

RESEARCH

Open Access



# Prognostic assessment capability of a five-gene signature in pancreatic cancer: a machine learning based-study

Xuanfeng Zhang<sup>1,2†</sup>, Lulu Yang<sup>3,4†</sup>, Dong Zhang<sup>1,5</sup>, Xiaochuan Wang<sup>1,2</sup>, Xuefeng Bu<sup>6</sup>, Xinhui Zhang<sup>1,2\*</sup> and Long Cui<sup>1,2\*</sup>

## Abstract

**Background** A prognostic assessment method with good sensitivity and specificity plays an important role in the treatment of pancreatic cancer patients. Finding a way to evaluate the prognosis of pancreatic cancer is of great significance for the treatment of pancreatic cancer.

**Methods** In this study, GTEx dataset and TCGA dataset were merged together for differential gene expression analysis. Univariate Cox regression and Lasso regression were used to screen variables in the TCGA dataset. Screening the optimal prognostic assessment model is then performed by gaussian finite mixture model. Receiver operating characteristic (ROC) curves were used as an indicator to assess the predictive ability of the prognostic model, the validation process was performed on the GEO datasets.

**Results** Gaussian finite mixture model was then used to build 5-gene signature (ANKRD22, ARNTL2, DSG3, KRT7, PRSS3). Receiver operating characteristic (ROC) curves suggested the 5-gene signature performed well on both the training and validation datasets.

**Conclusions** This 5-gene signature performed well on both our chosen training dataset and validation dataset and provided a new way to predict the prognosis of pancreatic cancer patients.

**Keywords** Pancreatic cancer, Bioinformatics, Prognosis, Gaussian finite mixture model, RNA-seq, Machine learning

<sup>†</sup>Xuanfeng Zhang and Lulu Yang contributed equally to this work.

\*Correspondence:

Xinhui Zhang  
zhaomengxi196335@2008.sina.com  
Long Cui  
lboclong@aliyun.com

<sup>1</sup> Center of Hepatobiliary Pancreatic Disease, XuZhou Central Hospital, Jiangsu, People's Republic of China

<sup>2</sup> Center of Hepatobiliary Pancreatic Disease, The Affiliated Xuzhou Hospital of Medical School of Southeast University, No.199 Jiefang South Road, Xuzhou, Jiangsu, People's Republic of China

<sup>3</sup> Department of Radiology, XuZhou Central Hospital, Jiangsu, People's Republic of China

<sup>4</sup> Department of Radiology, The Affiliated Xuzhou Hospital of Medical School of Southeast University, Jiangsu, People's Republic of China

<sup>5</sup> Bengbu Medical College, Anhui, People's Republic of China

<sup>6</sup> Department of General Surgery, Affiliated People's Hospital of Jiangsu University, Zhenjiang, Jiangsu, People's Republic of China



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Introduction

Pancreatic cancer (PC) has one of the worst prognoses among malignant tumors, the overall 5-year survival rate of patients with pancreatic cancer is less than 5% [1]. Due to atypical symptoms, no sensitive early diagnostic biomarkers, and exceptional anatomical structures, only about 20% of patients at diagnosis are on the verge of being resectable by surgery [2]. Investigators are currently pursuing a comprehensive medical treatment plan, which includes immunotherapy, targeted drugs, radiotherapy, and chemotherapy [3]. Due to tumor heterogeneity, different patients respond differently to the same treatment regimen. This requires clinicians to adjust treatment regimens based on each patient's response during treatment. Therefore, a prognostic assessment method with good sensitivity and specificity plays an important role in the treatment of patients.

In prognostic assessment, the use of prognosis-related gene expression is better than the use of patient clinical characteristics [4]. The development of next-generation sequencing technology and gene chip technology provides a convenient, accurate, and inexpensive way for the detection of prognosis-related genes. [5]. A growing number of researchers tend to use next-generation sequencing or gene chips to detect prognosis-related genes, then a prognostic model is established to guide the treatment of patients [6]. The mathematical modeling process includes the use of logistic regression, Poisson regression, Cox regression, lasso regression, and ridge regression [7]. The combined use of these bioinformatic modeling approaches can significantly improve the specificity and sensitivity of prognostic models.

Here, we obtained differential genes (DEGs) using the TCGA-PAAD dataset, GTEx and two GEO datasets. For further analysis, we normalized and de-batched all datasets. Eight prognosis-related genes were screened in TCGA-PAAD dataset using univariate Cox regression and lasso regression. The eight genes were permuted and combined, and the AUC value of each combination was calculated separately. The optimal AUC value is then screened using the Gaussian model and validated in the validation set. Ultimately, we found five genes that were excellent in evaluating the prognosis of pancreatic cancer patients in both the training set and validation set.

## Materials and methods

### Data collection and processing

The gene expression data were obtained from four public datasets, including TCGA, Pancreatic Cancer (PAAD) ( $n=182$ ), GTEx ( $n=167$ ), GSE62452 ( $n=130$ ) and GSE28735 ( $n=90$ ). PAAD was downloaded from UCSC Xena (<http://xena.ucsc.edu/>), the expression data

was normalized to  $\log_2(\text{FPKM}+1)$ . GTEx pancreatic cancer expression data was downloaded from UCSC Xena (<http://xena.ucsc.edu/>) and was normalized to  $\log_2(\text{FPKM}+0.001)$ . GSE62452 and GSE28735 were downloaded from GEO database (<https://www.ncbi.nlm.nih.gov/geo/>), the expression data was normalized by RMA. The expression levels of TCGA-PAAD and GTEx were rescaled using FPKM as the unit of measure for subsequent analyses. Differentially expressed genes (DEGs) was identified using Limma package (version 3.51.8) in R. The cut-off value was set to  $|\log_2\text{FC}|>1$  and  $p$  value  $<0.05$ .

### Screening for prognosis related DEGs

The univariate Cox regression, LASSO Cox regression models were performed with survival (version 3.1.1) and glmnet (version 4.1.4) R packages. The data screening process was conducted on TCGA-PAAD dataset.

### Data normalization and removing batch effects

The TCGA-PAAD gene expression values were then transformed to FPKM. FPKM were then transformed to TPM. After removing the mRNAs with low expression levels, the TCGA expression level were closer to gene chips. ComBat algorithm which is included in the sva R package (version 3.44.0) was used to remove batch effects.

### Gaussian mixture model

All selected prognosis-related genes were permuted and combined, and used multivariate Cox regression analysis to model, respectively. Calculated AUC for each model separately. AUC value was used as the classification basis and classification was conducted with model-based hierarchical agglomerative clustering which is based on Gaussian finite mixture model. This process used the mclust R package (version 5.4.9) which is a contributed R package for model-based clustering, classification, and density estimation based on finite normal mixture modelling. The associations between prognosis-related gene expression levels and survival information were estimated by the Kaplan–Meier method. The cut-off value was calculated with Kaplan–Meier method.

