Research Paper

# Development of a four-gene prognostic model for pancreatic cancer based on transcriptome dysregulation

**Jie Yan[1], Liangcai Wu[2], Congwei Jia[1], Shuangni Yu[1], Zhaohui Lu[1], Yueping Sun[3], Jie Chen[1]**

[1]Department of Pathology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China
[2]Department of Obstetrics and Gynecology, Obstetrics and Gynecology Hospital of Fudan University, Shanghai 200011, China
[3]Institute of Medical Information, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100020, China

**Correspondence to:** Jie Chen; **email:** chenjie@pumch.cn
**Keywords:** prognostic prediction model, pancreatic cancer, TCGA, robust rank aggregation, WGCNA
**Received:** November 2, 2019 **Accepted:** February 4, 2020 **Published:** February 20, 2020



## ABSTRACT

**We systematically developed a prognostic model for pancreatic cancer that was compatible across different transcriptomic platforms and patient cohorts. After performing quality control measures, we used seven microarray datasets and two RNA sequencing datasets to identify consistently dysregulated genes in pancreatic cancer patients. Weighted gene co-expression network analysis was performed to explore the associations between gene expression patterns and clinical features. The least absolute shrinkage and selection operator (LASSO) and Cox regression were used to construct a prognostic model. We tested the predictive power of the model by determining the area under the curve of the risk score for time-dependent survival. Most of the differentially expressed genes in pancreatic cancer were enriched in functions pertaining to the tumor immune microenvironment. The transcriptome profiles were found to be associated with overall survival, and four genes were identified as independent prognostic factors. A prognostic risk score was then proposed, which displayed moderate accuracy in the training and self-validation cohorts. Furthermore, patients in two independent microarray cohorts were successfully stratified into high- and low-risk prognostic groups. Thus, we constructed a reliable prognostic model for pancreatic cancer, which should be beneficial for clinical therapeutic decision-making.**

## INTRODUCTION

Risk stratification (also commonly called "prognostic modeling") is a useful tool in cancer management, since it enables timely interventions for high-risk patients while obviating unnecessary treatments for low-risk patients [1, 2]. Classical prognostic factors such as the clinical tumor-node-metastasis (cTNM) stage and pathological stage (pTNM) are not completely prognostically relevant in some patients [3–5]. Accordingly, the guidelines for prognostic assessments have been continually modified to improve their accuracy while reducing their complexity for daily clinical use [6–9].

Novel molecular factors such as immunoscores assessing *in situ* immune cell infiltration in tumors, abnormal DNA levels and mRNA levels are more accurate risk predictors than the existing tumor parameters [10–12]. High-throughput technologies provide an efficient means of measuring the molecular disruptions in tumors [13]. For example, a prognostic landscape of cancer was recently developed, which integrated the transcriptomes and clinical data of approximately 26,000 patients across 39 malignancies to establish the patterns and determinants of responses to targeted therapy [14]. Since numerous cancer-related microarrays and sequencing platforms have been generated in recent years, it is essential to integrate

the large amounts of available data and translate these molecular findings into clinical decision-making tools. To this end, clinical data from The Cancer Genome Atlas (TCGA) Pan-Cancer analysis project have been integrated [15], and genotype-to-phenotype databases have been developed [16] for clinical interpretation [17].

Pancreatic cancer has a dismal prognosis, with a five-year survival rate of only 9% [18]. It is characterized by desmoplastic stroma, perineural invasion [5], invasiveness and immune suppression [13], which are largely responsible for the early metastasis [19], chemoresistance [20] and cachexia [21] observed in patients. Based on the transcriptome data of pancreatic cancer cells, tumors can be classified into the squamous, pancreatic progenitor, aberrantly differential endocrine exocrine, and immunogenic subtypes [13]. The squamous subtype is associated with a poor prognosis, and the immunogenic subtype involves the upregulation of gene networks for acquired immune suppression. A better understanding of the molecular landscape of pancreatic cancer would enable the development of novel therapeutic strategies to improve clinical outcomes and facilitate the stratification of patients into prognostic groups to guide personalized treatment. However, a comprehensive prognostic model with compatibility across different transcriptomic platforms and patient cohorts has not been systematically developed.

To determine the prognostic significance of the pancreatic cancer transcriptome, we screened multiple RNA-Seq and microarray datasets for genes that were differentially expressed between normal and tumorous tissues, and identified genes that were significantly associated with overall survival. We then developed a prognostic risk score and successfully validated it in three independent pancreatic cancer cohorts. We thereby devised a prognostic model that can predict the post-surgical prognosis of pancreatic cancer patients with moderate accuracy.

## RESULTS

### Combined analyses of multiple pancreatic cancer microarray datasets

We searched the Gene Expression Omnibus (GEO) database for all the human tissue microarrays that included pancreatic cancer tissues and paired/unpaired normal pancreatic tissues. Then, we used Transcriptome Analysis Console software (Applied Biosystems, version 4.0.2) to evaluate the data for hybridization and labeling controls. Affy [22] was used to assess RNA degradation, and simpleAffy [23] was used to determine the 3'-to-5' ratios of *β-actin* and *GAPDH* (Supplementary Figure 2). Two pancreatic ductal adenocarcinoma (PDAC) datasets (GSE22780 and GSE27890) were thus excluded, and seven datasets (GSE32676, GSE16515, GSE71989, GSE41368, GSE15471, GSE28735 and GSE62452) were selected for further analysis (Table 1).

After seven cases were excluded from these datasets, the data of 177 normal pancreatic tissue samples and 226 PDAC tissue samples were included in subsequent analyses. A robust rank aggregation analysis [24] identified 616 differentially expressed genes (DEGs) between the normal and PDAC samples across all datasets, with an adjusted $p$ value < 0.05 and $|\log_2^{FC\ (fold\ change)}| > 1$ as the cut-offs. Among these genes, 403 were upregulated and 213 were downregulated in PDAC tissues. The heatmap displaying the top 10 significantly overexpressed or suppressed genes is shown in Figure 1A.

Gene Ontology (GO) analysis of the DEGs revealed significant enrichment in the GO terms for 158 biological processes (BPs), 26 cellular components (CCs) and 28 molecular functions (MFs) ($p$ value < 0.01 and $q$ value < 0.01 as cut-offs) (Figure 1B). The top BP terms were related to three aspects: i) extracellular stroma formation, including extracellular structure organization and extracellular matrix organization, which was not surprising, since stiffness is the defining characteristic of PDAC; ii) immune cell responses, such as the innate immune response, neutrophil activation and neutrophil mediated immunity; and iii) fundamental pancreatic functions, such as regulating pancreatic juice secretion and epithelial cell proliferation. Major CC terms included the extracellular matrix and various components of the intracellular lumen and the apical part of the cell. The most enriched MF terms were extracellular matrix formation, cell adhesion and receptor ligand activity.

Thus, through a combined analysis of seven high-quality GEO microarrays, we identified 616 genes that were consistently differentially expressed between normal pancreatic tissues and PDAC tissues. Most of the DEGs were associated with the pancreatic extra-cellular stroma and the tumor immune micro-environment.

### Combined analyses of TCGA and GTEx RNA-Seq datasets

To determine whether the DEGs were independent of the detection method, we also analyzed their expression in PDAC RNA-Seq datasets. Since TCGA only contains data from four normal pancreatic tissue samples [25], we also included normal tissue data from the Genotype-Tissue

**Table 1. Enrolled PDAC cases from seven GEO datasets after quality control.**

| Country | Organization name | Series | Platform | Normal | Tumor | Quality control | Publication |
|---|---|---|---|---|---|---|---|
| USA | University of Los Angeles | GSE32676 | GPL570 | 7 | 25 | Passed | [77] |
| USA | Mayo Clinic | GSE16515 | GPL570 | 16 | 36 | Passed | [78] |
| USA | University of Florida | GSE71989 | GPL570 | 8 | 13 | Excluded one non-tumor sample | [79] |
| Romania | ICI | GSE15471 | GPL570 | 35 | 36 | Excluded one normal tissue | [81] |
| Italy | Sapienza University of Rome | GSE41368 | GPL6244 | 6 | 6 | Passed | [80] |
| USA | NCI/NIH | GSE28735 | GPL6244 | 44 | 43 | Excluded one normal and two tumor samples | [82, 83] |
| USA | National Cancer Institute | GSE62452 | GPL6244 | 61 | 67 | Excluded two tumor samples | [84] |

ICI: National Institute for Research in Informatics

Expression (GTEx) database, which contains normal tissue samples from 54 human body sites and is maintained by The Broad Institute of MIT and Harvard [26, 27]. The GTEx and TCGA RNA-Seq data and phenotypic information were obtained from the University of California Santa Cruz (UCSC) Xena platform (https://xena.ucsc.edu/), which is routinely updated and integrated [28].

