendrograms; To avoid over-splitting, correlated modules (r < 0.25) were then merged and each module was added a corresponding color[15].

**Identification of Modules Associated with Clinical characteristics**
Module eigengenes (MEs) are considered to be the main components of principal component analysis of each gene module, and the expression patterns of all genes can be summarized as a single feature expression profile within a given module[16]. To show the correlation between genes and sample characteristics, the gene significance (GS) was calculated. And module significance (MS) refers to the average GS of all genes in the module. MS was used to measure the correlation between modules and sample characteristics. The module with higher correlation with MS were identified in the further analysis.

**Screening of the key Genes in red module**

The key genes in co-expressed network were highly interconnected with nodes in a module, and they were used to explore the biological functions of the identified dysregulated genes[17]. The module membership (MM) was defined as the correlation of gene expression profile with module eigengenes (MEs). The intramodular key genes were defined as cor.GS (gene significance) > 0.5 and cor. gene MM > 0.8. And then the "corrplot" R package was applied to describe the Pearson’s correlation among the key genes.

**Function and pathway enrichment analysis**

To investigate the biological function of key genes, GO and KEGG analyses were employed[18, 19]. The "clusterProfiler" R package used to implement this functional annotation and pathway enrichment analysis. FDR ≤ 0.05 were considered to determine statistical significance. The protein-protein interaction (PPI) network of the key genes in key modules was constructed using the STRING (https://www.string-db.org/)[20]. The criterion for selecting hub nodes was a combined score of ≥0.8 and a connectivity degree of ≥20.

**Establishment of a prognostic model**

We used K-M survival analysis curves and the stratified Cox proportional hazards analysis to evaluate the prognostic value of the key genes. Then risk score for each prognostic gene was calculated based on the sum of products of Si (expression level
of key genes) and β (corresponding coefficients) generated from Cox model, respectively. And the PADC patients were divided into high- and low-risk subgroups by median risk score for fitting K-M analysis to further validate whether they went through diametrically distinct prognosis. Predictive accuracy of each prognostic signature was accessed by calculating Uno’s inverse-probability of censoring weighting estimation of dynamic time-dependent receiver-operator characteristic (ROC) area under the curve (AUC) values (time spanning from 1 to 3 years) with "timeROC" package (version 0.3), according to the method proposed by Blanche et al[21]. Finally, the expression distribution of the signature genes was visualized in the heatmap using the "pheatmap" R package.

**Prognostic value of the five-gene prognostic model**

The risk scores and the clinicopathological parameters, including age, gender, TNM stage, grade, OS, postoperative radiotherapy (postoperative tx tx), radiotherapy, alcohol story, were included in the univariate and multivariate Cox model. Next, based on the result of multivariable Cox regression analysis, a nomogram was established using the "rms" package (version 5.1.2)[22]. The nanogram was applied to predict 1-, 2, 3-years OS of PADC patients in the TCGA dataset. Subsequently, using the "timeROC" package of R, the time-dependent ROC curve was utilized to assess the sensibility and specificity of the nomogram, and these was quantified by AUC. The predictive accuracy discrimination of the nomogram was assessed by the Harrell's C-index. The C-index was calculated to assess nomogram discrimination based on a bootstrap method with 1,000 resamples.

**Validation of the prognostic signature**

To minimize bias caused by small study cohorts, the prognostic capability of five-gene model was validated using dataset GSE62452 the GEO database (www.ncbi.nlm.nih.gov/geo/). The optimal cut-off for risk score for the GEO dataset was the same as for the primary TCGA cohort. K-M analysis was used to plot survival curves for the low-risk and high-risk groups. The predictive accuracy of the prognostic model was investigated using ROC curves, and the performance of the
diagnostic nomogram was assessed by time-dependent ROC curves. The Human Protein Atlas (HPA) (https://www.proteinatlas.org/) were applied to validated the expression of the real hub genes.

