Medicine®

OPEN

Construction of a 6-gene prognostic signature to assess prognosis of patients with pancreatic cancer

Jiayue Yang, MD[a], Wei Shi, MD[a], Shengwei Zhu, MD[a], Cheng Yang, MD[b,*]

Abstract

Pancreatic cancer (PaCa) is one of the most fatal cancers in the world. Although great efforts have made to explore the mechanisms of PaCa oncogenesis, the prognosis of PaCa patients is still unsatisfactory. Thus, it is imperative to further understand the potential carcinogenesis of PaCa and reliable prognostic models.

The gene expression profile and clinical information of GSE21501 were downloaded from the Gene Expression Omnibus (GEO) database. Weighted gene co-expression network analysis (WGCNA) was applied to explore the potent genes associated with the overall survival (OS) events of PaCa patients. Cox regression model was applied to selecting prognostic genes and establish prognostic model. The prognostic values of six-gene signature were validated in TCGA-PAAD cohort.

According to the WGCNA analysis, a total of 19 modules were identified and 115 hub genes in the mostly associated module were reserved for next analysis. According to the univariate and multivariate Cox regression analysis, we established a six-gene signature (*FTSJ3*, *STAT1*, *STX2*, *CDX2*, *RASSF4*, *MACF1*) which could effectively evaluate the overall survival (OS) of PaCa patients. In validated patients' cohorts, the six-gene signature exhibited excellent prognostic value in TCGA-PAAD cohort as well.

We developed a six-gene signature to exactly predict OS of PaCa patients and provide a novel personalized strategy for evaluating prognosis. The findings may be contributed to medical customization and therapeutic decision in clinical practice.

Abbreviations: 95%CI = 95% confidence interval, CDX2 = caudal type homeobox 2, DEGs = different expression genes, EMT = epithelial-mesenchymal translation, FTSJ3 = FTSJ homolog 3, GEO = the Gene Expression Omnibus, HR = hazard ratio, MACF1 = microtubule-actin crosslinking factor 1, OS = overall survival, PaCa = pancreatic cancer, RASSF4 = Ras association (RalGDS/AF-6) domain family member 4, STAT1 = signal transducer and activator of transcription 1, STX2 = syntaxin 2, TCGA = the Cancer Genome Atlas, WGCNA = weighted gene co-expression network analysis.

Keywords: Cox regression analysis, pancreatic cancer, prognosis, WGCNA

1. Introduction

Pancreatic cancer (PaCa) is a high-incidence tumor in the digestive system, which characterized with high degree of malignancy. At present, the incidence and mortality of PaCa are increasing yearly, and in 2019 PaCa is the fourth cause for tumor-related deaths in the United States, with 56,770 new cases and 45,750 deaths.[1] Due to the lack of typical clinical manifestations in patients with early PaCa, the faster growth of tumors and early tend of metastasis, more than 90% of patients miss the best treatment opportunity, which are diagnosed in the middle and/or advanced stages.[2,3] Besides, the treatments for PaCa are usually difficult. Thus, several reasons above contribute to the poor prognosis in patients with PaCa. In summary, for the early diagnosis and exact prognostic evaluation of PaCa, it is still significant to explore more effective biomarkers or gene models to exactly predict the prognostic of PaCa patients.

In the past decades, continuously developing computer analysis methods has significantly contributed to the wide application of "big data" in biomedical research. As an emerging cross-discipline subject in biomedicine and computer science, bioinformatics has been widely applied to auxiliary study of multifaceted clinical and/or basic medical research.[4–6] Weighted gene co-expression network analysis (WGCNA) is a comprehensive analysis strategy for assessing association between gene clusters and phenotypes in different samples.[7,8] Scholars applied

Editor: Raffaele Pezzilli.

The main results of this research is based on bioinformatics analysis, thus the ethics approval is not applicable.

This work was supported by the Natural Science Foundation for Young Scientists of Jiangsu Province (BK20170213), the Youth Foundation of Wuxi Health and Family Planning Commission (Q201719), the Wuxi Medical Youth Talents Foundation (QNRC062), and this research also funded by Wuxi Science and Technology Development Foundation (CSE31N1704) to Shengwei Zhu.

The authors have no conflicts of interest to disclose.

Supplemental Digital Content is available for this article.

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

[a] *Department of Endocrinology,* [b] *Department of Gastroenterology, Wuxi People's Hospital Affiliated to Nanjing Medical University, Wuxi 214023, China.*

[*] *Correspondence: Cheng Yang, Department of Gastroenterology, Wuxi People's Hospital Affiliated to Nanjing Medical University, 299 Qing Yang Road, Wuxi 214023, China (e-mail: yang_301@njmu.edu.cn).*

How to cite this article: Yang J, Shi W, Zhu S, Yang C. Construction of a 6-gene prognostic signature to assess prognosis of patients with pancreatic cancer. Medicine 2020;99:37(e22092).