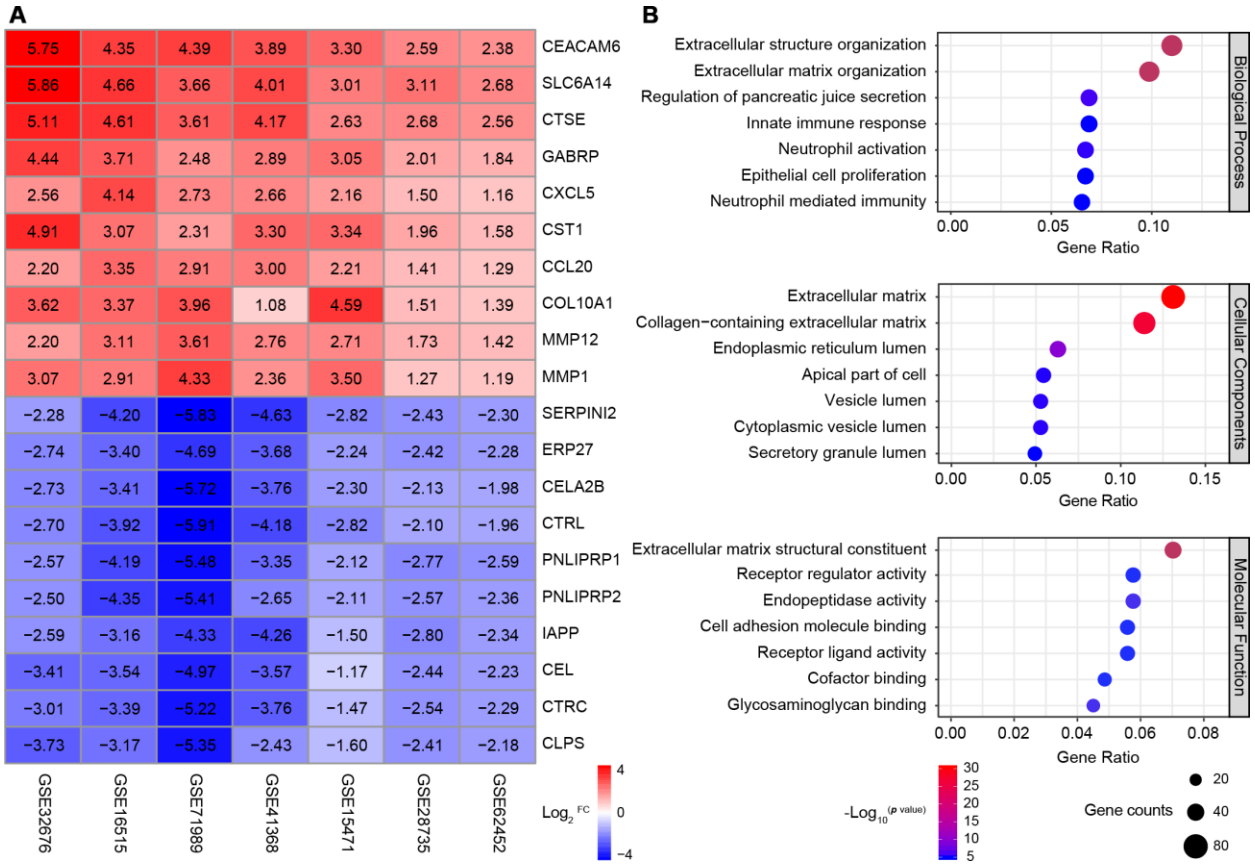



**Figure 1. Identification and GO analysis of DEGs in seven PDAC datasets.** (**A**) A heatmap of the top 10 significantly upregulated or downregulated genes. The expression of each gene in each dataset is shown in a colored box with the $log_2^{FC}$ inset. Red and blue boxes indicate upregulated and downregulated genes, respectively, and the color saturation correlates with the gene level. (**B**) GO analysis of all the DEGs. The seven most enriched GO terms in each category (BP, CC and MF) are listed next to the left axis. The size of the circle indicates the number of enriched genes, and the color is associated with the respective $-log_{10}{}^{p\ value}$.

In total, 171 normal and 178 pancreatic cancer samples were analyzed, and 5,886 DEGs were identified by the same cut-off criteria used to analyze the PDAC microarrays ($|\log_2^{FC}| > 1$ and false discovery rate < 0.05). Among these genes, 2,980 were upregulated and 2,906 were downregulated in pancreatic cancer tissues relative to normal pancreatic tissues. These DEGs were enriched in 600 BP, 106 CC and 32 MF terms, and the most significantly enriched BP terms were related to leukocyte function, extracellular matrix formation, inflammation, etc. (Figure 2A). Consistent with the DEGs in the microarrays, most of the genes were enriched in the extracellular matrix or junctions for adhesion and antigen-binding functions (Figure 2B). Thus, both sequencing and multi-microarray data indicated that the tumor immune and stroma microenvironment was significantly disturbed during pancreatic tumorigenesis.

### Identification of 542 genes that were consistently differentially expressed across independent platforms

We next investigated whether the 616 DEGs identified by the seven microarrays were consistent with the 5,886 DEGs identified by PDAC transcriptome sequencing. When we compared the DEG profiles obtained by these two methods, we detected 542 common genes. Surprisingly, all of the overlapping genes displayed consistent expression trends in the two types of profiles (Supplementary Table 1).

Since transcription factors (TFs) and kinases are key components of cancer regulatory networks and are preferred targets for drug development [29, 30], we cross-referenced the DEGs with both the Cistrome Cancer human TF database [31] and a list of 518 human kinases [32]. Of the DEGs, 19 displayed TF activity (Table 2) and 16 were kinases, of which six belonged to the Tyrosine Kinase group (Table 3). Thus, we identified 542 pancreatic-cancer-related genes that were consistently dysregulated in both multi-microarray datasets and sequencing datasets, including 35 well-defined protagonists harboring core regulatory functions.

### The dysregulated transcriptome is associated with overall survival in pancreatic cancer patients

To determine the phenotypic relevance of the DEGs, we performed a weighted gene co-expression network analysis (WGCNA) [33] to identify gene modules

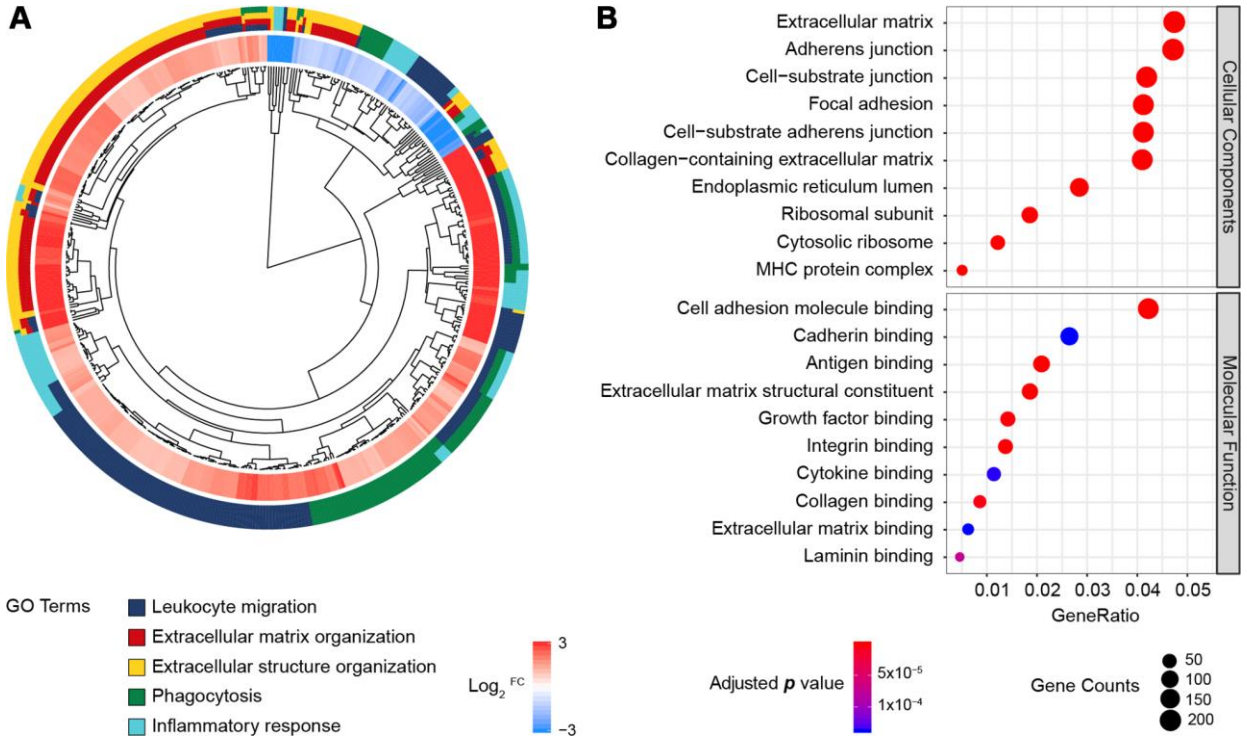



**Figure 2. GO enrichment analysis of the DEGs from the RNA-Seq datasets of TCGA and GTEx.** (**A**) GO Cluster. The inner dendrogram indicates the hierarchical clustering of the gene expression profiles, the outer circle represents the $\log_2^{FC}$ of each DEG, with the color corresponding to the gene level, and the outermost circle represents the GO BP terms assigned to the gene. (**B**) The 10 most significantly enriched CC and MF terms. The size of a circle indicates the number of enriched genes, and its color corresponds to the adjusted *p* value.

**Table 2. The 19 differentially expressed TFs in pancreatic cancer.**

| Gene | Seven GEO datasets | | | TCGA combined GTEx datasets | | |
|---|---|---|---|---|---|---|
| | $Log_2^{FC}$ | *p* value | Adjusted *p* value | $Log_2^{FC}$ | *p* value | FDR |
| *GCG* | -1.97 | 1.96E-11 | 4.90E-07 | -1.13 | 0.038 | 0.039 |
| *FOXQ1* | 1.61 | 5.62E-09 | 1.41E-04 | 3.39 | 1.54E-51 | 7.57E-51 |
| *DKK1* | 2.29 | 1.65E-12 | 4.13E-08 | 2.47 | 1.36E-42 | 3.29E-42 |
| *KLF5* | 1.46 | 1.79E-07 | 4.49E-03 | 3.17 | 7.69E-48 | 2.49E-47 |
| *AHR* | 1.16 | 7.41E-08 | 1.85E-03 | 2.96 | 3.53E-56 | 1.30E-54 |
| *ARNTL2* | 1.61 | 5.71E-10 | 1.43E-05 | 1.94 | 2.39E-54 | 3.24E-53 |
| *ID1* | 1.25 | 9.58E-09 | 2.40E-04 | 2.15 | 3.63E-43 | 9.03E-43 |
| *PPARG* | 1.27 | 8.74E-08 | 2.19E-03 | 2.00 | 4.27E-47 | 1.31E-46 |
| *LEF1* | 1.45 | 4.60E-08 | 1.15E-03 | 1.80 | 2.38E-49 | 8.74E-49 |
| *PTTG1* | 1.13 | 4.99E-09 | 1.25E-04 | 2.06 | 1.29E-53 | 1.15E-52 |
| *MXD1* | 1.07 | 2.05E-08 | 5.13E-04 | 1.68 | 3.89E-51 | 1.79E-50 |
| *DTL* | 1.22 | 1.18E-10 | 2.94E-06 | 1.13 | 7.81E-53 | 5.14E-52 |
| *ETV1* | 1.04 | 9.97E-07 | 2.49E-02 | 1.29 | 7.00E-52 | 3.63E-51 |
| *ZNF521* | 1.19 | 1.99E-06 | 4.98E-02 | 1.05 | 1.29E-41 | 3.00E-41 |
| *NCAPG* | 1.01 | 1.14E-07 | 2.85E-03 | 1.07 | 8.97E-49 | 3.14E-48 |
| *BCAT1* | -1.12 | 1.91E-08 | 4.78E-04 | -1.51 | 1.91E-36 | 3.79E-36 |
| *ZBTB16* | -1.51 | 6.22E-10 | 1.56E-05 | -1.88 | 5.90E-43 | 1.45E-42 |
| *NR5A2* | -2.12 | 3.54E-14 | 8.86E-10 | -2.27 | 1.65E-48 | 5.66E-48 |
| *CTRL* | -3.37 | 2.22E-19 | 5.55E-15 | -8.38 | 8.32E-55 | 1.45E-53 |