**Gene set enrichment analysis (GSEA)**

To identify potential function of the five-gene model, GSEA (http://software.broadinstitute.org/gsea/index.jsp) was performed with GSEA v4.0.3 software. GSEA can detect whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states [23]. Based on the molecular signature database v. 6.2, C2 (curated gene sets), C5 (GO gene sets), and C6 (oncogenic signatures) were searched to identify KEGG pathways, biological processes (BP), cellular components (CC), molecular functions (MF) and dysregulated oncogenic signatures. The samples in TCGA were divided into two groups (high-risk score group vs. low-risk score group) with median expression values of each genes. Enriched gene sets with a nominal P-value <0.05, |NES|> 1 and FDR q-value <0.25 were considered statistically significant.

**Statistical analysis**

Statistical analysis was performed using R software v3.6.1. For K-M curves, p-values and hazard ratio (HR) with 95% confidence interval (CI) were generated by log-rank tests and univariate Cox proportional hazards regression. Pearson’s chi-square tests were used to detect the variables difference. And, results of multivariate Cox regression analyses were visualized with nomogram. C-index and time-dependent ROC were also important indicators used to assess the nomogram. P value < 0.05 was considered statistically significant.

**Results**

**Data processing and survival analysis**

RNA-seq data related to CSCs were acquired from the TCGA PADC cohorts, including 142 PADC samples. The 169 normal samples were collected from TCGA
and GTEx normal tissues. Gene expression-based stemness indices for PADC were extracted by OCLR[5]. mRNAsi is an indicator used to quantify the degree of similarity between tumor cells and stem cells. mRNAsi is based on the gene expression. EGER-mRNAsi is based on training stem cell epigenetic regulation of the expression of related genes. As shown in Figure 1A, Significant differences were found for age. And survival analysis showed that the expression level of mRNAsi was significantly different from the survival time of PADC (Figure 1B) (adjusted P value < 0.017). It further suggested that the stemness features of CSCs was related to the survival rate of PADC patients.

**DEGs identification in UCEC**

According to cut-off criteria, by Wilcoxon-test, 1750 DEGs were screened, including 845 down-regulated genes and 905 up-regulated genes. We visualized the expression of DEGs using the "beeswarm" R package, and the results showed that DEGs were significant differentially expressed in PADC and normal groups with a clear display of volcano plot and heat map (Figure 1C).

**14 co-expression modules were identified by WGCNA**

Based on the DEGs, the WGCNA package was used to identify the highly stemness-related modules. In order to ensure accuracy of the results, β = 11 (scale-free R² = 0.95) as a soft threshold (power) was used to ensure a scale-free network (Figure 2A). and then we established a sample dendrogram and trait heatmap of mRNAsi and EGER-mRNAsi (Figure 2B). Based on this cluster analysis of PADC, there were a total of 14 different modules (module size ≥ 50, cut height ≥ 0.25) in the network (blue, tan, brown, greenyellow, cyan, green, magenta, red, purple, yellow, black, turquoise, grey) (Figure 2C). Three modules were identified as related to mRNAsi. The red module showed the highest positive correlation with mRNAsi (Figure 2C, r=0.56, p=2e-12). The blue module exhibited most significantly negative associations with mRNAsi (Figure 3C, r=-0.78, p=4e-28). Thus, the red module was selected to further explored the hub genes (Figure 2D).
Analysis of key gene in red module

Genes in the same module exhibit common gene expression patterns. And in this work, we obtained 36 key genes out of 1750 DEGs that were the relevant to mRNAsi in the red module. We also compared the differential expression of genes with minimum p-value (P<0.01) in the PADC and normal tissues, and found that these genes were upregulated in tumor tissues (Figure 3A). Meanwhile, in order to explicit the difference clearly, we drew a heat map by "ggpubr" R package (Figure 3B). Moreover, the interrelationships among key genes were visualized as a heatmap (Figure 3C, p<0.01).