### Statistical analyses

All the statistical analysis was performed in R software (version 4.2.0). In the survival analysis, Cox proportional hazards regression and Kaplan–Meier analysis were used. We conducted paired t tests on paired samples. All statistical tests with a p-value of less than 0.05 were considered significant.

## Results

### Identification of DEGs in TCGA, GTEX, GSE62452 and GSE28735

This study was conducted according to the flowchart in Fig. 1. TCGA-PAAD, GTEX, GSE62452, and GSE28735 were used to identify DEGs. Because of the lack of normal samples in TCGA-PAAD data, we combined TCGA-PAAD and GTEX-PAAD for analysis. The cut-off value is  $|\log_{2}FC| > 1$  and  $p$ -value  $< 0.05$ . As shown in Fig. 1A, there are 178 tumor samples and 4 normal samples in TCGA-PAAD dataset. 167 normal samples in GTEX-PAAD dataset, 45 tumor samples and 45 normal samples in GSE28735, 69 tumor samples and 61 normal samples in GSE62452. The TCGA-PAAD and GTEX-PAAD are RNA-seq data, the expression data in them were normalized by FPKM. GSE28735 and GSE62452 share the same platform which ID is GPL6244. The expression data in the array chips was normalized into RAM. Limma R package was used for differential expression analysis. After differential expression analysis, 208 mRNAs were selected by taking the intersection of these datasets (Fig. 1A). These results suggested that 208 mRNAs have the same expression trends in different datasets (TCGA-PAAD, GTEX-PAAD, GSE62452, and GSE28735). Thus, these initially screened genes might be a promising parameter in patients with pancreatic cancer.

### Identification of prognosis related genes

Firstly, univariate Cox analysis was performed by using TCGA-PAAD expression and clinical data. After excluding genes with  $P$  values greater than 0.05, 100 prognostic related genes were screened. To further conduct variable selection and regularization, lasso-penalized Cox analysis was used (Supplementary Fig. 2). The complexity adjustment in the lasso regression algorithm refers to controlling the complexity of the model through a series of parameters to avoid overfitting. As shown in Fig. 1A and Table 1, total 8 candidate genes were selected, which were prognosis related genes. Among these genes, ankyrin repeat domain 22 (ANKRD22), aryl hydrocarbon receptor nuclear translocator like 2 (ARNTL2), desmoglein 3 (DSG3), integrin subunit beta 6 (ITGB6), keratin 7 (KRT7), MET proto-oncogene receptor tyrosine kinase (MET) are up-regulated. The remaining two genes, serine protease 3 (PRSS3) and thyrotropin releasing hormone degrading enzyme (TRHDE), are downregulated (Fig. 1A). The expression trends of these 8 genes were consistent in the 3 datasets.

These results confirmed that 8 candidate genes were associated with the prognosis of pancreatic cancer patients. Further analysis of 8 candidate genes might lead to better tools for assessing prognosis.

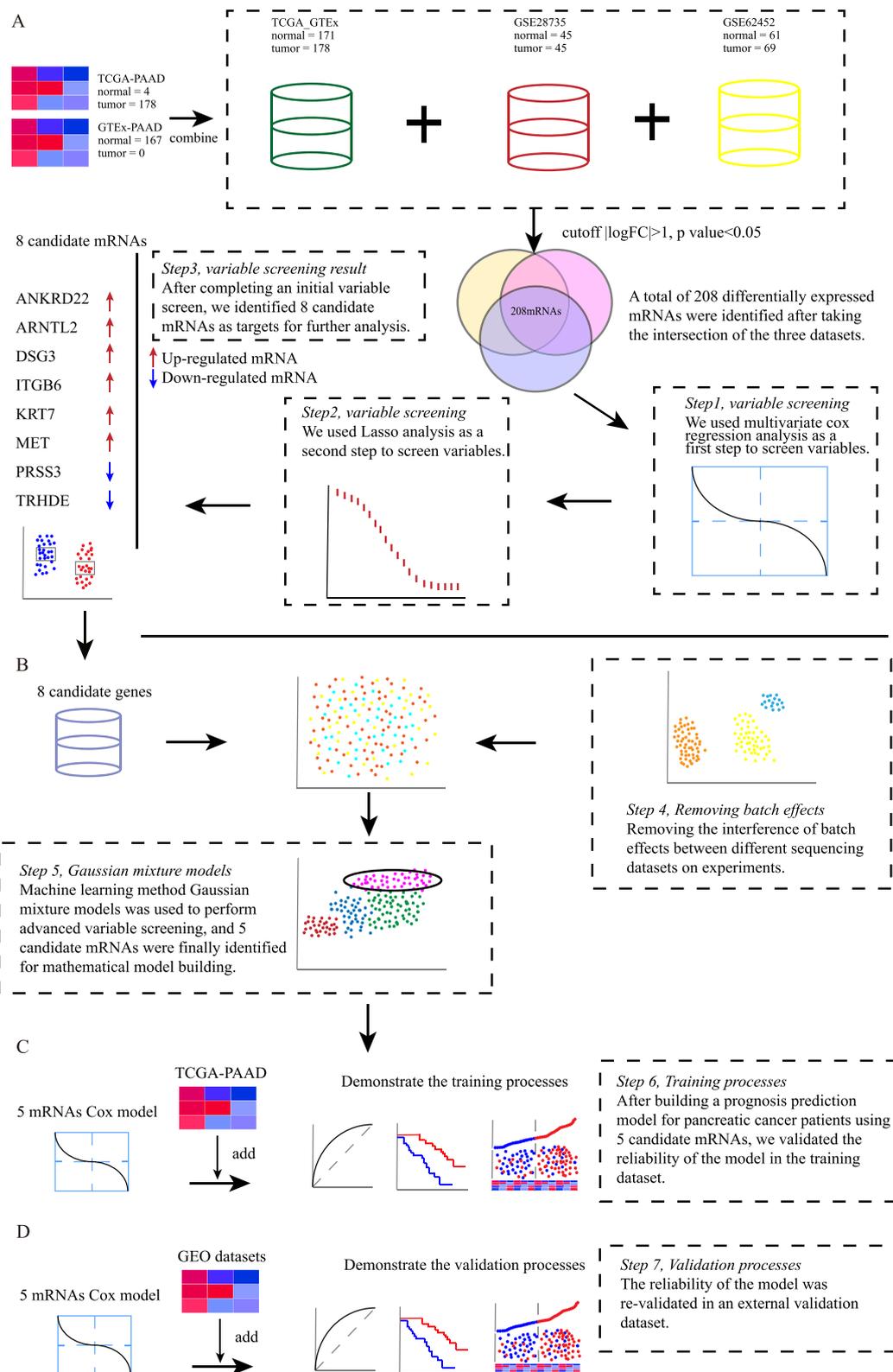
### Removing batch effects

Due to various factors such as experimental personnel, technology, environment, time point, chip processing, etc., the chip expression matrix contains differences in nonbiological factors. Especially when two or more datasets are integrated, even on the same platform, the integration is even more magnified. After adjusting the magnitude of TCGA-PAAD, we used the combat R function to remove the batch effects of these datasets (TCGA, GTEX, GSE62452, GSE28735). As shown in Fig. 2, after de-batching effects, the expression data of the 3 datasets were merged. The expression heatmaps and boxplots generated from the datasets after the batch effect has been removed are shown in Fig. 3.