Received: 13 May 2020 / Received in final form: 29 July 2020 / Accepted: 6 August 2020

http://dx.doi.org/10.1097/MD.0000000000022092

WGCNA to systematically identify gene clusters of significantly altered genes or all genes, and then perform association analyses with phenotypes. At present, numerous latent biomarkers in cancers have been authenticated based on WGCNA of RNA-sequencing data.[9–11] Besides, investigators tend to utilized WGCNA to systematically analyze cancer patients' phenotypes, especially for establishing potential prognostic gene signature.[12,13]

GSE21501, a microarray containing 132 PaCa samples, was contributed by Stratford et al in 2010.[14] A total 102 samples both contained gene expression data and survival time. Stratford et al. developed and validated a six-gene signature based on this microarray. Given the integrated survival information and gene expression data in GSE21501, we re-evaluated this microarray and finally established a novel six-gene signature predicting prognosis of PaCa patients through WGCNA and Cox regression analysis. We further performed validation through TCGA-PAAD cohort. In conclusion, we found out a six-gene signature (*FTSJ3*, *STAT1*, *STX2, CDX2*, *RASSF4*, *MACF1*) associated poor prognosis, suggesting a novel gene scoring model to evaluate the prognosis of PaCa patients.

2. Materials and methods

2.1. Acquisition of microarray data and pre-processing

The workflow of our study was shown in Fig. 1. The microarray profiles of GSE21501 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21501) were downloaded from the Gene Expression Omnibus (GEO) database. Besides, Level 3 mRNA expression and clinical data of 178 PaCa samples were downloaded from the Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/). Before further analysis, GSE21501 microarray profiles from GEO database were pre-processed by background correction, quantile normalization, and probe summarization. After matching the gene expression data and survival information, 102 samples from GSE21501 were retained to be used as training dataset in the current research. For further WGCNA analysis, the top 25% different expression genes (DEGs) from GSE21501 dataset based on analysis of variance (3951 genes) were retained.

2.2. Construction of co-expression network

After pre-processing the GSE21501 microarray data, the transcriptional expression profiles of these 3951 genes was sent to construct a gene co-expression network by the WGCNA package in R language.[8] The idea of a soft threshold is to continually elementize the elements in the Adjacency Matrix through a weight function and the choice of the soft threshold β is bound to influence the result of module identification. To create a network with a nearly scale-free topology, we installed the soft threshold power of $\beta=7$ (scale free $R^2=0.887$). Adjacency matrices were calculated and transformed into the topological overlap matrix. The dynamic tree cut algorithm was used to detect gene modules. Gene significance was defined as the correlation coefficient between gene expression and module traits. The module eigengene was calculated as a summary profile for each module. Module significance was defined as the correlation coefficient between a module's eigengene and traits, gene modules were considered statistically significant if the *Pearson P* values were less than 0.05.

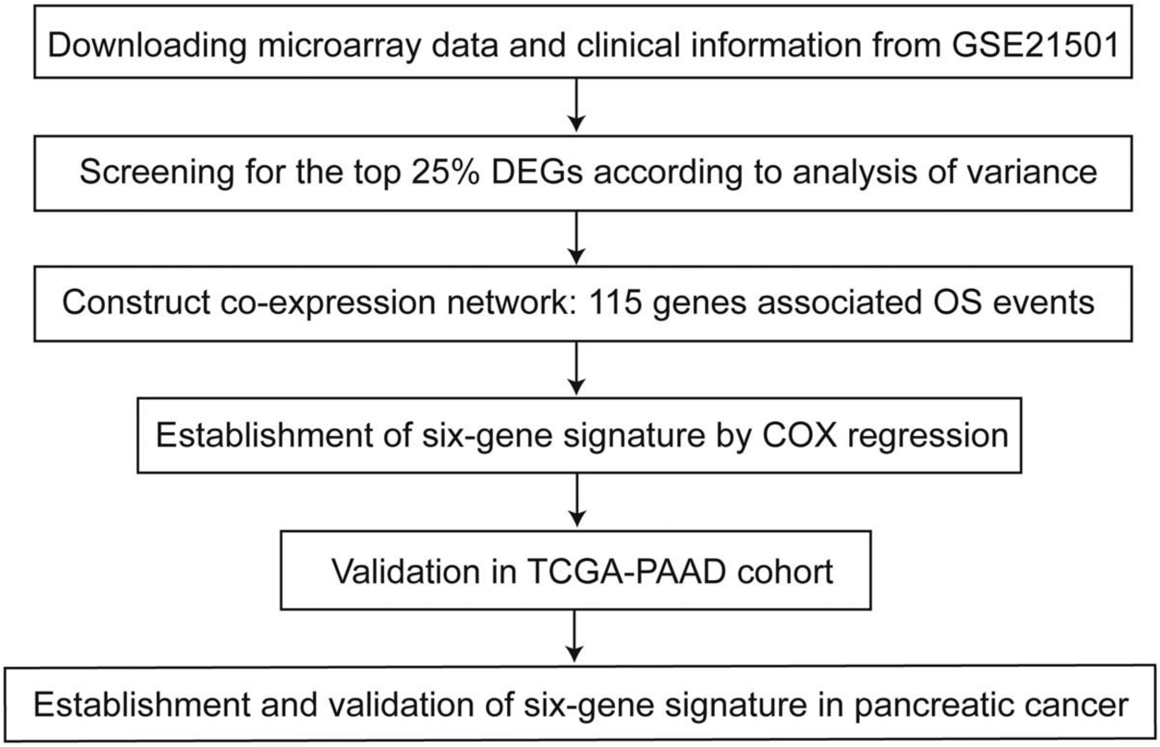
Downloading microarray data and clinical information from GSE21501
Screening for the top 25% DEGs according to analysis of variance
Construct co-expression network: 115 genes associated OS events
Establishment of six-gene signature by COX regression
Validation in TCGA-PAAD cohort
Establishment and validation of six-gene signature in pancreatic cancer