Functional Enrichment and PPI Network Analysis of the key genes

To fully understand biological functions of 36 key genes in PADC, GO and KEGG function and enrichment analyses were performed. The functions of these genes are mainly enriched to 30 pathways, including BP, CC, and MF (Figure 4A). BP was found to be relevant to mitotic nuclear division. CC was related to the spindle. MF includes microtubule binding. As shown in Figure 4B, KEGG pathway analysis indicated that the hub genes were enriched in cell-cycle signaling pathway. The signaling pathway with FDR corrected p-values < 0.01 was deemed significant. Furthermore, a PPI network of key genes was constructed to identify gene interactions (Figure 4C).

Identification and evaluation of stemness-related prognostic model

To find out the prognostic value of stemness-feature associated genes, a total of 36 key genes were identified to be associated with OS based on univariate Cox regression analysis, and the 35 genes related OS were identified as risk factors (Figure 5A). Next, a prognostic model composed of 5 genes was then developed based on the red module using the stepwise multivariate Cox regression analysis. The 5 genes contained TPX2, NCAPH, UBE2C, CCNB2, CEP55. Therefore, the 5 gene-based risk score was constructed based on their Cox coefficients: risk score = 1.268615*Exp(TPX2) − 1.29279*Exp(NCAPH) − 0.78723*Exp(UBE2C) + 0.734581
*Exp(CCNB2) + 0.60947*Exp(CEP55). The risk score of patients were divided into high-risk group (risk score ≥ 1.103) and low-risk group (risk score < 1.103). And the survival status of all PADC patients were showed in Figure 5B. the heatmap of 5 prognostic genes were also showed in Figure 5B. The K-M curves revealed significantly worse prognosis in the high-risk group (P<0.00017) (Figure 6A). The 1-, 2-, 3-years-AUC of the time-dependent ROC were 0.712, 0.7, 0.741, respectively (Figure 6B), confirming the satisfactory prediction efficiency of the five-gene model in OS. The C-index of the prognostic model was 0.706 (95% CI; 0.631–0.781).

**Building and validation of a prognostic nomogram**

A prognostic nomogram for predicting OS of PADC patients was developed based on 311 patients with clinical information from TCGA database. The parameters of the nomogram include risk score, gender, tumor size, TNM stage, grade, postoperative_tx_tx, radiotherapy, alcohol story. The forest plot of the multivariable Cox regression analysis, the forest plot of the univariable Cox regression analysis, based on 8 clinicopathologic characteristics, were applied to evaluate the independent prognostic force of the signature (Figure 7A, B). Results from univariate Cox regression analysis showed that risk group (HR = 1.510, p < 0.001), postoperative_rx_tx (HR = 5.600 p < 0.001), radiation therapy (HR = 2.353, p < 0.021) identified as the protective factors. Next, risk group (HR = 1.333, p < 0.003), tumor size (HR = 1.338, p < 0.036), postoperative_rx_tx (HR = 4.938, p < 0.001) were statistically significant in multivariable Cox regression analysis.

**Validation of the five-gene prognostic model**

We evaluated the predictive power of the prognostic gene signature in the PADC cohorts from the GEO database (GSE62452). Figure 6C showed that the high-risk cohort also was a significantly poorer prognosis than the low-risk cohort (P<0.0043). And the time-dependent ROC showed that AUC for 1-, 2-, 3-years were 0.551, 0.821, 0.825 (Figure 6D). The C-index in the external validation set was 0.6806(95% CI, 0.619-0.741).
The protein expression levels of these hub genes were obtained from HPA database. Notably, the proteins of CCNB2, TPX2, NCAPH2 and CEP55 were not or low expressed in normal pancreas tissues, whereas medium and high expression levels of these genes in PADC tissues (Figure 8). This may indicate that this five-gene prognostic model has some reference value for the diagnosis or treatment of patients with PADC.