These results suggested that 8 candidate genes had the same expression level in the 3 datasets.

### Gaussian mixture and Multivariate cox regression model

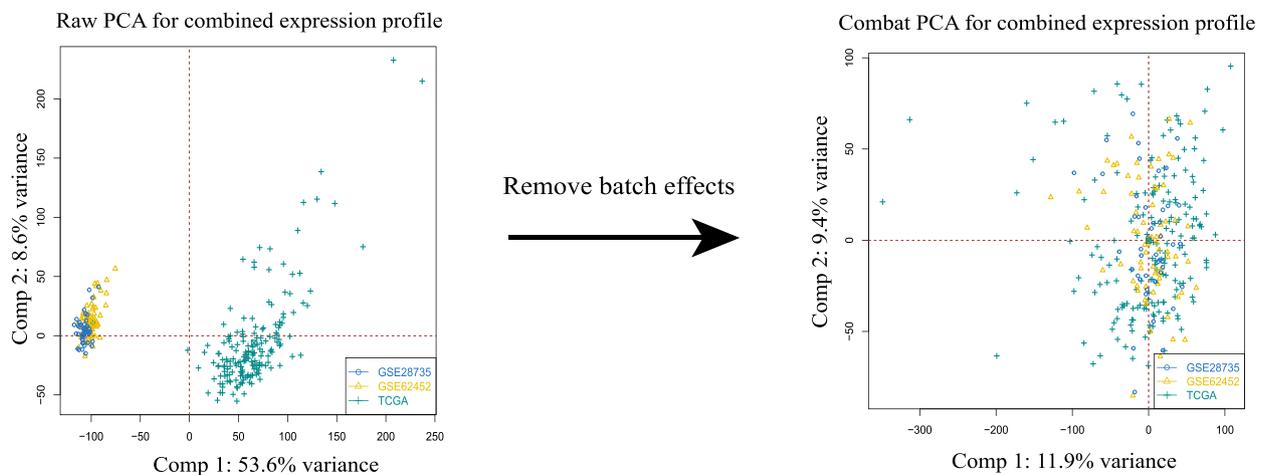
To further fit the model and predict patient prognosis more accurately, we conducted a classification based on Gaussian finite mixture model. After permutation and combination of 8 genes, a total of 255 multivariate Cox regression models were generated. All these models were evaluated for prediction accuracy using AUC. As shown in Fig. 4, 255 models were divided into 4 clusters, and cluster 4 had the highest AUCs. Model 168 that contained in cluster 4 had the highest AUC value of 0.755 (Supplementary table 1). The 5 genes contained in model 168 are ANKRD22, ARNTL2, DSG3, KRT7, and PRSS3. The results of multivariate regression analysis are shown in Table 2. Multivariate regression analysis showed that ANKRD22 (HR: 1.22, 95% CI 0.96–1.55;  $P=0.098$ ), ARNTL2 (HR: 1.71, 95% CI 1.19–2.46;  $P=0.003$ ), DSG3 (HR: 1.15, 95% CI 1.01–1.31;  $P=0.029$ ), KRT7 (HR: 1.19, 95% CI 0.97–1.46;  $P=0.0093$ ), PRSS3 (HR: 1.23, 95% CI 1.04–1.44;  $P=0.013$ ) were prognostic factors in pancreatic cancer patients. Survival analysis was performed on this 5-mRNA signature using TCGA training dataset and GSE validation datasets. Kaplan–Meier curves are shown in Fig. 5. According to the multivariate cox regression analysis results, the calculation formula of risk score is  $[0.19956 * \text{Exp}(\text{ANKRD22})] + [0.538767 * \text{Exp}(\text{ARNTL2})] + [0.141913 * \text{Exp}(\text{DSG3})] + [0.17315 * \text{Exp}(\text{KRT7})] + [0.205059 * \text{Exp}(\text{PRSS3})]$ . The best cut-off value was calculated based on Kaplan–Meier analysis. The calculating result of cut-off value is shown in Supplementary Fig. 1A. By dividing the risk score according to its cut-off value (cut-off=1.6), 176 patients were stratified into high-risk ( $n=61$ ) and low-risk ( $n=115$ ) groups (Fig. 6A). Kaplan–Meier curves showed that high-risk patients had worse survival outcome ( $P<0.0001$ ) (Fig. 6B). Additional ROC curves revealed that 5-mRNA signature model had the best AUC value (0.755),



**Fig. 1** Flowchart for the identification of 5-gene signature. **A** Eight genes associated with the prognosis of pancreatic cancer patients were screened by univariate cox regression and lasso regression. **B** The 5-gene signature was built with using of Gaussian finite mixture model. **C** The processes of training. **D** The processes of validation

**Table 1** Table of differential analysis results for 8 candidate genes in the 3 datasets

| TCGA and GTEx |                  |          |        | GSE62452 |                  |          |        | GSE28735 |                  |          |        |
|---------------|------------------|----------|--------|----------|------------------|----------|--------|----------|------------------|----------|--------|
| Gene          | log(fold-change) | p-value  | change | Gene     | log(fold-change) | p-value  | change | Gene     | log(fold-change) | p-value  | change |
| ANKRD22       | 1.28             | 4.38E-22 | up     | ANKRD22  | 1.14             | 2.78E-08 | up     | ANKRD22  | 1.58             | 1.10E-10 | up     |
| ARNTL2        | 1.94             | 2.39E-54 | up     | ARNTL2   | 1.08             | 3.90E-12 | up     | ARNTL2   | 1.26             | 9.10E-11 | up     |
| DSG3          | 1.14             | 6.86E-24 | up     | DSG3     | 1.04             | 1.97E-06 | up     | DSG3     | 1.01             | 1.17E-04 | up     |
| ITGB6         | 1.91             | 1.80E-37 | up     | ITGB6    | 2.02             | 2.66E-13 | up     | ITGB6    | 2.08             | 2.98E-11 | up     |
| KRT7          | 2.42             | 6.59E-36 | up     | KRT7     | 1.49             | 3.49E-10 | up     | KRT7     | 1.68             | 1.27E-08 | up     |
| MET           | 2.45             | 2.22E-49 | up     | MET      | 1.35             | 2.14E-13 | up     | MET      | 1.48             | 5.01E-12 | up     |
| PRSS3         | -4.86            | 2.34E-47 | down   | PRSS3    | -1.25            | 4.34E-06 | down   | PRSS3    | -1.21            | 1.02E-04 | down   |
| TRHDE         | -1.46            | 6.39E-48 | down   | TRHDE    | -1.61            | 1.77E-08 | down   | TRHDE    | -1.68            | 5.25E-07 | down   |