Figure 1. Flow chart of the research. The gene expression profiles of GSE21501 were downloaded from the GEO database. WGCNA was applied to investigate potential biomarkers associated with the OS events. Next, a six-gene signature was identified by univariate and multivariate Cox regression analysis. In addition, the prognostic value of the six-gene signature in PaCa was validated in TCGA-PAAD cohort.

2.3. Establishment and validation of the prognostic gene signature

According to the WGCNA analysis, 115 genes in all the modules associated overall survival (OS) events were reserved for next analysis. A total 55 genes showing significantly associated with prognosis of PaCa patients based univariate Cox regression analysis. Next, multivariate Cox regression analysis was applied to develop multiple-gene signature predicting PaCa prognosis and a six-gene signature was established. All the PaCa patients were scored by six-gene model: Risk score = 3.641 *FTSJ3 + 2.293 *STAT1 + 1.900 *STX2 – 1.214 *CDX2 – 1.381 *RASSF4 – 2.342 *MACF1. To verify the significant values of the six-gene signature, 178 samples from TCGA-PAAD cohort was treated as an independent testing dataset. For Kaplan-Meier analysis, all cases were ranked based on risk score and further divided into two groups according to the average score.

2.4. Statistical analysis

All statistical analyses were performed using SPSS 25.0 software and R 3.5.1 software. For Kaplan-Meier analysis, all cohorts were evaluated by Kaplan-Meier survival plots. The hazard ratio (HR), 95% confidence interval (95%CI), and log rank P value were calculated. For all analyses, differences were considered statistically significant if the P value was less than 0.05.

3. Results

3.1. Weighted co-expression network construction and crucial genes identification

We applied the R package for WGCNA in the construction of a co-expression network and then 3951 DEGs with similar expression features were submitted to modules through cluster analysis. In the current study, we selected the power of $\beta = 7$ (scale free $R^2 = 0.887$) as the soft threshold to construct a scale-free network (Fig. 2A-2D). Then, we identified nineteen modules for next analysis (Fig. 3A). Afterwards, we took advantage of a heatmap and meta-modules aiming to visualize the gene network (Fig. 3B). Among all modules, cyan gene module exhibited significant correlation with overall survival (OS) events ($R^2 = 0.27$; $P = 0.007$) (Fig. 3C, 3D), shown remarkable values in the evaluation of PaCa prognosis. Subsequently, we totally extracted 115 genes in the cyan module for further univariate Cox regression analysis.

3.2. Developing prognostic six-gene signature using COX regression analysis

A total of 115 genes were identified by WGCNA analysis. Then, Cox regression analysis was applied to develop prognosis related gene signature to establish novel model for assessing prognosis of PaCa patients. Univariate Cox regression model identified 55 genes that remarkably correlated with OS (Supplemental Digital Content [Table S1, http://links.lww.com/MD/E819]). Then, multivariate Cox analysis was conducted in the training set to further select prognosis-associated genes (Supplemental Digital Content [Table S2, http://links.lww.com/MD/E820]). Six genes were extracted and next applied to construct a prognostic gene signature.

The six genes identified were FTSJ homolog 3 (FTSJ3), signal transducer and activator of transcription 1 (STAT1), syntaxin 2 (STX2), caudal type homeobox 2 (CDX2), Ras association (RalGDS/AF-6) domain family member 4 (RASSF4), and microtubule-actin crosslinking factor 1 (MACF1). The risk score = 3.641 *FTSJ3 + 2.293 *STAT1 + 1.900 *STX2 – 1.214 *CDX2 – 1.381 *RASSF4 – 2.342 *MACF1 (Fig. 4A). Figure 4B exhibited the survival overview in the training cohort. Then, PaCa patients were divided into two groups: the low-risk group ($n = 51$) and the high-risk group ($n = 51$) based on median of risk value (Fig. 4C). The Kaplan-Meier curve and log-rank test uncovered that patients in the high-risk group had remarkably worse OS compared to those in the low-risk group (HR = 2.575, $P < 0.001$) (Fig. 4D). Taken together, our findings suggested a well performance of the six-gene signature for prognostic prediction in PaCa patients.