**Gene set enrichment analysis (GSEA)**

In order to further explore the potential functions on the 5 genes in PADC, GSEA was applied. These results suggest that these genes are mainly associated with cell cycle checkpoint, mitotic nuclear division, DNA integrity checkpoint (Figure 9).

**Discussion**

PDAC is a highly heterogeneous malignancy, with high morbidity and mortality. What’s worse, stem cells which show similar to cancer cells in some bioactivities are found to reside in some cancer tissues. And, there is evidence that stem cells and tumor cells are similar in the following three aspects: 1. The regulatory mechanism of self-renewal is similar; 2. Tumor cells may be derived from normal stem cells; 3. Tumor tissues may contain a small proportion of CSCs with unlimited proliferative potential[24-26]. Several years after the conceptually proposed existence of stem-like cancer cells, experimental evidence was first provided in leukemia models, confirming that CD34+CD38–eukemic cells show bone marrow hematopoietic stem cell characteristics[27, 28]. Solid tumor CSCs (CD44+ CD24–/lowLineage– cells) have been found in breast cancer, ovarian, prostate, colon, pancreatic, liver, and lung cancers. Therefore, CD34+, CD44+ have been suggested for common or unique CSC signatures[29, 30]. In addition, CSCs were identified in PDAC using some cell surface makers, including CD133, CD44, CD24, ESA/EpCAM (epithelial-specific antigen), etc.[31]. And it is now universally accepted that PDAC metastasis, chemoresistance, and disease relapse are driven by a subpopulation of highly plastic “stem-like” cells known as CSCs in tumors, and from a clinical perspective, understanding how these cells function and developing inhibitors to eliminate the
CSCs population should lead to tumor eradication [32-34]. However, the molecular mechanisms driving the stem-like state of pancreatic CSCs is still incomplete and requires further investigation.

The existence of CSCs in tumor tissues and the stem-like characteristics of CSCs suggest the feasibility to treat tumors by targeting stemness characteristics of CSCs. Tathiane et al. have shown that stemness index of CSCs can evaluate stemness characteristics, including mRNAsi and EGER-mRNAsi [5]. Although risk stratification by stemness index has been studied in pan-cancer populations, the comprehensive prognostic value of the stemness index for PADC has not been fully understood. Moreover, studies on genes related stemness characteristics of CSCs in PADC are sparse, and in this study aim to identify the stemness-related prognostic signature model.

In this study, significant differences were found in survival between low- and high mRNAsi (mRNAsi/tumor purity) score groups in the K-M curves. Meanwhile, stemness indices-related modules and genes were identified using WGCNA. And we selected the red module with the most stemness characteristics, in which we obtained 36 genes related to stemness of CSCs. Next, the GO, KEGG and GSEA analyses showed that these genes were enriched in mitotic nuclear division, nuclear chromosome segregation, spindle checkpoint function, microtubule binding and cell cycle. The results suggested these genes could have a role in enhancing the self-renewal and proliferative characteristics. Furthermore, 5 key mRNAsi-related genes model (i.e. TPX2, NCAPH, UBE2C, CCNB2, CEP55) was constructed by univariate Cox regression and multivariate Cox regression analyses. Then, we successfully verified that survival model accurately predicts the prognosis of PADC patients in the TCGA database and GEO database, respectively. Evaluation of the model for OS in the PADC by ROC curve displayed better predictive power. Furthermore, some prognostic parameters (tumor size, risk score, radiotherapy) were significantly correlated with OS for PADC patients. Overall, these results indicated that the 5 genes might be related to the occurrence, development of PADC
by regulating cell cycle, and were consistent with the view of Bai et al.[35]. Thus, new prospective studies are necessary to further verify prognostic value of the stemness index-associated risk signature PADC patients.