**Fig. 2** PCA diagrams for the sample distribution of the 3 datasets before and after removing batch effects

compared to ANKRD22 (0.655), ARNTL2 (0.688), DSG3 (0.671), KRT7 (0.696), PRSS3 (0.652) (Fig. 6C).

As noted above, these results suggested that 5-mRNA signature model was able to effectively predict the survival of pancreatic patients in the training dataset (TCGA-PAAD). The 5-mRNA signature model was better than the single gene prediction model.

#### External validation of 5-mRNA signature model

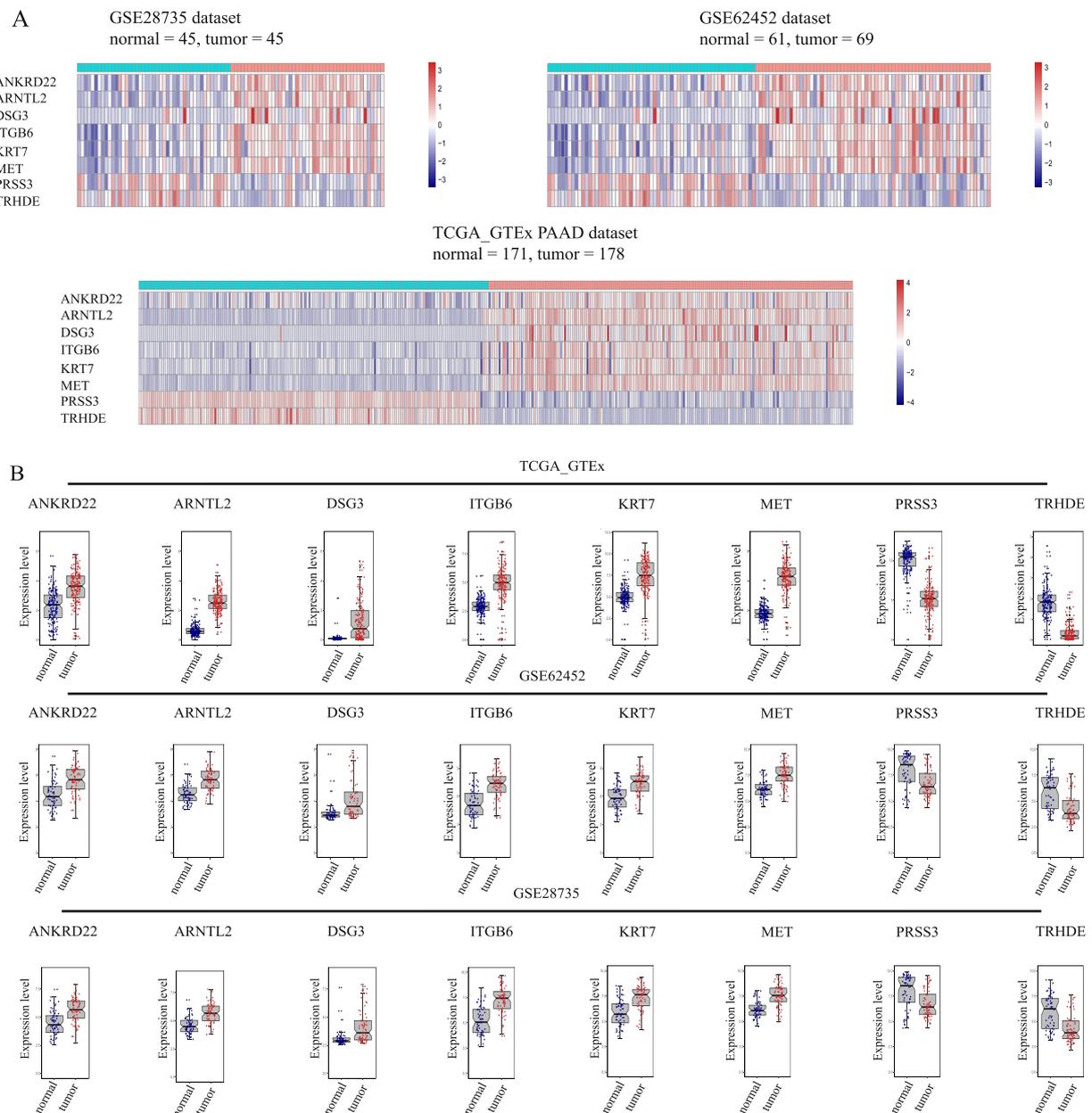
Two GEO datasets, GSE62452 and GSE28735, were used for external validation. The sample size of a single dataset is limited. To improve the prediction accuracy, the two datasets were combined after removing batch effects. Risk scores were calculated with the same formula for each patient. The best cut-off value was calculated based on Kaplan–Meier analysis (supplementary Fig. 1B). Patients were divided into high-risk ( $n=78$ ) and low-risk ( $n=29$ ) groups according to the optimal cut-off value (Fig. 7A). Kaplan–Meier analysis with the 5-mRNA

signature was used to compare the survival outcomes of patients in high-risk and low-risk groups. As shown in Fig. 7B, we confirmed that a higher risk score was associated with shorter survival time. We further analyzed the AUC value in the validation dataset, which was 0.91, significantly higher than AUC in the training dataset (0.755) (Fig. 7C).

In summary, the 5-mRNA signature's accuracy in predicting the prognosis of pancreatic cancer patients was validated, and the model performed better in the validation dataset.