3.3. Validation of the six-gene signature for OS prediction

To confirm the six-gene signature developed in the training set, we validated the prognostic value of the signature in the TCGA cohort. A total of 178 PaCa patients with survival information were retained, we calculated the risk score for each patient by using the same formula in the training set. Figure 5A showed the survival overview in the validation cohort, and Fig. 5B showed the risk score distribution. According to the median of risk value, the cohort of patients were grouped into high- ($n = 89$) and low-risk ($n = 89$) groups. The Kaplan-Meier curve revealed a notable worse OS in the low-risk group compared to the high-risk group (HR = 1.556, $P = 0.035$) (Fig. 5C). The result was in accordance with our findings in the training cohort based on GSE21501 dataset, suggesting that the six-gene signature was validated as a promising indicator for OS in PaCa patients.

4. Discussion

PaCa is a common malignant tumor in the digestive system with high degree of malignancy. The prognostic evaluation of PaCa has always been a hot topic for concerned scholars. Establishing gene scoring models contribute to quantify the prognostic evaluation criteria and increasing studies have performed successful precedents in this regard.[12,15,16] Our current study established a six-gene prognosis predictive model in patients with PaCa and validated the prognostic value of this model in total PaCa patients and several subgroups, confirming that the six-gene prognosis model could exactly predict the OS of patients with PaCa.

In our study, we applied integrated bioinformatics strategies to establish the six-gene signature. To obtain prognosis-related genes, we used the WGCNA analysis to explore potential gene modules associated with OS event. Afterwards, univariate Cox regression analysis was conducted to further verify the prognostic values of 115 prognosis-related genes. As the core procedure of this study, multivariate Cox regression model were constructed based on statistically-significant genes in univariate Cox regression analysis and finally developed a novel six-gene (*FTSJ3*, *STAT1*, *STX2, CDX2*, *RASSF4*, *MACF1*) prognostic signature.

FTSJ3, STAT1, and STX2 were identified as risky markers in our research. FTSJ3 is a new-found gene and the function of this gene is not known. Morello et al indicated that FTSJ3 interacted between NIP7 during pre-rRNA processing and revealed that FTSJ3 participated in ribosome synthesis in human cells.[17] Recently, FTSJ3 was identified as an RNA 2′-O-methyltransferase and can be recruited by HIV to avoid innate immune