**Table 3. The 16 differentially expressed kinases in pancreatic cancer.**

| Entrez ID | Gene | SK | Group | Family | Seven GEO datasets | | | TCGA combined GTEx datasets | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $Log_2^{FC}$ | *p* value | Adjusted *p* value | $Log_2^{FC}$ | *p* value | FDR |
| 1969 | *EPHA2* | SK122 | TK | Eph | 1.02 | 1.10E-06 | 0.028 | 2.77 | 1.21E-48 | 4.18E-48 |
| 4233 | *MET* | SK227 | TK | Met | 1.17 | 1.07E-06 | 0.027 | 2.45 | 2.22E-49 | 8.18E-49 |
| 55359 | *STYK1* | SK530 | TK | TK-Unique | 1.47 | 3.72E-09 | 9.30E-05 | 1.85 | 1.25E-53 | 1.13E-52 |
| 4486 | *MST1R* | SK332 | TK | Met | 1.58 | 1.10E-07 | 0.003 | 1.66 | 4.74E-37 | 9.57E-37 |
| 9833 | *MELK* | SK298 | CAMK | CAMKL | 1.54 | 3.07E-10 | 7.67E-06 | 1.57 | 2.93E-50 | 1.19E-49 |
| 983 | *CDK1* | SK065 | CMGC | CDK | 1.03 | 1.99E-07 | 0.005 | 1.86 | 5.32E-52 | 2.82E-51 |
| 4751 | *NEK2* | SK251 | Other | NEK | 1.08 | 1.05E-08 | 0.000 | 1.61 | 5.66E-53 | 3.91E-52 |
| 9448 | *MAP4K4* | SK437 | STE | STE20 | 1.01 | 2.85E-07 | 0.007 | 1.56 | 3.46E-53 | 2.55E-52 |
| 55872 | *PBK* | SK529 | Other | TOPK | 1.10 | 3.76E-08 | 0.001 | 1.35 | 1.74E-53 | 1.47E-52 |
| 9891 | *NUAK1* | SK195 | CAMK | CAMKL | 1.02 | 5.09E-08 | 0.001 | 1.38 | 3.78E-50 | 1.52E-49 |
| 701 | *BUB1B* | SK053 | Other | BUB | 1.28 | 3.42E-09 | 8.56E-05 | 1.10 | 5.88E-49 | 2.09E-48 |
| 2043 | *EPHA4* | SK124 | TK | Eph | 1.06 | 2.30E-07 | 0.006 | 1.27 | 5.14E-48 | 1.69E-47 |
| 4915 | *NTRK2* | SK378 | TK | Trk | -1.06 | 4.81E-07 | 0.012 | -1.06 | 9.65E-30 | 1.65E-29 |
| 5166 | *PDK4* | SK280 | Atypical | PDHK | -1.76 | 6.11E-13 | 1.53E-08 | -1.49 | 5.54E-19 | 7.88E-19 |
| 5063 | *PAK3* | SK269 | STE | STE20 | -1.59 | 2.81E-12 | 7.02E-08 | -2.09 | 8.26E-45 | 2.22E-44 |
| 8569 | *MKNK1* | SK235 | CAMK | MAPKAPK | -1.01 | 1.82E-08 | 0.000 | -4.04 | 2.02E-55 | 4.75E-54 |

(groups of highly interconnected genes) that were significantly associated with the clinico-pathological features of pancreatic cancer. Since sufficient sample sizes and adequate phenotypic data (including prognostic data) are prerequisites for analyzing gene co-expression networks that are associated with clinical characteristics, we only extracted the gene expression profiles and corresponding clinical data of the PDAC patients from TCGA who met these criteria (N = 135). Fourteen clusters (modules) of highly interconnected genes with co-expression similarity values > 0.75 for the module eigengenes were identified (Supplementary Figure 3, Supplementary Table 2).

Since we had already identified the characteristic gene co-expression profiles of each module, we searched for significant correlations between the module eigengenes and clinical traits. Overall survival status was associated with three modules, although the Pearson correlations were weak; for example, the black module correlated positively ($r$ = 0.27, $p$ = 0.001) and the pink module correlated negatively with overall survival ($r$ = -0.26, $p$ = 0.002) (Figure 3). Additionally, age and tumor grade were associated with one module each, while the other four clinical traits we examined (gender, tumor stage, T classification and N classification) were not associated with any module. Thus, WGCNA analysis indicated that overall survival was the main clinical trait associated with the pancreatic cancer profiles in TCGA, so we focused on overall survival in our subsequent investigations.

## Construction of a prognostic model for pancreatic cancer

Since GSE62452 and TCGA harbored an adequate number of cases and sufficient clinical follow-up data, we performed a univariate Cox regression analysis of

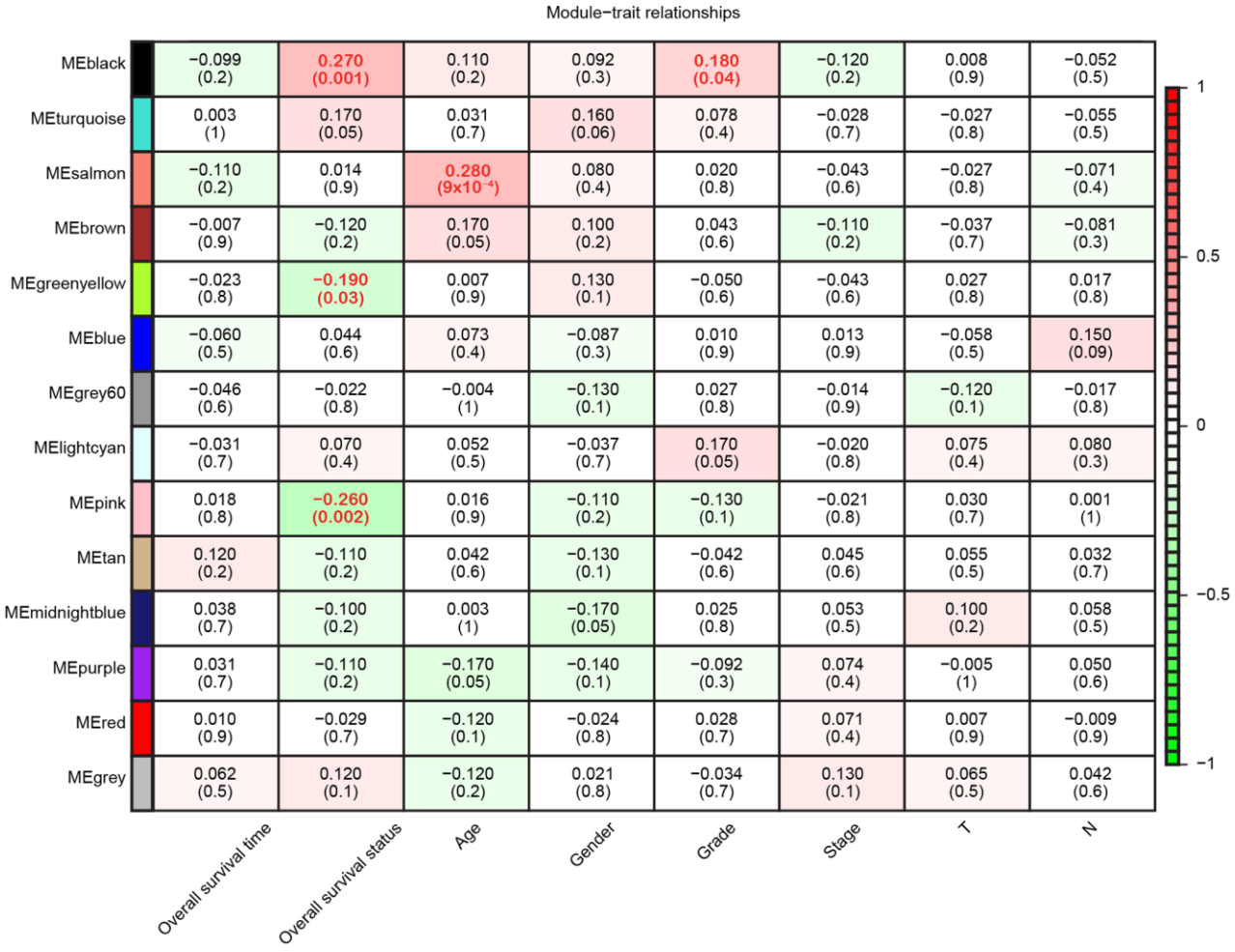



**Figure 3. Module-trait relationships.** Each row represents a color-coded module eigengene, each column represents a clinical trait, and each cell represents the Pearson correlation coefficient (top number) and $p$ value (in parentheses) of the corresponding module-trait. The color of each cell indicates the degree of correlation, as shown in the key. Abbreviations: T, Primary tumor; N, Regional lymph nodes.

these datasets to investigate the prognostic significance of the 542 consistently identified DEGs described above. A prognosis-associated gene was defined as impacting overall survival with a hazard ratio (HR) and 95% confidence interval (CI) greater than or less than 1. While 92 of the 542 DEGs had prognostic value in GSE62452 ($p < 0.05$), 76 of these 92 genes were still prognostically relevant in the cohort from TCGA ($p < 0.05$) (Figure 4).

A least absolute shrinkage and selection operator (LASSO) regression model [34] was then used to select key prognosis-associated genes. In the LASSO-penalized Cox regression, as log λ (a tuning parameter) changed, the corresponding coefficients of certain genes were reduced to zero, indicating that their effects on the model could be omitted because they were shrinking parameters (Figure 5A). Following cross-validation, nine genes achieved the minimum partial likelihood deviance (Figure 5B). Moreover, at this point, log λ was approximately -2.16, and the nine genes displayed non-zero effects, all contributing positive HRs to the model. The nine genes that were thus fit into the Cox model were *PBK*, which encodes MAPKK-like protein kinase; *DLGAP5*, which encodes a mitotic phosphoprotein; *RACGAP1*, also called Rac GTPase activating protein 1; *DSG3*, also called cadherin family member 6; *ARNTL2*, which functions as a TF; *NUSAP1*, also called nucleolar and spindle associated protein 1; *DKK1*, also called Dickkopf WNT signaling pathway inhibitor 1; *KRT7*, which encodes a cytokeratin; and *C15orf48*, a protein-coding gene that is also called chromosome 15 open reading frame 48.

Then, we randomly divided the 176 pancreatic cancer patients in TCGA (two cases in the enrolled TCGA pancreatic cancer cohort (N=178) were excluded due to insufficient follow-up data) into a training cohort (N = 88) for the construction of a prognostic model, and a validation cohort (N = 88) for internal self-validation (Table 4). Multivariate Cox regression analysis of the training cohort indicated that *ARNTL2*, *NUSAP1*, *DSG3* and *KRT7* were independent prognostic factors (Figure 5C). The prognostic risk score was calculated as: expression of *DSG3* × 0.17 + expression of *ARNTL2* × 0.58 + expression of *NUSAP1* × 0.92 + expression of *KRT7* × 0.22, where the numbers indicate the respective multivariate Cox regression coefficients.

Thus, through univariate Cox regression analysis, LASSO model shrinkage and multivariate Cox model construction, we obtained four DEGs (*DSG3*, *ARNTL2*, *NUSAP1* and *KRT7*) for prognostic risk evaluation.