In agreement with previous reports[36-38], dysregulation of the 5 key genes that we identified were found to result in the development of tumors. In addition, these hub genes were found to be closely associated with cell cycle. These 5 genes functioned to disrupt mitotic cyclins and cell cycle progression and might be involved in cancer progression. TPX2 was found to encode a microtubule-associated protein. Up-regulation TPX2 levels led to increased proliferation and invasion of renal cancer cells[39]. CEP55 were reported to play an important role in promoting self-renewal and stemness maintains of glioma stem like cells[40]. Dysregulation of UBE2C was associated with proliferative marker Ki-67 and a poor overall survival in colorectal carcinoma[41, 42]. In advanced prostate cancer, Aaron et al showed that the abnormal expression of CCNB2 was closely related to cell cycle-driven subpopulation[43]. And NCAPH may regulate cell cycle progression and DNA damage via activation of the DNA damage response (Chk1/Chk2) signaling pathways in pancreatic cancer [44]. In PADC, there are few studies on these 5 genes related to stemness characteristics of CSCs. In the study, we identified the key genes as having functions in regulating cell division cycle in PADC. These findings suggest that these genes may promote tumor initiation, metastasis, as well as recurrence. Therefore, the study of the genes related to the stemness characteristics of CSCs provides a new idea for the targeted therapy of PADC in the future. Future studies should focus on the role of targeting the key genes to direct adjuvant treatment of PADC and further exploration about molecular mechanism that sensitive to tailored targeted therapies.

**Conclusion**

In summary, we established a stemness index-related signature and developed a prognostic nomogram in combination with prognosis-related clinical parameters. And the model had predictive value for predicting the OS of patients with PADC.
Mechanically, TPX2, ZWINT, UBE2C, CCNB2, CDK1, BUB1, BIRC5 could orchestrate the stemness, cell proliferation and cell invasion. Therefore, these genes might be candidate prognostic biomarkers and potential therapeutic targets in PADC.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

In this study, the privacy of individual participants was not involved, such as images, videos or other details. All authors have contributed to, read, and approved this submitted manuscript in its current form.

**Availability of Data and materials**

All data generated or analyzed during this study are included in the manuscript.

**Abbreviations**

| Abbreviation | Description |
|--------------|-------------|
| PADC         | Pancreatic ductal adenocarcinoma |
| mRNAsi       | mRNAsi expression base-index |
| EREG-mRNAsi  | epigenetic regulation based-index |
| DEGs         | differentially expressed genes |
| WGCNA        | Weighted gene co-expression network analysis |
| CSCs         | Cancer stem cells |
| FDR          | false discovery rate |
| TOM          | topological overlap measure |
| GS           | gene significance |
| MS           | Module significance |
| ME           | Module Eigengenes |
| MM           | Module Membership |
| GO           | Gene Ontology analysis |
| KEGG         | Kyoto Encyclopedia of Genes and Genomes |
Analyses

| Acronym | Description                             |
|---------|-----------------------------------------|
| GSEA    | Gene set enrichment analysis            |
| PPI     | protein-protein interaction             |
| MF      | Molecular Function                      |
| BP      | biological process                      |
| CC      | cellular component                      |
| OS      | overall survival                        |
| ROC     | receiver operating characteristic curves|
| C-index | concordance index                       |

Authors' contributions

Xiao-Yan Huang conceived and designed experiments, analyzed the data, prepare the figures, authored or reviewed drafts of the manuscript, approved the final draft.

Wen-Tao Qin conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, approved the final draft.

Qi-Sheng Su contributed reagents/materials/analysis tools, approved the final draft.

Ruo-Chuan Liu conceived and designed the experiments, approved the final draft.

Cheng-Cheng Qiu, Shan-Shan Xie and Yang Hu performed the data curation, approved the final draft.

*Shang-Yong Zhu conceived and designed the experiments, contributed reagents/materials/analysis tools, approved the final draft.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare that they have no competing interests in this work.

Acknowledgments
We would like to thank Ms. Hu Yan-Ling for providing us with the solution to the technical question of co-expression network analysis. And in writing this paper, we would like to express our sincere thanks to all those who have helped us.