#### Discussion

In this study, we used TCGA and GEO datasets to construct 5-mRNA signature (ANKRD22, ARNTL2, DSG3, KRT7, PRSS3) that is associated with the prognosis of pancreatic cancer patients. The superiority of 5-mRNA signature was verified using the validation dataset. Our



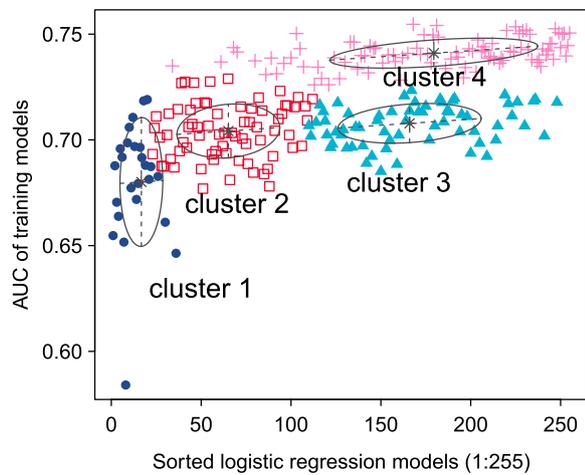
**Fig. 3** The expression levels of 8 candidate genes. **A** Heatmap. **B** Boxplot

results suggested that patients with higher risk score that calculated on 5-mRNA signature, had shorter survival time.

Ankyrin repeat domain 22 (ANKRD22), a novel mitochondrial membrane protein. Several studies show that the expression of ANKRD22 is significantly elevated in various tissues and cells. Such as macrophages of patients with an acute rejection reaction after a renal transplant [8], peripheral blood mononuclear cells of pancreatic cancer patients [9], basal type I basal-like breast cancer

tissues [10], non-small cell lung cancer (NSCLC) tissues [11]. In colorectal cancer cells, ANKRD22 plays a role in promoting glycolysis and reducing ATP levels [12]. Several studies suggest that the expression level of ANKRD22 is related to prognosis of pancreatic cancer [13], endometrial carcinoma [14], hepatocellular carcinoma [15].

Aryl hydrocarbon receptor nuclear translocator like 2 (ARNTL2), which encodes a basic helix-loop-helix transcription factor, is a member of PAS (PER ARNT, SIM)



**Fig. 4** The diagram of gaussian mixture model cluster analysis result. There are 4 clusters of 255 combinations

**Table 2** Multivariate COX regression analysis results of 5-gene signature

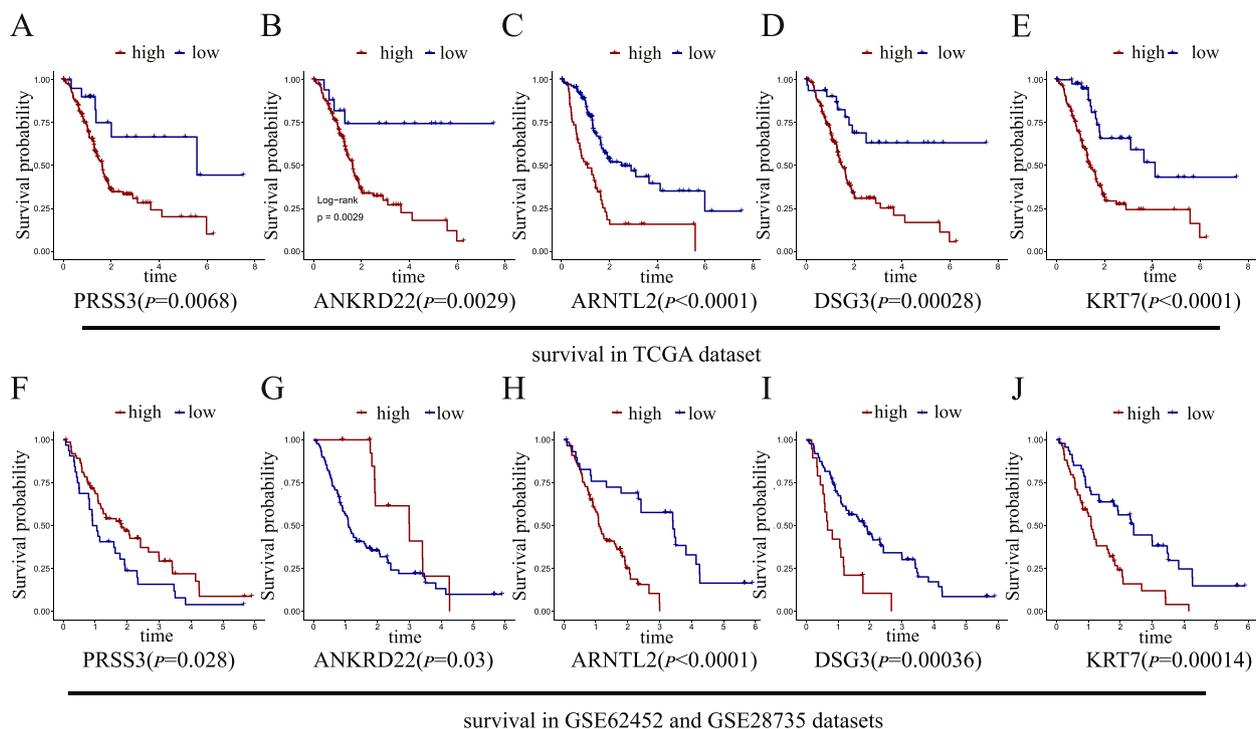
| Gene    | coefficient | HR   | 95% CI    | P        |
|---------|-------------|------|-----------|----------|
| ANKRD22 | 0.19956     | 1.22 | 0.96–1.55 | 0.098033 |
| ARNTL2  | 0.5387671   | 1.71 | 1.19–2.46 | 0.003633 |
| DSG3    | 0.141913    | 1.15 | 1.01–1.31 | 0.029027 |
| KRT7    | 0.17315     | 1.19 | 0.97–1.46 | 0.093604 |
| PRSS3   | 0.2050587   | 1.23 | 1.04–1.44 | 0.012608 |

superfamily. ARNTL2 plays a role in biological processes like hypoxia and circadian [16]. Serval malignant tumors, including lung adenocarcinoma [17], colorectal cancer [18], breast cancer [19], are associated with dysfunction of ARNTL2.