**Stratification of TCGA training and self-validation cohorts using the four-gene signature**

We then divided the training cohort into high-risk and low-risk groups (Figure 6A). We used the median risk score (6.12) as the cut-off because the risk score usually exhibits a skewed distribution. The high-risk group displayed a higher frequency of poor survival outcomes than the low-risk group (Figure 6B). The three-year survival rates of the high-risk and low-risk groups were 7.78% and 51.3%, respectively (Figure 6C). To determine the predictive accuracy of this prognostic model, we performed a receiver operating characteristic (ROC) curve analysis, which demonstrated that the area under the curve (AUC) was 0.805 for one-year survival and 0.839 for three-year survival in the training cohort of TCGA (Figure 6D).

The model was further tested in the self-validation cohort by the same protocol (Figure 6E), which indicated that higher scores corresponded to worse overall survival (Figure 6F). The three-year survival rate was 28.6% in the high-risk group and 50.4% in the low-risk group (Figure 6G). Moreover, the AUCs for one-year and three-year survival in the validation cohort were 0.747 and 0.695, respectively (Figure 6H). Thus, the prognostic model successfully stratified pancreatic cancer patients from TCGA into high- and low-risk groups with moderate predictive power.

**The four-gene prognostic model is reliable in independent cohorts**

The predictive capacity of the four-gene model was further tested on two independent GEO microarray datasets (GSE28735 and GSE62452) that also included clinical data. The risk scores were calculated as described above (Figure 7A), and the expression of each gene was found to be greater in the high-risk group than in the low-risk group (Figure 7B). Consistent with the results in the entire TCGA cohort (Figure 7C), the risk score accurately stratified the patients of both datasets in terms of their survival outcomes. The three-year overall survival rates of the high- and low-risk groups in GSE28735 were 14.8% and 41.3%, respectively (Figure 7D), while those in GSE62452 were 5.15% and 45.3%, respectively, displaying an even greater prognostic difference (Figure 7E). Therefore, the four-gene signature is an accurate, reliable and independent predictive tool for determining the prognosis of pancreatic cancer patients.

## DISCUSSION

In this study, a pancreatic cancer prognostic model based on transcriptome dysregulation was developed, which was

compatible across the microarray and RNA-Seq platforms and among different patient cohorts. Our prognostic model containing four DEGs (*DSG3*, *ARNTL2*, *NUSAP1* and *KRT7*) was used to determine the risk scores of pancreatic cancer patients, and patients with high risk scores were found to exhibit poor overall survival.

Molecular predictors such as the multi-gene expression assays in lung cancer [35] and renal cell carcinoma [36] or the immunoscore in colon cancer [10] have shown promise in facilitating clinical decision-making in large international validation studies. However, commonly mutated genes are not the primary determinants of

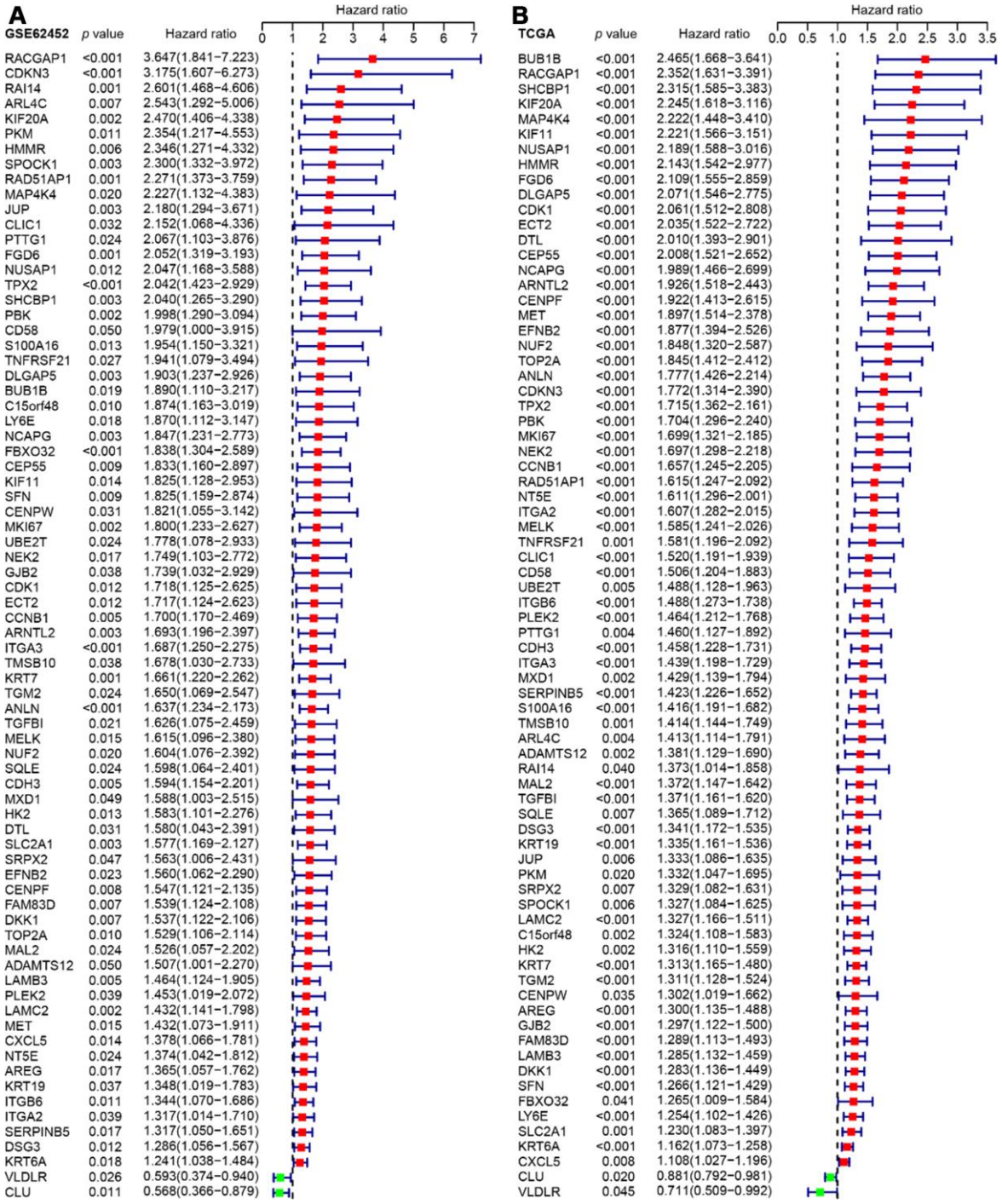



**Figure 4.** Forest plots visualizing the HRs of 76 prognostic DEGs identified by univariate Cox analysis of GSE62452 (**A**) and TCGA (**B**). The first three columns display the gene name, *p* value, and HR and 95% CI, respectively. In the forest plot, protective associations are shown in green and risk factors are shown in red.

PDAC prognosis [4]. On the other hand, transcriptome profiles have been successfully translated into prognostic markers in some prospective studies; for instance, a 21-gene signature was recently developed for breast cancer (the Oncotype DX breast cancer assay) [37]. To this end, we constructed a prognostic model for pancreatic cancer by analyzing the transcriptomes of a large number of pancreatic cancer patients across multiple datasets, and subsequently developed a four-gene risk score.

Since gene expression platforms are based on different analytical and data processing methods, it is often challenging to compare and integrate results from multiple datasets. In some previous studies, researchers have obtained integrated results by intersecting the results from different cohorts, which may lead to bias. Therefore, we used the robust rank aggregation method to screen for significantly altered genes across seven microarray datasets in an unbiased manner, and subsequently identified 542 DEGs that overlapped between the microarray datasets and the RNA-Seq datasets from TCGA and GTEx. A similar approach has been used to identify DEGs in prostate cancer [38] and bladder cancer patients [39], as well as in Pan-cancer [40] and multi-omics analyses [41]. The pancreatic-cancer-related DEGs were functionally enriched in the tumor immune microenvironment, with particular influence on the desmoplastic stroma, immune cell infiltration and perineural invasion, which contribute to cancer progression [5], metastasis [42] and chemotherapy resistance [43].

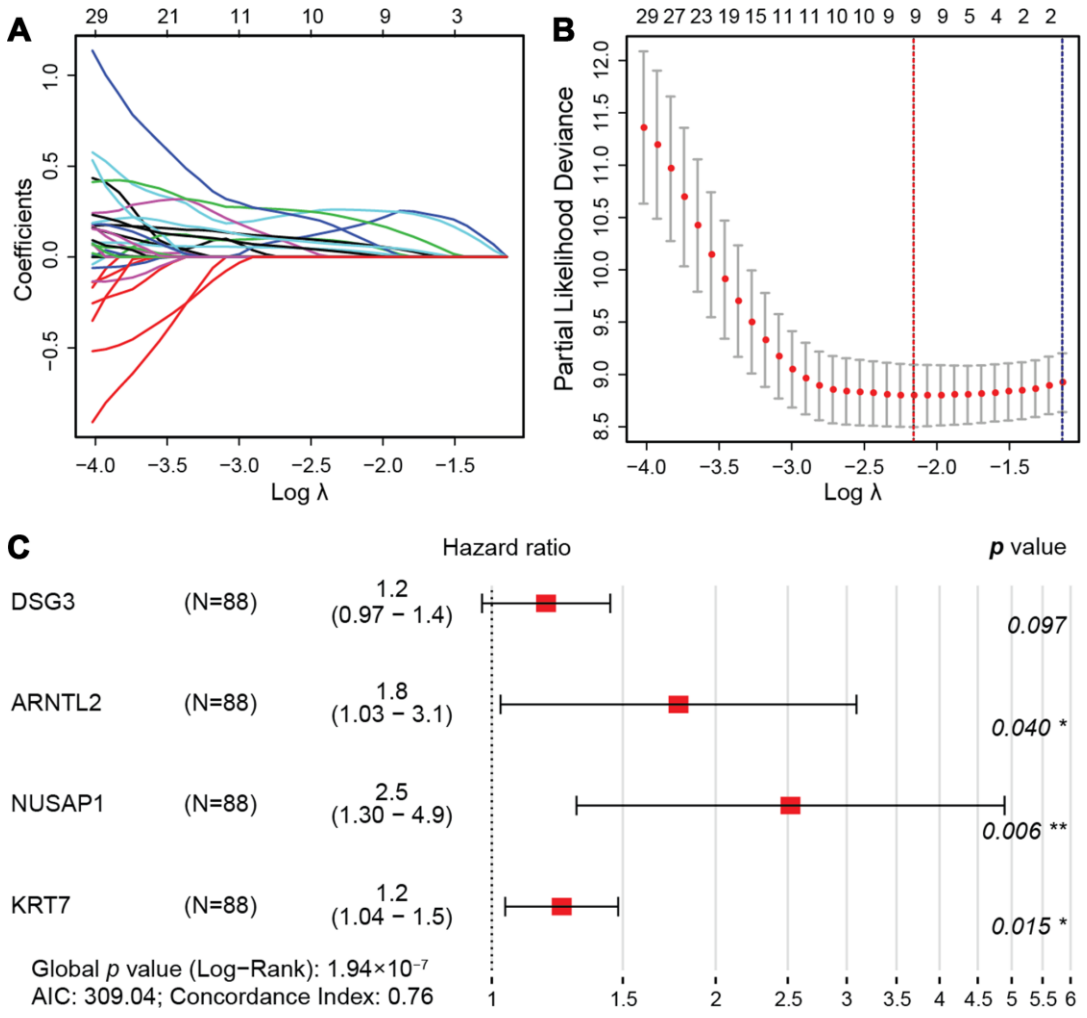



**Figure 5. LASSO regression model.** (**A**) LASSO coefficient profiles of the 76 prognostic DEGs. Each curve represents a coefficient, and the x-axis represents the regularization penalty parameter. As λ changes, a coefficient that becomes non-zero enters the LASSO regression model. (**B**) Cross-validation to select the optimal tuning parameter (λ). The red dotted vertical line crosses over the optimal log λ, which corresponds to the minimum value for multivariate Cox modeling. The two dotted lines represent one standard deviation from the minimum value. (**C**) HRs and 95% CIs of the four genes based on multivariate Cox regression analysis of the training cohort from TCGA.