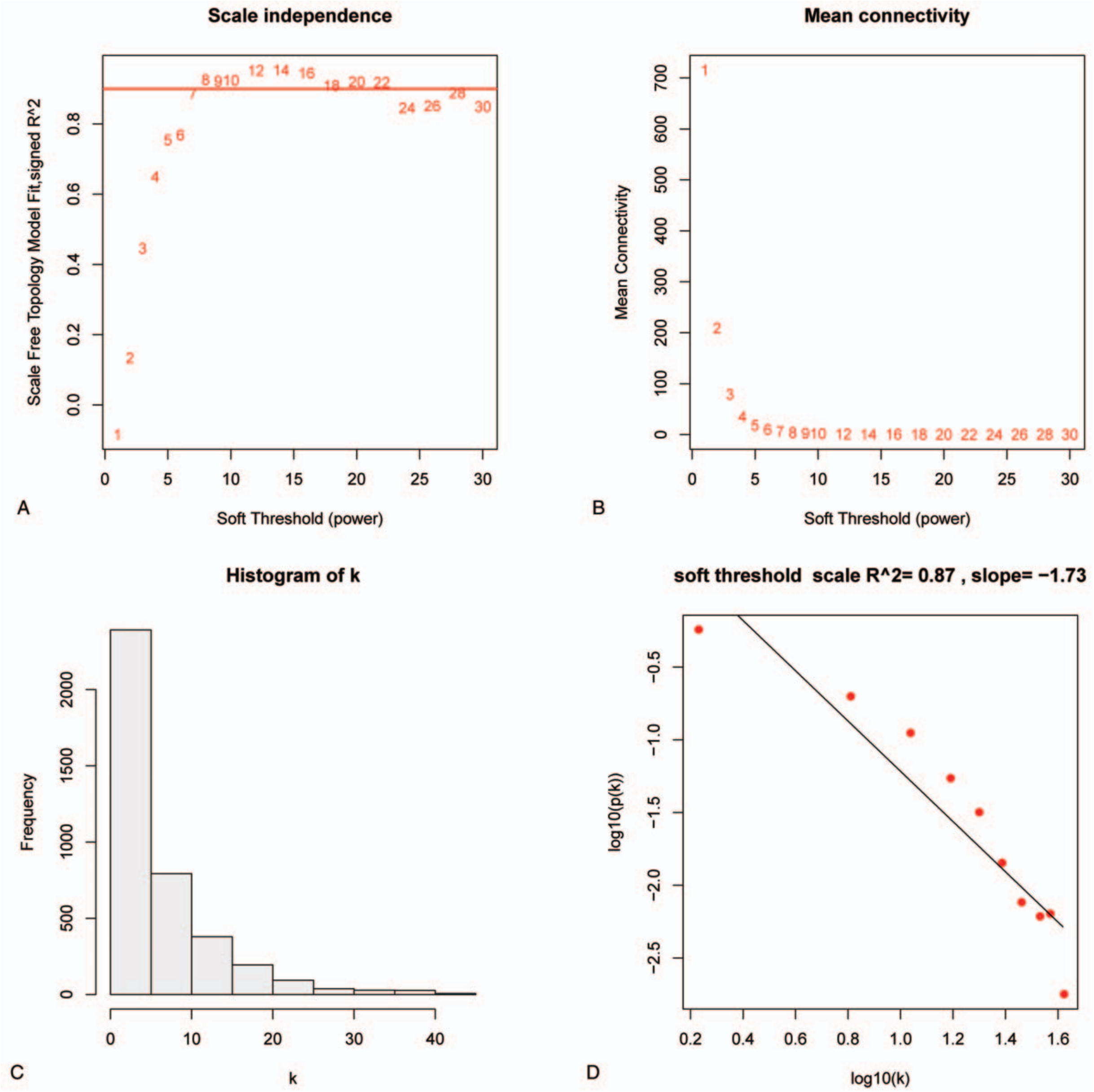

Figure 2. Determination of soft-thresholding power in WGCNA. (A) Analysis of the scale-free fitting indices for various soft-thresholding powers (β). (B) Mean connectivity analysis of various soft-thresholding powers. (C) Histogram of the connection distribution when $\beta = 7$. (D) Checking the scale-free topology when $\beta = 7$. According to Fig. 2C and D, k and p(k) were negatively correlated (correlation coefficient is 0.87), indicating that a gene scale-free network can be resumed.

sensing.[18] The role of STAT1 in cancers has been well established. Most evidence showed that activating STAT1 acted as a tumor suppressor role in multiple cancer cells. However, STAT1 also functioned as oncogene under specific conditions.[19] STX2 regulates epithelial-mesenchymal translation (EMT) and epithelial cell morphogenesis and activation. Several studies showed STX2 acted as oncogene in cancers. STX2 promoted metastasis of colorectal cancer through a positive feedback loop which could activate the NF-κB pathway.[20]

Other three genes, CDX2, RASSF4, and MACF1 were identified as protective markers. CDX2 is a member of the caudal-related homeobox transcription factor gene family. The encoded protein is an important mediator of intestine-specific genes participating in cell growth and differentiation. CDX2 has been uncovered to be a prognostic indicator in multiple cancers, particularly in gastrointestinal cancers.[21–23] The function of RASSF4 has not yet been determined but may play a role in tumor suppression. Studies manifested that RASSF4 was downregulated in lung cancer as well as and exogenously overexpressed RASSF4 significantly inhibited cancer cell proliferation and invasion.[24,25] MACF1 played a significant function in cytoskeleton organization. Aberrant expression of MACF1 was observed in several cancers, such as colon cancer, breast cancer, lung cancer, and glioblastoma, which initiated cancer cell proliferation, migration, and invasion.[26]

We further proved that six-gene signature is a powerful prognostic model for OS evaluation in PaCa patients using the TCGA cohort, confirming the gene signature was a promising