**Author’ information**

Xiao-Yan Huang, Email: huangxiaoyan691@163.com.

Wen-Tao Qin, Email: qwt1713974028@163.com

Qi-Sheng Su, Email: 1753359993@qq.com

Cheng-Cheng Qiu, Email: qiuchengcheng@stu.gxmu.edu.cn

Ruo-Chuan Liu, Email: 1659402429@qq.com

Shan-Shan Xie, Email: 1092081241@qq.com

Yang Hu, email: 529656913@qq.com

Shang-Yong Zhu, Email: zhushangyongsound@163.com.

**Reference**

1. Siegel RL, Miller KD and Jemal A (2019) Cancer statistics, 2019. CA Cancer J Clin 69:7-34. doi: 10.3322/caac.21551

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

3. Mizrahi JD, Surana R, Valle JW and Shroff RT (2020) Pancreatic cancer. The Lancet 395:2008-2020. doi: 10.1016/s0140-6736(20)30974-0

4. Lee YT, Tan YJ and Oon CE (2018) Molecular targeted therapy: Treating cancer with specificity. Eur J Pharmacol 834:188-196. doi: 10.1016/j.ejphar.2018.07.034

5. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kaminska B, Huelsken J, Omberg L, Gevaert O, Colaprico A, Czerwinska P, Mazurek S, Mishra L, Heyn H, Krasnitz A, Godwin AK, Lazard AJ, Cancer Genome Atlas Research N, Stuart JM, Hoadley KA, Laird PW, Noushmehr H and
Wiznerowicz M (2018) Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. Cell 173:338-354 e15. doi: 10.1016/j.cell.2018.03.034

6. Miranda A, Hamilton PT, Zhang AW, Pattnaik S, Becht E, Mezheyeuski A, Bruun J, Micke P, de Reynies A and Nelson BH (2019) Cancer stemness, intratumoral heterogeneity, and immune response across cancers. Proc Natl Acad Sci U S A 116:9020-9029. doi: 10.1073/pnas.1818210116

7. Valle S, Martin-Hijano L, Alcala S, Alonso-Nocelo M and Sainz B, Jr. (2018) The Ever-Evolving Concept of the Cancer Stem Cell in Pancreatic Cancer. Cancers (Basel) 10. doi: 10.3390/cancers10020033

8. Barabasi AL, Gulbahce N and Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12:56-68. doi: 10.1038/nrg2918

9. Li C, Heidt DG, Dalerba P, Burant CF, Zhang L, Adsay V, Wicha M, Clarke MF and Simeone DM (2007) Identification of pancreatic cancer stem cells. Cancer Res 67:1030-7. doi: 10.1158/0008-5472.CAN-06-2030

10. Dalla Pozza E, Dando I, Biondani G, Brandi J, Costanzo C, Zoratti E, Fassan M, Boschi F, Melisi D, Cecconi D, Scupoli MT, Scarpa A and Palmieri M (2015) Pancreatic ductal adenocarcinoma cell lines display a plastic ability to bidirectionally convert into cancer stem cells. Int J Oncol 46:1099-108. doi: 10.3892/ijo.2014.2796

11. Di Carlo C, Brandi J and Cecconi D (2018) Pancreatic cancer stem cells: Perspectives on potential therapeutic approaches of pancreatic ductal adenocarcinoma. World J Stem Cells 10:172-182. doi: 10.4252/wjsc.v10.i11.172

12. Pan S, Zhan Y, Chen X, Wu B and Liu B (2019) Identification of Biomarkers for Controlling Cancer Stem Cell Characteristics in Bladder Cancer by Network Analysis of Transcriptome Data Stemness Indices. Front Oncol 9:613. doi: 10.3389/fonc.2019.00613