**Table 4. Clinico-pathological characteristics of patients in the training and self-validation cohorts of TCGA and two independent GEO validation datasets.**

| Characteristics | TCGA training cohort (N = 88) | TCGA validation cohort (N = 88) | GSE62452 validation cohort (N = 64) | GSE28735 validation cohort (N = 40) |
|---|---|---|---|---|
| **Age at initial diagnosis (year)** | 64.7 ± 1.2 | 64.6 ± 1.1 | NA | NA |
| **Gender** | | | NA | NA |
| Male | 53 (60.2%) | 43 (48.9%) | | |
| Female | 35 (39.8%) | 45 (51.1%) | | |
| **Neoplasm histological grade** | | | | NA |
| G1 | 13 (14.8%) | 17 (19.3%) | 2 (3.1%) | |
| G2 | 50 (56.8%) | 44 (50.0%) | 31 (48.4%) | |
| G3 | 22 (25.0%) | 26 (29.5%) | 29 (45.3%) | |
| G4 | 1 (1.1%) | 1 (1.1%) | 1 (1.6%) | |
| Not report | 2 (2.3%) | 0 (0.0%) | 1 (1.6%) | |
| **Primary tumor (T)** | | | NA | NA |
| T1 | 3 (3.4%) | 4 (4.5%) | | |
| T2 | 13 (14.8%) | 11 (12.5%) | | |
| T3 | 71 (80.7%) | 69 (78.4%) | | |
| T4 | 1 (1.1%) | 2 (2.3%) | | |
| Not report | 0 (0.0%) | 2 (2.3%) | | |
| **Tumor stage at diagnosis** | | | | NA |
| Stage I | 7 (8.0%) | 14 (15.9%) | 4 (6.3%) | |
| Stage II | 76 (86.4%) | 69 (78.4%) | 45 (70.3%) | |
| Stage III | 1 (1.1%) | 2 (2.3%) | 9 (14.1%) | |
| Stage IV | 3 (3.4%) | 1 (1.1%) | 6 (9.4%) | |
| Not report | 1 (1.1%) | 2 (2.3%) | 0 (0.0%) | |
| **Overall survival time (years)** | 1.31 (0.79–1.83) | 1.25 (0.67–1.95) | 1.27 (0.74–2.27) | 1.28 (0.6–2.03) |
| **Overall survival status** | | | | |
| Alive | 41 (46.6%) | 43 (48.9%) | 16 (25.0%) | 13 (32.5%) |
| Dead | 47 (53.4%) | 45 (51.1%) | 48 (75.0%) | 27 (67.5%) |

NA: Not available.

High-throughput data tend to be interpreted from a clinical transformation perspective in the precision oncology era [44, 45]. It is necessary to integrate all the available information to identify the most relevant markers in a critical and comprehensive analysis. WGCNA is a powerful bioinformatics tool that detects clusters of functionally correlated genes and therefore can identify clinically relevant markers [13, 46]. WGCNA has been successfully used to identify molecular signatures in brain cells with distinct spatial distributions [47], to determine the key factors promoting hepatic ischemia-reperfusion injury [46] and to demarcate the molecular subtypes of pancreatic cancer [13]. The LASSO regression algorithm is another genotype-to-phenotype "bridge" that has been used to construct prognostic models from key radiomic [48] and immunohistochemical [49] features. Using these approaches, we found that overall survival was the main clinical trait associated with the transcriptome profiles of pancreatic cancer patients, and that *DSG3*, *ARNTL2*, *NUSAP1* and *KRT7* were independent prognostic factors.

Desmoglein 3 (DSG3) is a component of desmosomes, the button-like structures in the cytomembrane that facilitate intercellular and cell-to-matrix adhesion. DSG3 is overexpressed in head and neck cancer, where it functions as an oncogene [50, 51]. Previous studies have demonstrated that DSG3 is an accurate biomarker for staging sentinel lymph nodes in head and neck cancer [52, 53], and for distinguishing lung squamous cell carcinoma from other subtypes of lung cancer [54].

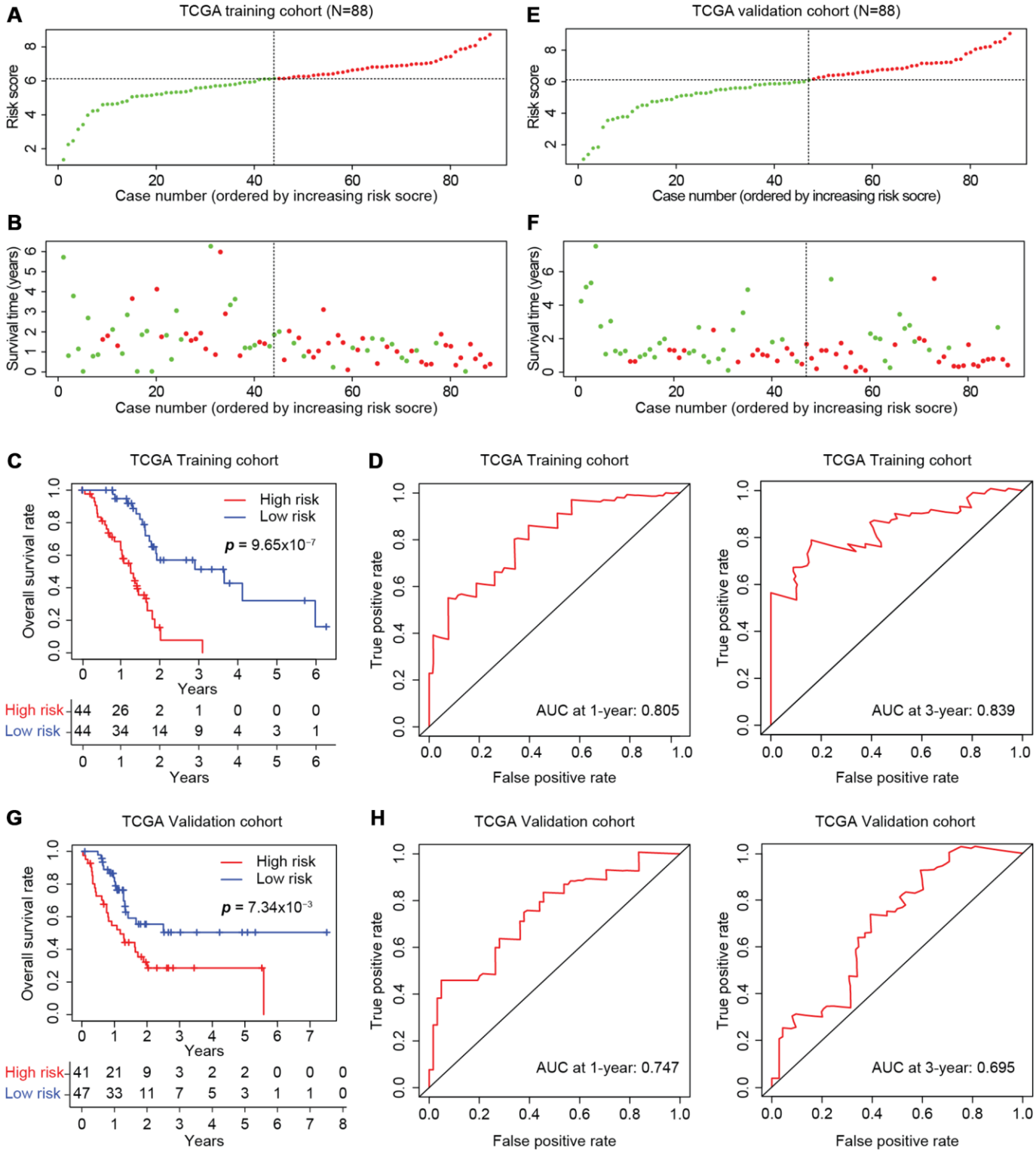



**Figure 6. Development of the prognostic scoring model in TCGA cohorts.** The distribution of risk scores is shown for the training (**A**) and validation (**E**) cohorts from TCGA. The dotted horizonal line indicates the cut-off level of the risk score used to stratify patients, and the dotted vertical line separates patients on the basis of low-risk (green) or high-risk (red). (**B**, **F**) The distribution of overall outcomes in the training (**B**) and validation (**F**) cohorts from TCGA. Surviving patients are shown in green, while deaths are shown in red. (**C**, **G**) Kaplan-Meier survival plots of patients predicted to be at risk for poor outcomes in the training (**C**) and validation (**G**) cohorts from TCGA. The number of patients remaining at a particular timepoint is shown at the bottom. (**D**, **H**) Time-dependent ROC curves for predicting one-year and three-year survival in the training (**D**) and validation (**H**) cohorts from TCGA.

Similarly, high expression of the pro-metastatic transcription factor ARNTL2 predicts poor survival in lung adenocarcinoma [55]. Forced expression of *Arntl2* in estrogen receptor-negative breast cancer cells was found to increase their metastatic potential and thus portend a poor prognosis [56]. NUSAP1 promotes mitosis, cell cycle progression and the DNA damage response as a substrate of Cyclin F [57–59]. Over-expression of NUSAP1 correlates with poor survival in melanoma [60], cervical carcinoma [61], prostate cancer [62] and glioblastoma multiforme [63]. KRT7, a membrane-cytoskeleton linker required for cell adhesion, is overexpressed in colon cancer [64] and esophageal squamous cell carcinoma [65], and is associated with poor survival and metastasis in colon cancer [64]. In an *in vivo* model, KRT7 was found to promote the transition of basal cells into the multi-layered epithelium and Barrett's esophagus [66]. Thus, all of the above genes have displayed pro-metastatic effects associated with poor outcomes in multiple cancers.