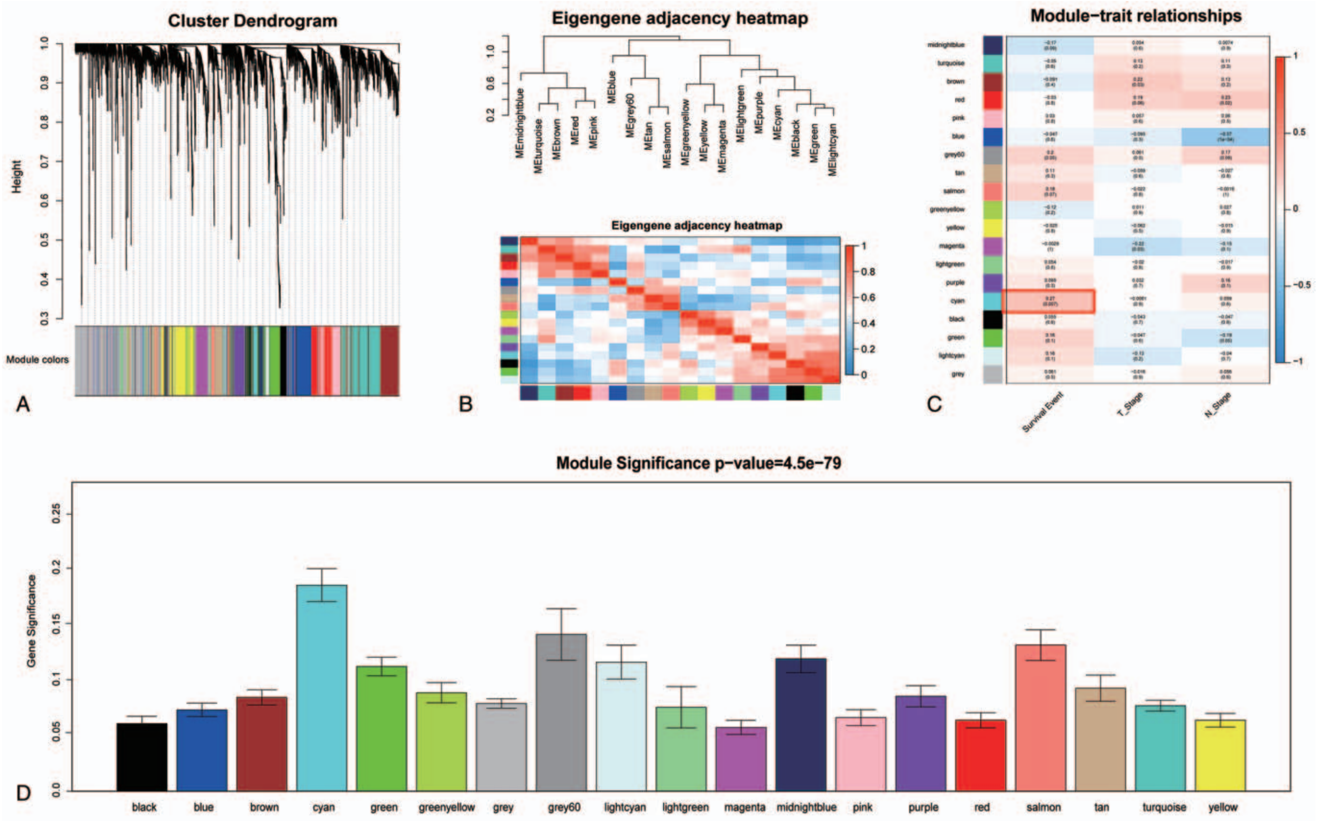

Figure 3. Identification of relevant modules associated with PaCa clinical traits. (A) Clustering dendrograms of genes were based on dissimilarity topological overlap and module colours. As a result, 19 co-expression modules were constructed and are shown in different colours. These modules were arranged from large to small according to the number of genes included. (B) The eigengene dendrogram and heatmap identify groups of correlated with eigengenes termed meta-modules. (C) Heatmap of the correlation between module eigengenes and clinical traits of PaCa. The cyan gene module was revealed to exhibit the highest correlation with OS events. (D) Distribution of average gene significance and errors in the modules associated with OS event of PaCa.