13. Pei G, Chen L and Zhang W (2017) WGCNA Application to Proteomic and Metabolomic Data Analysis. Methods Enzymol 585:135-158. doi:
14. Langfelder P and Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559. doi: 10.1186/1471-2105-9-559

15. Langfelder P, Zhang B and Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24:719-20. doi: 10.1093/bioinformatics/btm563

16. Niemira M, Collin F, Szalkowska A, Bielska A, Chwialkowska K, Reszec J, Niklinski J, Kwasniewski M and Kretowski A (2019) Molecular Signature of Subtypes of Non-Small-Cell Lung Cancer by Large-Scale Transcriptional Profiling: Identification of Key Modules and Genes by Weighted Gene Co-Expression Network Analysis (WGCNA). Cancers (Basel) 12. doi: 10.3390/cancers12010037

17. Wang CCN, Li CY, Cai JH, Sheu PC, Tsai JJP, Wu MY, Li CJ and Hou MF (2019) Identification of Prognostic Candidate Genes in Breast Cancer by Integrated Bioinformatic Analysis. J Clin Med 8. doi: 10.3390/jcm8081160

18. Kanehisa M, Sato Y, Furumichi M, Morishima K and Tanabe M (2019) New approach for understanding genome variations in KEGG. Nucleic Acids Res 47: D590-D595. doi: 10.1093/nar/gky962

19. Mazandu GK, Chimusa ER and Mulder NJ (2017) Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. Brief Bioinform 18:886-901. doi: 10.1093/bib/bbw067

20. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ and von Mering C (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45: D362-D368. doi: 10.1093/nar/gkw937

21. Blanche P, Dartigues JF and Jacqmin-Gadda H (2013) Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med 32:5381-97. doi:
10.1002/sim.5958

22. Zhang Z and Kattan MW (2017) Drawing Nomograms with R: applications to categorical outcome and survival data. Ann Transl Med 5:211. doi: 10.21037/atm.2017.04.01

23. Subramanian A, Kuehn H, Gould J, Tamayo P and Mesirov JP (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics 23:3251-3. doi: 10.1093/bioinformatics/btm369

24. Tochimoto M, Oguri Y, Hashimura M, Konno R, Matsumoto T, Yokoi A, Kodera Y and Saegusa M (2019) S100A4/non-muscle myosin II signaling regulates epithelial-mesenchymal transition and stemness in uterine carcinosarcoma. Lab Invest. doi: 10.1038/s41374-019-0359-x

25. Reya T, Morrison SJ, Clarke MF and IL W (2001) Stem cells, cancer stem cells. Nature 414:105-11. doi: 10.1038/35102167

26. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, Brooks M, Reinhard F, Zhang CC, Shipitsin M, Campbell LL, Polyak K, Brisken C, Yang J and Weinberg RA (2008) The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell 133:704-15. doi: 10.1016/j.cell.2008.03.027

27. <Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell.pdf>.

28. <A cell initiating human acute myeloid leukaemia after transplantation into SCID mice.pdf>.

29. <Prospective identification of tumorigenic breast cancer cells..pdf>.

30. Medema JP (2013) Cancer stem cells: the challenges ahead. Nat Cell Biol 15:338-44. doi: 10.1038/ncb2717

31. Yang MC, Wang HC, Hou YC, Tung HL, Chiu TJ and Shan YS (2015) Blockade of autophagy reduces pancreatic cancer stem cell activity and potentiates the tumoricidal effect of gemcitabine. Mol Cancer 14:179. doi: 10.1186/s12943-015-0449-3
32. Visvader JE and Lindeman GJ (2008) Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nat Rev Cancer 8:755-68. doi: 10.1038/nrc2499

33. Zhang M, Wang X, Chen X, Guo F and Hong J (2020) Prognostic Value of a Stemness Index-Associated Signature in Primary Lower-Grade Glioma. Front Genet 11:441. doi: 10.3389/fgene.2020.00441