According to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement, internal validation (also called self-validation) is a necessary part of model development, and external validation to evaluate the performance of a model with other datasets is strongly recommended [67]. Therefore, we formulated our four-gene risk score based on the expression and Cox regression coefficient of each gene. Our model stratified pancreatic cancer patients of three independent cohorts into high- and low-risk groups with moderate accuracy. This three-cohort validation, together with the fact that our study was conducted with nine pancreatic cancer datasets in a platform-independent manner, indicates that our model is compatible across different platforms.

Several bioinformatic investigations in pancreatic cancer have previously been conducted from different perspectives. A nine-gene signature (*MET*, *KLK10*, *COL17A1*, *CEP55*, *ANKRD22*, *ITGB6*, *ARNTL2*,

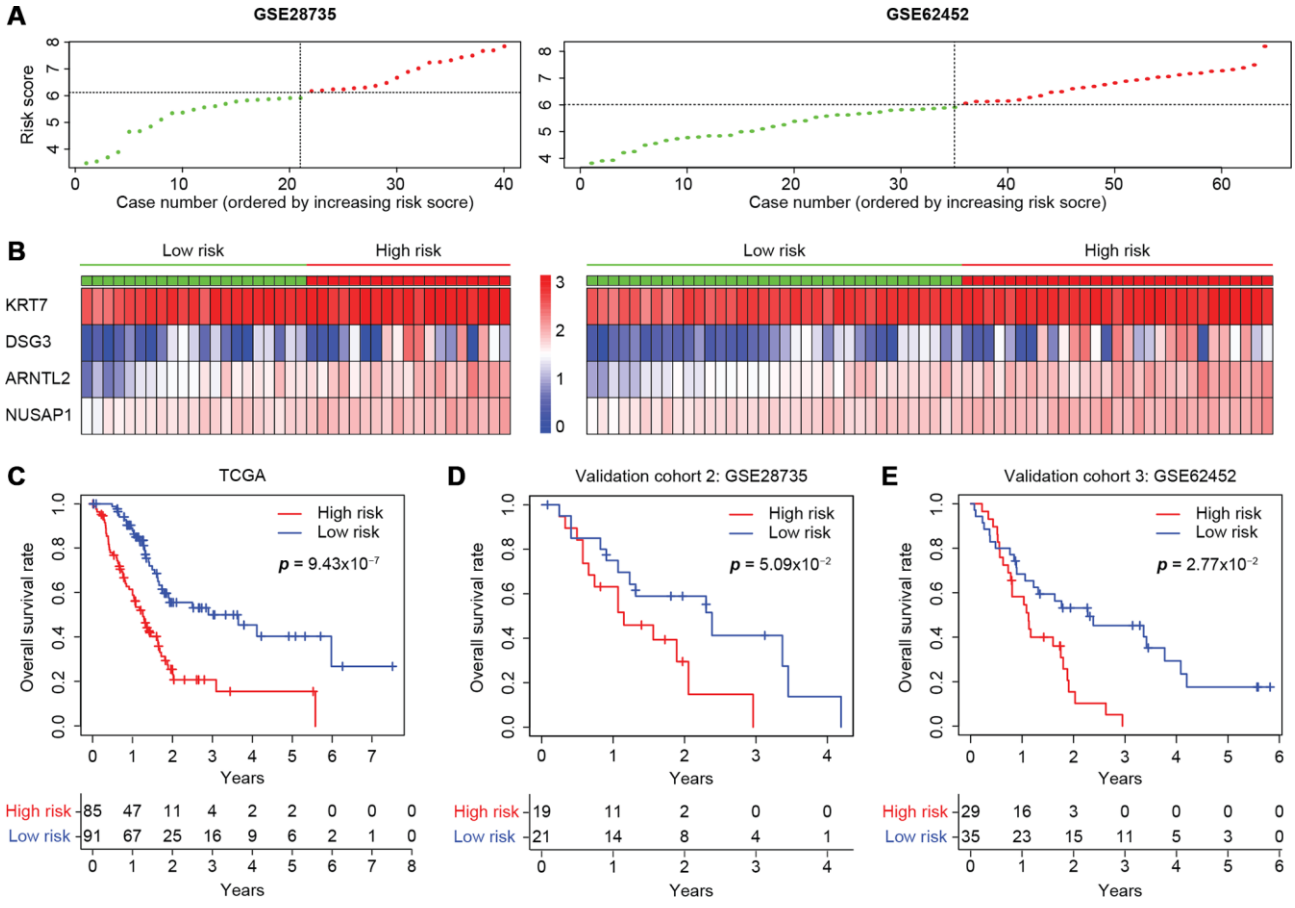



**Figure 7. Validation of the four-gene model in two independent microarray datasets.** (**A**) The risk score distribution in the GSE28735 and GSE62452 cohorts. (**B**) Heatmap displaying the levels of the four genes in the high- and low-risk groups. The color of each case corresponds to the $\log_2^{FC}$ of the gene level, as shown in the key. (**C**–**E**) Kaplan-Meier survival plots of high- or low-risk patients in the cohorts from TCGA (**C**), GSE28735 (**D**) and GSE62452 (**E**). The number of patients remaining at a particular timepoint is shown at the bottom.

*MCOLN3* and *SLC25A45*) [68] and a four-gene signature (*LYRM1*, *KNTC1*, *IGF2BP2* and *CDC6*) [69] were proposed to predict the overall survival of pancreatic cancer patients. Four of the genes identified by these two studies (*MET*, *CEP55*, *ITGB6* and *ARNTL2*) were also identified as prognosis-associated genes in our study. The AUCs for three-year overall survival prediction with the previously reported nine-gene model were 0.621, 0.814 and 0.670 in one training and two external validation cohorts, respectively, while the AUCs for three-year overall survival prediction with our four-gene model were 0.839, 0.695, 0.747 and 0.872 (the latter two are data not shown) in one training, one internal and two external validation cohorts, respectively. Thus, our four-gene prognostic model achieved moderate accuracy with less complexity than the nine-gene model.

On the other hand, certain proposed models, such as a five-gene [70] or 25-gene signature [71], were developed through the comparison of gene expression profiles between long-term and short-term survival groups of pancreatic cancer patients. In these studies, genes involved in extracellular matrix organization, cell adhesion and immune response suppression tended to be activated in the short-term survival group, consistent with our findings (Figures 4 and 7B). Other previous bioinformatics studies included only one or two PDAC datasets, and thus could not characterize common prognosis-associated genes across various datasets [72–74]. In contrast, our risk score was based on multiple datasets, a significantly larger number of patients and both microarray and sequencing platforms. This difference in patient groups and data processing methods may account for the different results.

This was the first study to combine comprehensive pancreatic cancer cohorts in a systematic analysis strategy. The major strengths of this study were the performance of quality control measures, the cross-validation of the results and the use of multiple cohorts for consistency. The use of publicly available data from millions of assays is a challenge [75], and one prerequisite is retaining the quality of the raw data. Since issues with technique or RNA degradation may occur, we checked the quality of the enrolled cases and removed any cases with potential problems. Another concern when integrating different platforms is balancing the different batch effects or detection methods, so we adopted robust rank aggregation to perform an unbiased analysis.

Our prognostic model, which was derived from multiple cohorts and validated by three cohorts, deserves further validation and translation into clinical practice, since the four genes in our model encode proteins and can be examined in clinical practice through routine cost-effective methods. However, since this model was developed from microarray and sequencing expression profiles, more common and practical methods (e.g., quantitative real-time PCR or immunohistochemistry) need to be used to translate the model into clinical practice. Additional studies are needed to elucidate whether the protein levels of these genes are consistent with their transcriptional levels in pancreatic cancer. For oncological research, these consistently altered genes are worthy of further investigation, and the prognosis-associated genes should be further characterized for both their involvement in cancer progression and their value as therapeutic targets.

In conclusion, we constructed a four-gene prognostic model for pancreatic cancer that can predict post-surgical prognosis with moderate accuracy and facilitate therapeutic decision-making and clinical monitoring.

## MATERIALS AND METHODS

### Study design and cohorts

In order to avoid biases caused by single or small numbers of cohorts, such as those based on a specific race, detection method or analysis technique, here we conducted a systematic retrospective analysis. We screened all the available high-throughput Affymetrix microarray datasets (up to August 2, 2019) in the GEO database to identify datasets that included both normal and cancerous pancreatic tissues and passed our quality control assessment for all the enrolled raw data. The analysis strategy is summarized in Supplementary Figure 1. Ultimately, we enrolled two RNA-Seq cohorts and seven microarray datasets. We first sought to identify common DEGs across all nine enrolled cohorts. WGCNA was then performed to connect clinical traits with pancreatic cancer expression profiles in TCGA. Based on the results of the WGCNA, we focused on determining the significance of the DEGs in predicting post-operation overall survival.

To establish the molecular prognostic model, we combined the prognosis-associated genes from TCGA and GSE62452 in a univariate Cox regression analysis. Next, the intersecting results were shrunk with the LASSO regression algorithm, such that highly interconnected genes were alternated to avoid overfitting. The filtered genes were entered into the multivariate Cox regression analysis, and a scoring model was built to predict overall survival.

The cohort from TCGA was randomly divided into two comparable sub-cohorts (each N = 88). The training cohort was used to optimize the parameters in the

LASSO and multivariate Cox regression analyses to build the risk score model, while the validation cohort was used to self-validate its performance. Two independent GEO microarray cohorts (GSE28735 and GSE62452) were also used to further validate the prognostic prediction model.

### Data acquisition

RNA-Seq data and clinical information from pancreatic cancer patients were downloaded from GTEx [26, 27] and TCGA [25, 76] via the UCSC Xena platform (https://xena.ucsc.edu/) [28]. The normal pancreatic and pancreatic-cancer-related RNA-Seq datasets are named 'GTEX_RSEM_gene_fpkm' and 'TCGA-PAAD/Xena_Metrices/TCGA-PAAD.htseq_fpkm', while the downloaded clinical data are entitled 'GTEX_phenotype' and 'TCGA-PAAD/Xena_Metrices/TCGA-PAAD.GDC_phenotype'. In addition, seven pancreatic cancer microarray datasets – GSE32676 [77], GSE16515 [78], GSE71989 [79], GSE41368 [80], GSE15471 [81], GSE28735 [82, 83] and GSE62452 [84] – were downloaded from the GEO database [85]. The human TF list was retrieved from the Cistrome Cancer website (http://cistrome.org/CistromeCancer/CancerTarget/) [31], and the list of human protein kinases was obtained from the Kinome of *Homo sapiens* [32].

### Microarray raw data quality control and identification of DEGs

Since all the selected GEO datasets were based on the Affymetrix platform, quality control was performed with Transcriptome Analysis Console software (Applied Biosystems, version 4.0.2) and the R 'simpleAffy' [23] and 'Affy' [22] packages. After averaging the expression values of the genes corresponding to the multi-microarray probes, we calculated the $\log_2^{FC}$ ratios between the normal and tumorous samples in each dataset, and determined their statistical significance with the R 'limma' package [86]. The 'RobustRankAggreg' R package was then used to identify the DEGs across the multi-microarray datasets based on a prioritized gene list, with a numerical core and *p* value determined through Bonferroni correction [24].