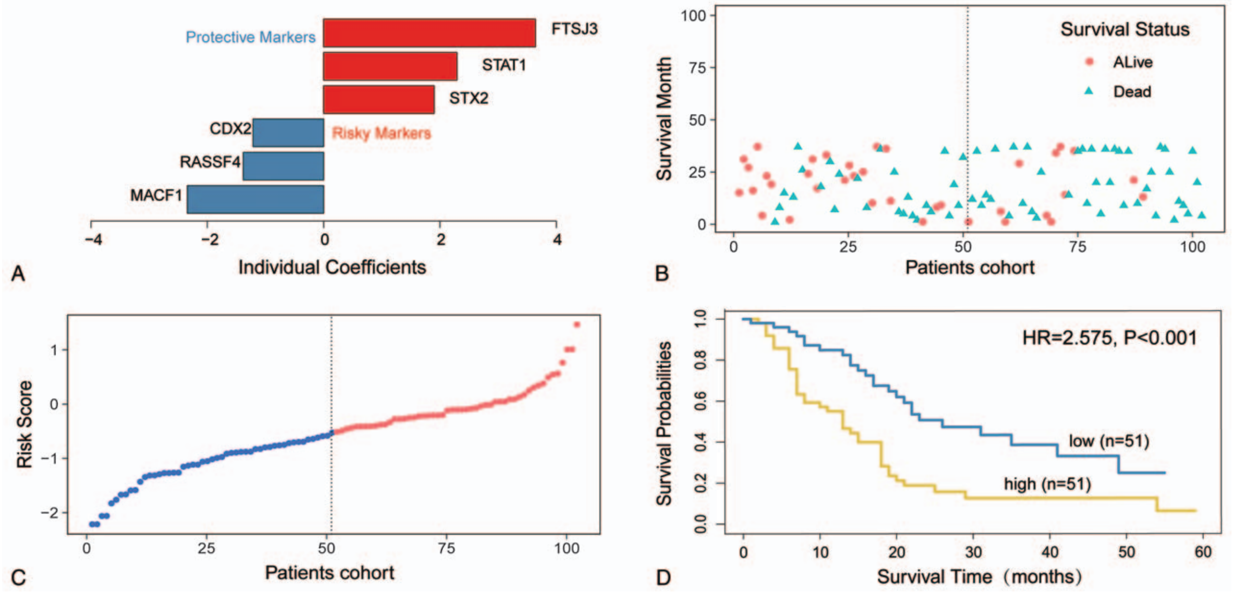

Figure 4. Identification of a six-gene signature as promising predictor in PaCa. (A) Cox coefficients distribution of the six-gene signature. (B) Survival overview in the training cohort. (C) Risk score distribution in the training cohort. (D) Patients in the high-risk group exhibited worse overall survival compared to those in the low-risk group.

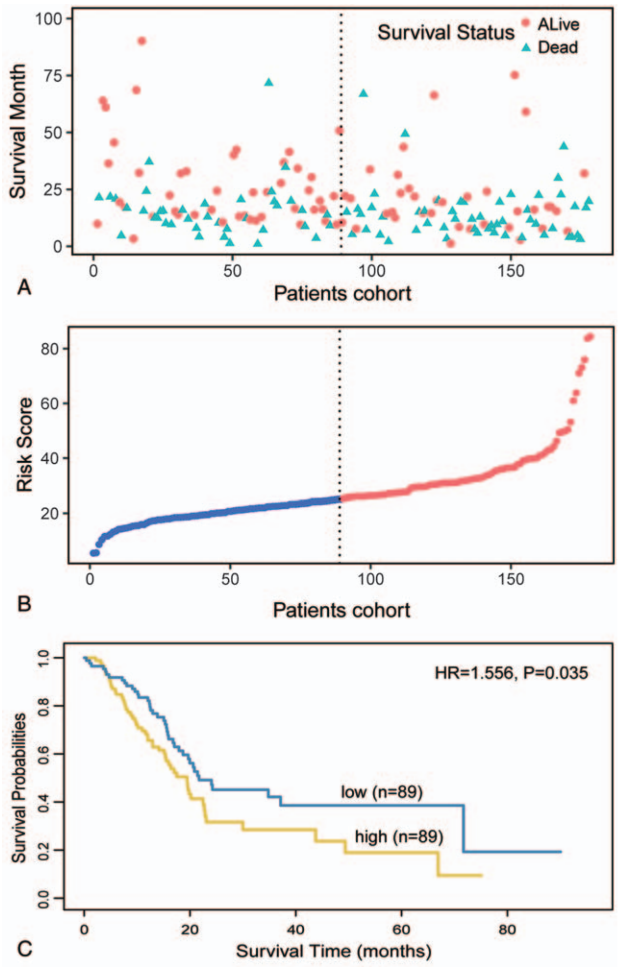

Figure 5. Validation of the signature in TCGA-PAAD cohort. Survival overview in the validated cohort. (B) Risk score distribution in the validated cohort. (C) Patients in the high-risk group exhibited worse overall survival compared to those in the low-risk group.

indicator for OS in PaCa patients. Compared to other established signatures to assess OS in PaCa, our model was constructed and validated from more comprehensive databases and it seemed to be more convenient to be applied in clinical practice with fewer numbers of genes. For example, Birnbaum et al identified a 25-gene signature to predict OS in resectable PaCa. They used supervised analysis to identify 1400 genes differentially expressed between 17 long-term survivors and 22 short-term survivors, then a 25-gene prognostic indicator was established. Obviously, Birnbaum et al established a complex multi-gene signature including 25 genes which were not suitable for clinical practice.[27] However, we identified a six-gene signature, which was easier to be used in clinical practice.

The current study also has several limitations. First of all, due to the limited number of PaCa patients, subgroup analysis of the prognostic value of the signature cannot be conducted; second, the study is based on bioinformatics analysis, and there is no recruited cohorts for prognostic verification; finally, the roles of six hub genes in PaCa have not been fully understood. Thus, in further research, we will validate the expressions of six hub genes and focus on their functions.

5. Conclusion

To sum up, we constructed a novel six-gene prognostic model to predict OS in PaCa patients. Although several limitations to this study exist, developing a multi-gene prognosis model could provide exact prognostic evaluation strategy than clinical pathological indicators, which has significant guidance value for the selection of individualized therapies.

Author contributions

Designed the experiments: CY and JY; Acquisition of data: JY and WS; Analysis and interpretation of data: JY, WS, and SZ; Draft of the manuscript: JY; Critical revision of the manuscript for intellectual content: CY; Funding Acquisition: CY.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69:7–34.