34. Sancho P, Alcala S, Usachov V, Hermann PC and Sainz B, Jr. (2016) The ever-changing landscape of pancreatic cancer stem cells. Pancreatology 16:489-96. doi: 10.1016/j.pan.2016.04.004

35. Bai L, Ren Y and Cui T (2020) Overexpression of CDCA5, KIF4A, TPX2, and FOXM1 Coregulated Cell Cycle and Promoted Hepatocellular Carcinoma Development. J Comput Biol 27:965-974. doi: 10.1089/cmb.2019.0254

36. Wang F, Zhao W, Gao Y, Zhou J, Li H, Zhang G, Guo D, Xie C, Li J, Yin Z and Zhang J (2019) CDK5-mediated phosphorylation and stabilization of TPX2 promotes hepatocellular tumorigenesis. J Exp Clin Cancer Res 38:286. doi: 10.1186/s13046-019-1297-6

37. Liu D, Xu X, Wen J, Xie L, Zhang J, Shen Y, Jiang G, Chen J and Fan M (2018) Integrated Genome-Wide Analysis of Gene Expression and DNA Copy Number Variations Highlights Stem Cell-Related Pathways in Small Cell Esophageal Carcinoma. Stem Cells Int 2018:3481783. doi: 10.1155/2018/3481783

38. Song Y, Kim S, Lee H, No JH, Ryu HC, Kim J, Lim JW, Kim M, Choi I and Seo HR (2020) Chromenopyrimidinone Controls Stemness and Malignancy by suppressing CD133 Expression in Hepatocellular Carcinoma. Cancers (Basel) 12. doi: 10.3390/cancers12051193

39. Liu B, Xiao Y, Li H, Zhang AL, Meng LB, Feng L, Zhao ZH, Ni XC, Fan B, Zhang XY, Zhao SB and Liu YB (2020) Identification and Verification of Biomarker in Clear Cell Renal Cell Carcinoma via Bioinformatics and Neural Network Model. Biomed Res Int 2020:6954793. doi: 10.1155/2020/6954793

40. F. Li, D. Jin, L. Guan, C.C. Zhang, T. Wu, Y.J. Wang, D.S. Gao (2019) CEP55
promoted the migration, invasion and neurosphere formation of the glioma cell line U251, Neurosci. Lett. 705:80-86. doi: 10.1016/j.neulet.2019.04.038

41. Bavi P, Uddin S, Ahmed M, Jehan Z, Bu R, Abubaker J, Sultana M, Al-Sanea N, Abduljabbar A, Ashari LH, Alhomoud S, Al-Dayel F, Prabhakaran S, Hussain AR and Al-Kuraya KS (2011) Bortezomib stabilizes mitotic cyclins and prevents cell cycle progression via inhibition of UBE2C in colorectal carcinoma. Am J Pathol 178:2109-20. doi: 10.1016/j.ajpath.2011.01.034

42. Zhou Z, Cheng Y, Jiang Y, Liu S, Zhang M, Liu J and Zhao Q (2018) Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. Int J Biol Sci 14:124-136. doi: 10.7150/ijbs.22619

43. Horning AM, Wang Y, Lin CK, Louie AD, JadHAV RR, Hung CN, Wang CM, Lin CL, KirmA NB, Liss MA, Kumar AP, Sun L, Liu Z, Chao WT, Wang Q, Jin VX, Chen CL and Huang TH (2018) Single-Cell RNA-seq Reveals a Subpopulation of Prostate Cancer Cells with Enhanced Cell-Cycle-Related Transcription and Attenuated Androgen Response. Cancer Res 78:853-864. doi: 10.1158/0008-5472.CAN-17-1924

44. Kim, J.H., Youn, Y., Kim, K.T., Jang, G., and Hwang, J.H. (2019). Non-SMC condensin I complex subunit H mediates mature chromosome condensation and DNA damage in pancreatic cancer cells. Sci. Rep. 9, 17889.doi: 10.1038/s41598-019-54478-3.