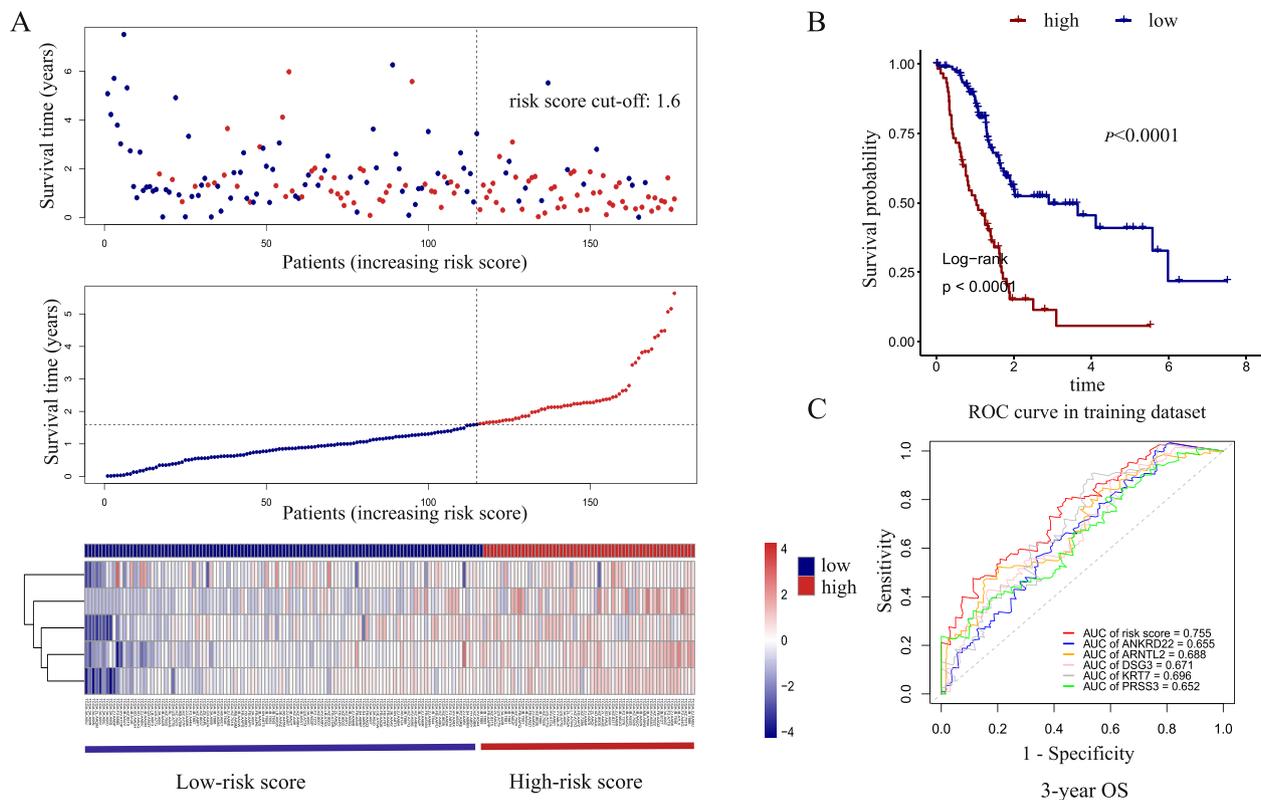
Desmoglein 3 (DSG3) is an adhesion protein in desmosomes and is a member of the cadherin superfamily. Recent studies identify that DSG3 is a key role in several pathways, like cell adhesion and proliferation, morphogenesis, differentiation and migration [20, 21]. Recent evidences suggest that DSG3 might play an important role in the prognostic assessment of head and neck squamous cell carcinoma [22], skin cutaneous melanoma [23], and triple negative breast cancer [24].

Integrin subunit beta 6 (ITGB6), as one of Integrins family, has an increased expression level in some biological processes like wound healing, fibrosis, and malignant tumor formation [25]. ITGB6 regulates many basic pathways of the cell, such as ECM degradation, proliferation [26]. ITGB6 tends to be identified as an oncogene, which is upregulated in several solid tumors, and is associated with poorer prognosis and increased invasiveness [27].

Thyrotropin releasing hormone degrading enzyme (TRHDE), the only downregulated gene in 5-mRNA signature, the protein translated by it, has the function as extracellular inactivation of TRH (Thyrotropin releasing hormone) [28]. However, the role in



**Fig. 5** Kaplan–Meier analysis estimates the OS of pancreatic cancer patients according to the expression levels of genes contained in 5-gene signature. **A** PRSS3 in TCGA. **B** ANKRD22 in TCGA. **C** ARNTL2 in TCGA. **D** DSG3 in TCGA. **E** KRT7 in TCGA. **F** PRSS3 in GSE. **G** ANKRD22 in GSE. **H** ARNTL2 in GSE. **I** DSG3 in GSE. **J** KRT7 in GSE



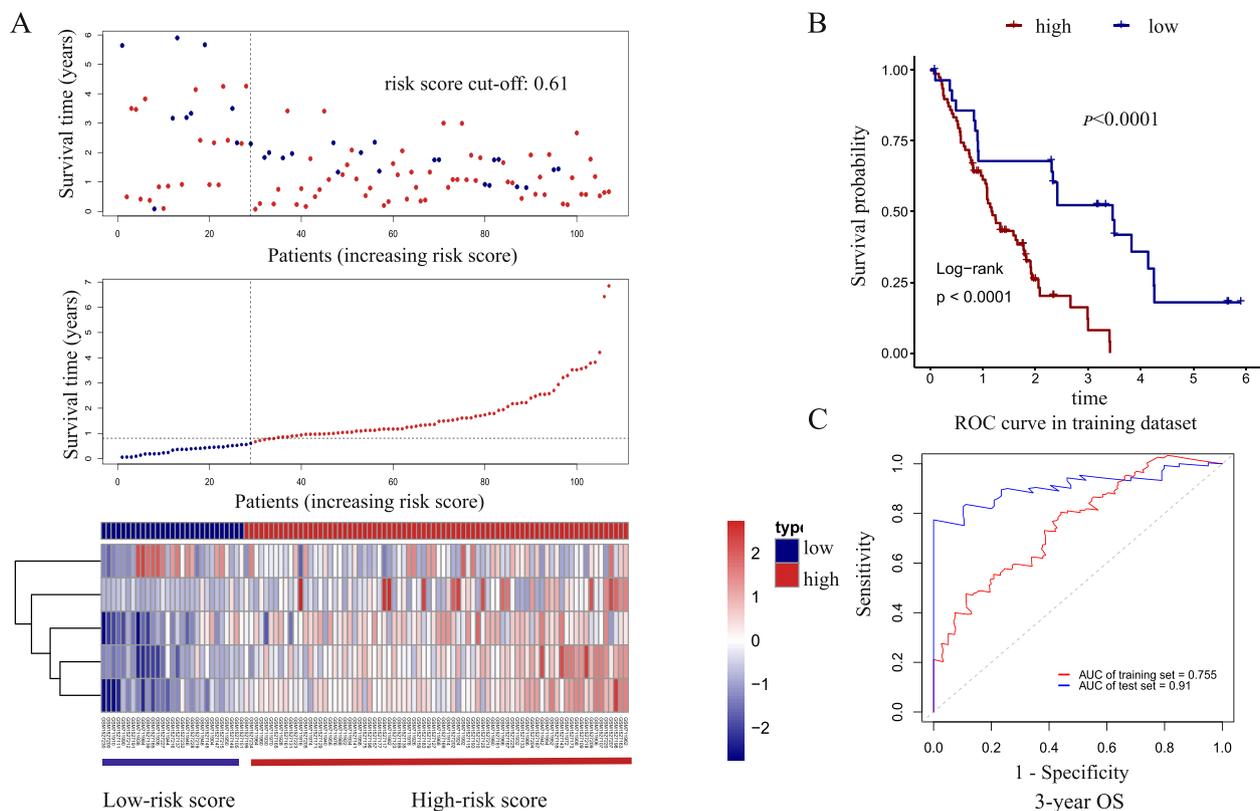
**Fig. 6** Survival predictive value of the 5-gene signature in TCGA-PAAD patients (training dataset). **A** Patients in TCGA-PAAD dataset were divided into high-risk score and low-risk score groups according to the risk scores calculated with 5-gene signature. **B** Kaplan–Meier analysis estimates the OS of high-risk and low-risk groups in training dataset. **C** ROC curves for 3-year OS by the combined or respective genes