### TCGA and GTEx sequencing data integration and DEG identification

After the UCSC Toil RNA-Seq Recompute data were downloaded, the FPKM (fragments per kilobase of transcript per million mapped reads) values from GTEx were $\log_2^{(x+0.001)}$ transformed, and the values from TCGA were $\log_2^{(x+1)}$ transformed. Both forms were unified as $\log_2^{(x+1)}$, and the 'limma' package was used to screen for DEGs between normal and tumorous pancreatic tissues, with a $|\log_2^{FC}| > 1$ and a false discovery rate < 0.05 as the cut-offs. For both the sequencing and microarray platforms, a $\log_2^{FC} > 0$ indicated gene overexpression in the tumor tissues.

### Construction of the heatmap and GO plot

The DEGs in each sample were plotted with the R 'pheatmap' package [87]. Entrez gene annotations were referred to 'org.Hs.eg.db' (Carlson M, 2019. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2.), and the GO analysis was performed in the R 'clusterProfiler' package [88] with an adjusted *p* value < 0.01 and a *q* value < 0.01 as the cut-offs. The GO cluster was plotted with the R 'GOplot' package [89].

### WGCNA

The R 'WGCNA' package was used to detect gene modules and evaluate the correlation of each module with clinico-pathological factors. To be specific, (i) we extracted data from TCGA on pancreatic cancer patients with complete clinical data regarding age, gender, tumor histological grade, clinical stage, TNM classification (except for M, as there were numerous cases lacking metastasis information), overall survival rate and duration, along with their gene expression profiles; (ii) sample clustering was performed to detect any outliers; (iii) the power (soft thresholding) β value was set to 10 so that we could achieve a scale-free topology fit index (scale-free $R^2$) greater than 0.9 and maintain optimal mean connectivity; (iv) the adjacency matrix was transformed into a topological overlap matrix (TOM) to define gene co-expression similarity; (v) the 'hclust' algorithm was used to create a gene hierarchical clustering based on the TOM dissimilarity measure; (vi) the optimal module size was set to 30, and a dynamic tree cut was used to identify the modules; (vii) after the dissimilarity of the module eigengenes was calculated, the similarity cut-off was set to 0.75 in order to merge the modules; (viii) since we had already identified the featured gene expression profiles for each module and the clinical traits of the patients, the correlations of the module eigengenes with the clinico-pathological factors were determined.

### Construction of the prognostic model

Prognosis-associated genes were identified by the R 'survival' package [90] as those impacting overall survival with HRs and 95% CIs greater than or less than 1. Identified genes were then subjected to univariate Cox regression analysis with $p < 0.05$ as the significance threshold. The list was further narrowed down by the LASSO algorithm with the R 'glmnet'

package [91], and the optimal tuning parameter (λ) was chosen to achieve the minimal partial likelihood deviance in the cross-validation plot. The genes still harboring non-zero corresponding coefficients were entered into the multivariate Cox model. Finally, the expression of each gene was multiplied by its Cox regression coefficient, and these values were summed to calculate the risk score.

### Training and validation of the prognostic model

The R 'caret' package [92] was used to divide the cohort from TCGA randomly into training and self-validation sets (N = 88 each). Two microarray datasets (GSE28735 and GSE62452) were selected as independent validation cohorts. The risk score was calculated for each patient, and the patients were then divided into high-risk and low-risk groups based on the median risk score of the training cohort. The performance of the model was evaluated in terms of its ability to predict one-year and three-year overall survival in the high- and low-risk groups.

### Statistical analysis

Continuous variables with normal distributions are reported as the mean ± standard deviation, while those with skewed distributions are reported as the median ($25^{th}$ percentile - $75^{th}$ percentile). Categorical variables are reported as frequencies (proportions). All analyses were conducted with the R foundation for statistical computing (version 3.6.1). Pearson correlation analysis was used to determine the correlation between a module eigengene and a clinico-pathological factor, with $p < 0.05$ indicating statistical significance. Kaplan-Meier survival analysis and the log-rank test were performed with the R 'survival' package. The time-dependent ROC curve from censored survival data was plotted with the R 'survivalROC' package [93].

## Abbreviations

WGCNA: Weighted gene co-expression network analysis; LASSO: Least absolute shrinkage and selection operator; ROC curve: Receiver operating characteristic curve; AUC: Area under the ROC curve; DEGs: Differentially expressed genes; cTNM: Clinical tumor-node-metastasis staging; pTNM: Pathological tumor-node-metastasis staging; GEO: Gene Expression Omnibus; GO: Gene Ontology; BP: Biological process; CC: Cellular component; MF: Molecular function; TCGA: The Cancer Genome Atlas; GTEx: Genotype-tissue expression; TFs: Transcription factors; HR: Hazard ratio; CI: Confidence interval; FC: Fold change; DSG: Desmoglein; TRIPOD: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; TOM: Topological overlap matrix; PDAC: pancreatic ductal adenocarcinoma.

## AUTHOR CONTRIBUTIONS

Study design: Jie Chen and Jie Yan. Data collection, analysis and interpretation: Jie Yan, Liangcai Wu and Congwei Jia. Manuscript writing and figure preparation: Jie Yan and Shuangni Yu. Paper revision: Zhaohui Lu, Yueping Sun and Jie Chen.

All authors participated in the discussion and approved of the submission of this manuscript.

## ACKNOWLEDGMENTS

The results shown were based in part on the data of GSE32676 [77], GSE16515 [78], GSE71989 [79], GSE41368 [80], GSE15471 [81], GSE28735 [82, 83] and GSE62452 [84] available on the GEO platform [85], TCGA Research Network [25], GTEx project [26, 27], UCSC Xena [28], Cistrome Cancer [31] and human protein kinase database [32]. We thank all the researchers for their contributions and generous sharing.

## CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

## FUNDING

This study was supported by the CAMS Innovation Fund for Medical Sciences (Grant No. 2016-I2M-1-001), the National Natural Science Foundation of China (Grant No. 81672648), the PUMC Innovation Grant for Doctoral Students (Grant No. 1001-03) and the Youth Foundation of the National Natural Science Foundation of China (Grant No. 81902620).

## REFERENCES

1. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, Zackrisson S, Senkus E, and ESMO Guidelines Committee. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019; 30:1194–220. https://doi.org/10.1093/annonc/mdz173 PMID:31161190

2. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology, Hepatobiliary Cancers (version 1.2018). 2019. https://oncolife.com.ua/doc/nccn/Hepatobiliary_Cancers.pdf

3. Stark AP, Sacks GD, Rochefort MM, Donahue TR, Reber HA, Tomlinson JS, Dawson DW, Eibl G, Hines OJ. Long-

term survival in patients with pancreatic ductal adenocarcinoma. Surgery. 2016; 159:1520–27. https://doi.org/10.1016/j.surg.2015.12.024 PMID:26847803

4. Dal Molin M, Zhang M, de Wilde RF, Ottenhof NA, Rezaee N, Wolfgang CL, Blackford A, Vogelstein B, Kinzler KW, Papadopoulos N, Hruban RH, Maitra A, Wood LD. Very long-term survival following resection for pancreatic cancer is not explained by commonly mutated genes: results of Whole-Exome Sequencing analysis. Clin Cancer Res. 2015; 21:1944–50. https://doi.org/10.1158/1078-0432.CCR-14-2600 PMID:25623214

5. Hruban RH, Gaida MM, Thompson E, Hong SM, Noë M, Brosens LA, Jongepier M, Offerhaus GJ, Wood LD. Why is pancreatic cancer so deadly? The pathologist's view. J Pathol. 2019; 248:131–41. https://doi.org/10.1002/path.5260 PMID:30838636

6. Fassnacht M, Johanssen S, Quinkler M, Bucsky P, Willenberg HS, Beuschlein F, Terzolo M, Mueller HH, Hahner S, Allolio B, and German Adrenocortical Carcinoma Registry Group, and European Network for the Study of Adrenal Tumors. Limited prognostic value of the 2004 International Union Against Cancer staging classification for adrenocortical carcinoma: proposal for a Revised TNM Classification. Cancer. 2009; 115:243–50. https://doi.org/10.1002/cncr.24030 PMID:19025987

7. Kreppel M, Eich HT, Kübler A, Zöller JE, Scheer M. Prognostic value of the sixth edition of the UICC's TNM classification and stage grouping for oral cancer. J Surg Oncol. 2010; 102:443–9. https://doi.org/10.1002/jso.21547 PMID:20872947

8. Nitsche U, Maak M, Schuster T, Künzli B, Langer R, Slotta-Huspenina J, Janssen KP, Friess H, Rosenberg R. Prediction of prognosis is not improved by the seventh and latest edition of the TNM classification for colorectal cancer in a single-center collective. Ann Surg. 2011; 254:793–800. https://doi.org/10.1097/SLA.0b013e3182369101 PMID:22042471

9. Huang SH, Xu W, Waldron J, Siu L, Shen X, Tong L, Ringash J, Bayley A, Kim J, Hope A, Cho J, Giuliani M, Hansen A, et al. Refining American Joint Committee on Cancer/Union for International Cancer Control TNM stage and prognostic groups for human papillomavirus-related oropharyngeal carcinomas. J Clin Oncol. 2015; 33:836–45. https://doi.org/10.1200/JCO.2014.58.6412 PMID:25667292

10. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, Lugli A, Zlobec I, Rau TT, Berger MD, Nagtegaal ID, Vink-Börger E, Hartmann A, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. Lancet. 2018; 391:2128–39. https://doi.org/10.1016/S0140-6736(18)30789-X PMID:29754777

11. Bird-Lieberman EL, Dunn JM, Coleman HG, Lao-Sirieix P, Oukrif D, Moore CE, Varghese S, Johnston BT, Arthur K, McManus DT, Novelli MR, O'Donovan M, Cardwell CR, et al. Population-based study reveals new risk-stratification biomarker panel for Barrett's esophagus. Gastroenterology. 2012; 143:927–35.e3. https://doi.org/10.1053/j.gastro.2012.06.041 PMID:22771507

12. Dosaka-Akita H, Hommura F, Mishina T, Ogura S, Shimizu M, Katoh H, Kawakami Y. A risk-stratification model of non-small cell lung cancers using cyclin E, Ki-67, and ras p21: different roles of G1 cyclins in cell proliferation and prognosis. Cancer Res. 2001; 61:2500–04. PMID:11289121

13. Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, Miller DK, Christ AN, Bruxner TJ, Quinn MC, Nourse C, Murtaugh LC, Harliwong I, et al, and Australian Pancreatic Cancer Genome Initiative. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016; 531:47–52. https://doi.org/10.1038/nature16965 PMID:26909576