[2] Nkembo AT, Salako O, Poku RA, et al. Disruption of actin filaments and suppression of pancreatic cancer cell viability and migration following treatment with polyisoprenylated cysteinyl amides. Am J Cancer Res 2016;6:2532–46.

[3] Balan BJ, Zygmanowska E, Radomska-Lesniewska DM. Disorders noticed during development of pancreatic cancer: potential opportunities for early and effective diagnostics and therapy. Cent Eur J Immunol 2017;42:377–82.

[4] Tang RX, Chen WJ, He RQ, et al. Identification of a RNA-Seq based prognostic signature with five lncRNAs for lung squamous cell carcinoma. Oncotarget 2017;8:50761–73.

[5] Toiyama Y, Hur K, Tanaka K, et al. Serum miR-200c is a novel prognostic and metastasis-predictive biomarker in patients with colorectal cancer. Ann Surg 2014;259:735–43.

[6] Mei J, Wang H, Wang R, et al. Evaluation of X-ray repair cross-complementing family members as potential biomarkers for predicting progression and prognosis in hepatocellular carcinoma. Biomed Res Int 2020;2020:5751939.

[7] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4: Article 17.

[8] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.

[9] Luo Y, Coskun V, Liang A, et al. Single-cell transcriptome analyses reveal signals to activate dormant neural stem cells. Cell 2015;161:1175–86.

[10] Cai Y, Mei J, Xiao Z, et al. Identification of five hub genes as monitoring biomarkers for breast cancer metastasis in silico. Hereditas 2019;156:20.

[11] Chen J, Cai Y, Xu R, et al. Identification of four hub genes as promising biomarkers to evaluate the prognosis of ovarian cancer in silico. Cancer Cell Int 2020;20:270.

[12] Kim J, Jo YH, Jang M, et al. PAC-5 gene expression signature for predicting prognosis of patients with pancreatic adenocarcinoma. Cancers (Basel) 2019;11:1749.

[13] Xue M, Shang J, Chen B, et al. Identification of prognostic signatures for predicting the overall survival of uveal melanoma patients. J Cancer 2019;10:4921–31.

[14] Stratford JK, Bentrem DJ, Anderson JM, et al. A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. PLoS Med 2010;7:e1000307.

[15] Chen X, Hu L, Wang Y, et al. Single cell gene co-expression network reveals FECH/CROT signature as a prognostic marker. Cells 2019; 8:698.

[16] Zhou C, Zhao Y, Yin Y, et al. A robust 6-mRNA signature for prognosis prediction of pancreatic ductal adenocarcinoma. Int J Biol Sci 2019;15:2282–95.

[17] Morello LG, Coltri PP, Quaresma AJ, et al. The human nucleolar protein FTSJ3 associates with NIP7 and functions in pre-rRNA processing. PLoS One 2011;6:e29174.

[18] Ringeard M, Marchand V, Decroly E, et al. FTSJ3 is an RNA 2'-O-methyltransferase recruited by HIV to avoid innate immune sensing. Nature 2019;565:500–4.

[19] Zhang Y, Liu Z. STAT1 in cancer: friend or foe? Discov Med 2017;24:19–29.

[20] Wang Y, Xu H, Jiao H, et al. STX2 promotes colorectal cancer metastasis through a positive feedback loop that activates the NF-kappaB pathway. Cell Death Dis 2018;9:664.

[21] Masood MA, Loya A, Yusuf MA. CDX2 as a prognostic marker in gastric cancer. Acta Gastroenterol Belg 2016;79: 197–200.

[22] Asgari-Karchekani S, Karimian M, Mazoochi T, et al. CDX2 protein expression in colorectal cancer and its correlation with clinical and pathological characteristics, prognosis, and survival rate of patients. J Gastrointest Cancer 2019;51:844–9.

[23] Joo MK, Park JJ, Chun HJ. Impact of homeobox genes in gastrointestinal cancer. World J Gastroenterol 2016;22:8247–56.

[24] Zhang M, Wang D, Zhu T, et al. RASSF4 overexpression inhibits the proliferation, invasion, EMT, and Wnt signaling pathway in osteosarcoma cells. Oncol Res 2017;25:83–91.

[25] Han Y, Dong Q, Hao J, et al. RASSF4 is downregulated in nonsmall cell lung cancer and inhibits cancer cell proliferation and invasion. Tumour Biol 2016;37:4865–71.

[26] Miao Z, Ali A, Hu L, et al. Microtubule actin cross-linking factor 1, a novel potential target in cancer. Cancer Sci 2017;108:1953–8.

[27] Birnbaum DJ, Finetti P, Lopresti A, et al. A 25-gene classifier predicts overall survival in resectable pancreatic cancer. BMC Med 2017;15:170.