cancer has not been elucidated. Only limited studies have demonstrated its role in tumor prognosis [29].

In clinical work, clinicians tend to use tumor TMN stage to evaluate the prognosis of pancreatic cancer patients. With the development of imaging technology, prognostic assessment combined with imaging data is also a feasible method. In any case, the current prognostic assessment method requires a high level of diagnosis capability for doctors and requires a certain amount of time for learning and training. At the beginning of the project, we wanted to find a simple, low-cost, universally adaptable way to perform prognostic assessment. Our prognostic assessment model includes 5 genes, making it easier and cheaper to test. By calculating the risk score, it is easier for clinicians to assess the prognosis of patients, to make clinical decisions and drug selection.

Since the prognosis evaluation of pancreatic cancer patients is important for the treatment of pancreatic cancer patients, many studies have focused on the role of prognosis-related genes in the prognosis evaluation of

pancreatic cancer. Luo et al. identified 7-gene signature (ARNTL2, DSG3, PTPRR, ANLN, S100A14, ANKRD22, and TSPAN7) by using of TCGA, ICGC and GEO datasets. The assessment of 7-gene signature was carried out using ROC curves, which is same to our study [13]. Wu et al. conducted a 5-gene signature (AADAC, DEF8, HIST1H1C, MET, and CHFR) which was potential molecular targets for overall surviving of resectable pancreatic cancer patients [30]. Not only gene expression data, but DNA methylation data can also be used to evaluate the prognosis of pancreatic cancer patients, and it has achieved good results in the prognosis evaluation of pancreatic cancer patients [31]. On the other hand, noncoding RNA expression data can also be used to assess the prognosis of pancreatic cancer patients with the same accuracy [32]. Our study not only used lasso regression and multivariate Cox regression commonly used by other researchers, but also used a Gaussian mixed model to further screen variables. The results of our study were evaluated using the ROC curve and showed that 5-gene signature had good performance in



**Fig. 7** External validation of the 5-gene signature in GSE dataset (combined GSE62452 and GSE28735). **A** Patients in GSE dataset were divided into high-risk score and low-risk score groups according to the risk scores calculated with 5-gene signature. **B** Kaplan–Meier analysis estimates the OS of high-risk and low-risk groups in validation dataset. **C** ROC curves for 3-year OS by the 5-gene signature in training and validation dataset

both the training set and the validation set. The comparison of AUC values showed that our 5-gene signature was superior and comparable to previous studies and has not been reported by other researchers yet. This shows that our prognostic gene screening method is superior and provides a better model for pancreatic cancer prognosis evaluation.

However, our study also has limitations, the ultimate of which is that we did not use a larger external validation dataset for testing, all validation datasets are from public databases. In addition, we did not explore the biological functions of the five genes, which need to be verified by further *in vivo* and *in vitro* experiments, which is also the focus of our future research.

## Conclusion

Our study used a novel prognostic-related gene screening method and identified 5-gene signature as a prognostic assessment model for pancreatic cancer patients. This 5-gene signature performed well on both our chosen training dataset and validation dataset. These results provided a new way to predict the prognosis of pancreatic cancer patients.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12876-023-02700-y>.

**Additional file 1.**

**Additional file 2.**

**Additional file 3.**

## Acknowledgements

We would like to thank the peer reviewers and editors for their work, which helped us significantly improve the quality of the manuscript.

## Authors' contributions

Xuanfeng Zhang wrote the main manuscript text. Lulu Yang and Dong Zhang completed data analysis. Xiaochuan Wang and Xuefeng Bu prepared figures 1-7. Xinhui Zhang prepared supplementary material. Long Cui revised the manuscript. The author(s) read and approved the final manuscript.

## Funding

This work is supported by the Xuzhou Science and Technology Bureau [grant numbers KC20124].

## Availability of data and materials

The expression data of TCGA pancreatic cancer patients were downloaded from UCSC Xena (<https://xenabrowser.net/datapages/>). GTEx pancreatic cancer expression data was downloaded from UCSC Xena (<https://xenabrowser.net/datapages/>). GSE62452 and GSE28735 were downloaded from Gene Expression Omnibus database (<https://www.ncbi.nlm.nih.gov/geo/>).

## Declarations

### Ethics approval and consent to participate

Ethical approval was not required for this study because we used a public database.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

Received: 12 November 2022 Accepted: 27 February 2023

Published online: 11 March 2023

## References

- Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin.* 2022;72:7–33.
- Mizrahi JD, Surana R, Valle JW, Shroff RT. Pancreatic cancer. *Lancet.* 2020;395:2008–20.
- Wu M, Li X, Zhang T, Liu Z, Zhao Y. Identification of a nine-gene signature and establishment of a prognostic nomogram predicting overall survival of pancreatic cancer. *Front Oncol.* 2019;9:996.
- Kamarajah SK, Burns WR, Frankel TL, Cho CS, Nathan H: Validation of the American Joint Commission on Cancer (AJCC) 8th Edition Staging System for Patients with Pancreatic Adenocarcinoma: A Surveillance, Epidemiology and End Results (SEER) Analysis. *Ann Surg Oncol.* 2017;24:2023–2030.
- Xuan J, Yu Y, Qing T, Guo L, Shi L. Next-generation sequencing in the clinic: promises and challenges. *Cancer Lett.* 2013;340:284–95.
- Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med.* 2021;13:152.
- Sui S, An X, Xu C, Li Z, Hua Y, Huang G, Sui S, Long Q, Sui Y, Xiong Y, et al. An immune cell infiltration-based immune score model predicts prognosis and chemotherapy effects in breast cancer. *Theranostics.* 2020;10:11938–49.
- Venner JM, Famulski KS, Badr D, Hidalgo LG, Chang J, Halloran PF. Molecular landscape of T cell-mediated rejection in human kidney transplants: prominence of CTLA4 and PD ligands. *Am J Transplant.* 2014;14:2565–76.
- Caba O, Prados J, Ortiz R, Jimenez-Luna C, Melguizo C, Alvarez PJ, Delgado JR, Irigoyen A, Rojas I, Perez-Florido J, et al. Transcriptional profiling of peripheral blood in pancreatic adenocarcinoma patients identifies diagnostic biomarkers. *Dig Dis Sci.* 2014;59:2714–20.
- Milioli HH, Tishchenko I, Riveros C, Berretta R, Moscato P. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC Med Genomics.* 2017;10:19.
- Yin J, Fu W, Dai L, Jiang Z, Liao H, Chen W, Pan L, Zhao J. ANKRD22 promotes progression of non-small cell lung cancer through transcriptional up-regulation of E2F1. *Sci Rep.* 2017;7:4430.
- Pan T, Liu J, Xu S, Yu Q, Wang H, Sun H, Wu J, Zhu Y, Zhou J, Zhu Y. ANKRD22, a novel tumor microenvironment-induced mitochondrial protein promotes metabolic reprogramming of colorectal cancer cells. *Theranostics.* 2020;10:516–36.
- Luo L, Li Y, Huang C, Lin Y, Su Y, Cen H, Chen Y, Peng S, Ren T, Xie R, Zeng L. A new 7-gene survival score assay for pancreatic cancer patient prognosis prediction. *Am J Cancer Res.* 2021;11:495–512.
- Wang A, Guo H, Long Z. Integrative Analysis of Differently Expressed Genes Reveals a 17-Gene Prognosis Signature for Endometrial Carcinoma. *Biomed Res Int.* 2021;2021:4804694.
- Li J, Chen N, Gong X. Prognostic implications of aberrantly expressed methylation-driven genes in hepatocellular carcinoma: a study based on The Cancer Genome Atlas. *Mol Med Rep.* 2019;20:5304–14.
- Lebailly B, Langa F, Boitard C, Avner P, Rogner UC. The circadian gene *Arntl2* on distal mouse chromosome 6 controls thymocyte apoptosis. *Mamm Genome.* 2017;28:1–12.
- Brady JJ, Chuang CH, Greenside PG, Rogers ZN, Murray CW, Caswell DR, Hartmann U, Connolly AJ, Sweet-Cordero EA, Kundaje A, Winslow MM. An *Arntl2*-Driven Secretome Enables Lung Adenocarcinoma Metastatic Self-Sufficiency. *Cancer Cell.* 2016;29:697–710.
- Mazzoccoli G, Paziienza V, Panza A, Valvano MR, Benegiamo G, Vinciguerra M, Andriulli A, Piepoli A. *ARNTL2* and *SERPINE1*: potential biomarkers for tumor aggressiveness in colorectal cancer. *J Cancer Res Clin Oncol.* 2012;138:501–11.
- Ha NH, Long J, Cai Q, Shu XO, Hunter KW. The Circadian Rhythm Gene *Arntl2* Is a Metastasis Susceptibility Gene for Estrogen Receptor-Negative Breast Cancer. *PLoS Genet.* 2016;12:e1006267.
- Brown L, Wan H. Desmoglein 3: a help or a hindrance in cancer progression? *Cancers (Basel).* 2015;7:266–86.
- Mannan T, Jing S, Foroushania SH, Fortune F, Wan H. RNAi-mediated inhibition of the desmosomal cadherin (desmoglein 3) impairs epithelial cell proliferation. *Cell Prolif.* 2011;44:301–10.
- Bunbanjerdusuk S, Vorasan N, Saethang T, Pongrujijorn T, Pangpunyakulchai D, Mongkonsiri N, Arsa L, Thokanit N, Pongpachit W, Anekpuritanang T, et al. Oncoproteomic and gene expression analyses identify prognostic biomarkers for second primary malignancy in patients with head and neck squamous cell carcinoma. *Mod Pathol.* 2019;32:943–56.
- Sheng Z, Han W, Huang B, Shen G. Screening and identification of potential prognostic biomarkers in metastatic skin cutaneous melanoma by bioinformatics analysis. *J Cell Mol Med.* 2020;24:11613–8.
- Fei H, Chen S, Xu C. RNA-sequencing and microarray data mining revealing: the aberrantly expressed mRNAs were related with a poor outcome in the triple negative breast cancer patients. *Ann Transl Med.* 2020;8:363.
- Meecham A, Marshall JF. The *ITGB6* gene: its role in experimental and clinical biology. *Gene X.* 2020;5:100023.
- Thomas GJ, Lewis MP, Hart IR, Marshall JF, Speight PM.  $\alpha$ V $\beta$ 6 integrin promotes invasion of squamous carcinoma cells through up-regulation of matrix metalloproteinase-9. *Int J Cancer.* 2001;92:641–50.
- Niu J, Li Z. The roles of integrin  $\alpha$ v $\beta$ 6 in cancer. *Cancer Lett.* 2017;403:128–37.
- Charli JL, Rodriguez-Rodriguez A, Hernandez-Ortega K, Cote-Velez A, Uribe RM, Jaimes-Hoy L, Joseph-Bravo P. The Thyrotropin-Releasing Hormone-Degrading Ectoenzyme, a Therapeutic Target? *Front Pharmacol.* 2020;11:640.
- Qin W, Qi F, Li J, Li P, Zang YS. Prognostic biomarkers on a competitive endogenous RNA network reveals overall survival in triple-negative breast cancer. *Front Oncol.* 2021;11:681946.
- Wu C, Wu Z, Tian B. Five gene signatures were identified in the prediction of overall survival in resectable pancreatic cancer. *BMC Surg.* 2020;20:207.
- Deng GC, Sun DC, Zhou Q, Lv Y, Yan H, Han QL, Dai GH. Identification of DNA methylation-driven genes and construction of a nomogram to predict overall survival in pancreatic cancer. *BMC Genomics.* 2021;22:791.
- Wei C, Liang Q, Li X, Li H, Liu Y, Huang X, Chen X, Guo Y, Li J. Bioinformatics profiling utilized a nine immune-related long noncoding RNA signature as a prognostic target for pancreatic cancer. *J Cell Biochem.* 2019;120:14916–27.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