14. He B, Gao R, Lv D, Wen Y, Song L, Wang X, Lin S, Huang Q, Deng Z, Wang Z, Yan M, Zheng F, Lam EW, et al. The prognostic landscape of interactive biological processes presents treatment responses in cancer. EBioMedicine. 2019; 41:120–33. https://doi.org/10.1016/j.ebiom.2019.01.064 PMID:30799199

15. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, Omberg L, Wolf DM, Shriver CD, et al, and Cancer Genome Atlas Research Network. An integrated TCGA Pan-Cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018; 173:400–416.e11. https://doi.org/10.1016/j.cell.2018.02.052 PMID:29625055

16. Thorisson GA, Muilu J, Brookes AJ. Genotype-phenotype databases: challenges and solutions for the post-genomic era. Nat Rev Genet. 2009; 10:9–18. https://doi.org/10.1038/nrg2483 PMID:19065136

17. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007; 3:140–140. https://doi.org/10.1038/msb4100180 PMID:17940530

18. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019; 69:7–34. https://doi.org/10.3322/caac.21551 PMID:30620402

19. Rhim AD, Mirek ET, Aiello NM, Maitra A, Bailey JM, McAllister F, Reichert M, Beatty GL, Rustgi AK, Vonderheide RH, Leach SD, Stanger BZ. EMT and dissemination precede pancreatic tumor formation. Cell. 2012; 148:349–61. https://doi.org/10.1016/j.cell.2011.11.025 PMID:22265420

20. Zheng X, Carstens JL, Kim J, Scheible M, Kaye J, Sugimoto H, Wu CC, LeBleu VS, Kalluri R. Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. Nature. 2015; 527:525–30. https://doi.org/10.1038/nature16064 PMID:26560028

21. Yang J, Zhang Z, Zhang Y, Ni X, Zhang G, Cui X, Liu M, Xu C, Zhang Q, Zhu H, Yan J, Zhu VF, Luo Y, et al. ZIP4 promotes muscle wasting and cachexia in mice with orthotopic pancreatic tumors by stimulating RAB27B-regulated release of extracellular vesicles from cancer cells. Gastroenterology. 2019; 156:722–734.e6. https://doi.org/10.1053/j.gastro.2018.10.026 PMID:30342032

22. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004; 20:307–15. https://doi.org/10.1093/bioinformatics/btg405 PMID:14960456

23. Wilson CL, Miller CJ. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics. 2005; 21:3683–85. https://doi.org/10.1093/bioinformatics/bti605 PMID:16076888

24. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012; 28:573–80. https://doi.org/10.1093/bioinformatics/btr709 PMID:22247279

25. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, and Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45:1113–20. https://doi.org/10.1038/ng.2764 PMID:24071849

26. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, et al, and GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–85. https://doi.org/10.1038/ng.2653 PMID:23715323

27. Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, et al, and GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–60. https://doi.org/10.1126/science.1262110 PMID:25954001

28. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, Pfeil J, Narkizian J, Deran AD, Musselman-Brown A, Schmidt H, Amstutz P, Craft B, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017; 35:314–16. https://doi.org/10.1038/nbt.3772 PMID:28398314

29. Bushweller JH. Targeting transcription factors in cancer - from undruggable to reality. Nat Rev Cancer. 2019; 19:611–24. https://doi.org/10.1038/s41568-019-0196-7 PMID:31511663

30. Bhullar KS, Lagarón NO, McGowan EM, Parmar I, Jha A, Hubbard BP, Rupasinghe HP. Kinase-targeted cancer therapies: progress, challenges and future directions. Mol Cancer. 2018; 17:48–48. https://doi.org/10.1186/s12943-018-0804-2 PMID:29455673

31. Mei S, Meyer CA, Zheng R, Qin Q, Wu Q, Jiang P, Li B, Shi X, Wang B, Fan J, Shih C, Brown M, Zang C, Liu XS. Cistrome Cancer: A web resource for integrative gene regulation modeling in cancer. Cancer Res. 2017; 77:e19–22. https://doi.org/10.1158/0008-5472.CAN-17-0327 PMID:29092931

32. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002; 298:1912–34. https://doi.org/10.1126/science.1075762 PMID:12471243

33. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559. https://doi.org/10.1186/1471-2105-9-559 PMID:19114008

34. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997; 16:385–95. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3 PMID:9044528

35. Kratz JR, He J, Van Den Eeden SK, Zhu ZH, Gao W, Pham PT, Mulvihill MS, Ziaei F, Zhang H, Su B, Zhi X, Quesenberry CP, Habel LA, et al. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. Lancet. 2012; 379:823–32. https://doi.org/10.1016/S0140-6736(11)61941-7 PMID:22285053

36. Rini B, Goddard A, Knezevic D, Maddala T, Zhou M, Aydin H, Campbell S, Elson P, Koscielny S, Lopatin M, Svedman C, Martini JF, Williams JA, et al. A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies. Lancet Oncol. 2015; 16:676–85. https://doi.org/10.1016/S1470-2045(15)70167-1 PMID:25979595

37. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, Geyer CE Jr, Dees EC, Perez EA, Olson JA Jr, Zujewski J, Lively T, Badve SS, et al. Prospective validation of a 21-Gene Expression Assay in breast cancer. N Engl J Med. 2015; 373:2005–14. https://doi.org/10.1056/NEJMoa1510764 PMID:26412349

38. Song ZY, Chao F, Zhuo Z, Ma Z, Li W, Chen G. Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis. Aging (Albany NY). 2019; 11:4736–56. https://doi.org/10.18632/aging.102087 PMID:31306099

39. Zhou H, Tang K, Xiao H, Zeng J, Guan W, Guo X, Xu H, Ye Z. A panel of eight-miRNA signature as a potential biomarker for predicting survival in bladder cancer. J Exp Clin Cancer Res. 2015; 34:53. https://doi.org/10.1186/s13046-015-0167-0 PMID:25991007

40. Zhang Y, Tao Y, Ji H, Li W, Guo X, Ng DM, Haleem M, Xi Y, Dong C, Zhao J, Zhang L, Zhang X, Xie Y, et al. Genome-wide identification of the essential protein-coding genes and long non-coding RNAs for human pan-cancer. Bioinformatics. 2019; 35:4344–49. https://doi.org/10.1093/bioinformatics/btz230 PMID:30923830

41. Dimitrakopoulos C, Hindupur SK, Häfliger L, Behr J, Montazeri H, Hall MN, Beerenwinkel N. Network-based integration of multi-omics data for prioritizing cancer genes. Bioinformatics. 2018; 34:2441–48. https://doi.org/10.1093/bioinformatics/bty148 PMID:29547932

42. Ren B, Cui M, Yang G, Wang H, Feng M, You L, Zhao Y. Tumor microenvironment participates in metastasis of pancreatic cancer. Mol Cancer. 2018; 17:108. https://doi.org/10.1186/s12943-018-0858-1 PMID:30060755

43. Feig C, Gopinathan A, Neesse A, Chan DS, Cook N, Tuveson DA. The pancreas cancer microenvironment. Clin Cancer Res. 2012; 18:4266–76. https://doi.org/10.1158/1078-0432.CCR-11-3114 PMID:22896693

44. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL, McKenna A, Sivachenko A, Rosenberg M, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med. 2014; 20:682–88. https://doi.org/10.1038/nm.3559 PMID:24836576

45. Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, Kalyana-Sundaram S, Sam L, Balbin OA, Quist MJ, Barrette T, Everett J, Siddiqui J, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med. 2011; 3:111ra121. https://doi.org/10.1126/scitranslmed.3003161 PMID:22133722

46. Zhang XJ, Cheng X, Yan ZZ, Fang J, Wang X, Wang W, Liu ZY, Shen LJ, Zhang P, Wang PX, Liao R, Ji YX, Wang JY, et al. An ALOX12-12-HETE-GPR31 signaling axis is a key mediator of hepatic ischemia-reperfusion injury. Nat Med. 2018; 24:73–83. https://doi.org/10.1038/nm.4451 PMID:29227475

47. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, Abajian C, Beckmann CF, Bernard A, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012; 489:391–99. https://doi.org/10.1038/nature11405 PMID:22996553

48. Huang YQ, Liang CH, He L, Tian J, Liang CS, Chen X, Ma ZL, Liu ZY. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. J Clin Oncol. 2016; 34:2157–64. https://doi.org/10.1200/JCO.2015.65.9128 PMID:27138577

49. Meng J, Zhang J, Xiu Y, Jin Y, Xiang J, Nie Y, Fu S, Zhao K. Prognostic value of an immunohistochemical signature in patients with esophageal squamous cell carcinoma undergoing radical esophagectomy. Mol Oncol. 2018; 12:196–207. https://doi.org/10.1002/1878-0261.12158 PMID:29160958

50. Chen YJ, Chang JT, Lee L, Wang HM, Liao CT, Chiu CC, Chen PJ, Cheng AJ. DSG3 is overexpressed in head neck cancer and is a potential molecular target for inhibition of oncogenesis. Oncogene. 2007; 26:467–76. https://doi.org/10.1038/sj.onc.1209802 PMID:16878157

51. Brown L, Waseem A, Cruz IN, Szary J, Gunic E, Mannan T, Unadkat M, Yang M, Valderrama F, O'Toole EA, Wan H. Desmoglein 3 promotes cancer cell migration and invasion by regulating activator protein 1 and protein kinase C-dependent-Ezrin activation. Oncogene. 2014; 33:2363–74.

https://doi.org/10.1038/onc.2013.186
PMID:23752190

52. Solassol J, Burcia V, Costes V, Lacombe J, Mange A, Barbotte E, de Verbizier D, Cartier C, Makeieff M, Crampette L, Boulle N, Maudelonde T, Guerrier B, Garrel R. Pemphigus vulgaris antigen mRNA quantification for the staging of sentinel lymph nodes in head and neck cancer. Br J Cancer. 2010; 102:181–87. https://doi.org/10.1038/sj.bjc.6605470 PMID:19997107

53. Ferris RL, Xi L, Seethala RR, Chan J, Desai S, Hoch B, Gooding W, Godfrey TE. Intraoperative qRT-PCR for detection of lymph node metastasis in head and neck cancer. Clin Cancer Res. 2011; 17:1858–66. https://doi.org/10.1158/1078-0432.CCR-10-3110 PMID:21355082

54. Savci-Heijink CD, Kosari F, Aubry MC, Caron BL, Sun Z, Yang P, Vasmatzis G. The role of desmoglein-3 in the diagnosis of squamous cell carcinoma of the lung. Am J Pathol. 2009; 174:1629–37. https://doi.org/10.2353/ajpath.2009.080778 PMID:19342368

55. 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. https://doi.org/10.1016/j.ccell.2016.03.003 PMID:27150038

56. 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. https://doi.org/10.1371/journal.pgen.1006267 PMID:27656887

57. Emanuele MJ, Elia AE, Xu Q, Thoma CR, Izhar L, Leng Y, Guo A, Chen YN, Rush J, Hsu PW, Yen HC, Elledge SJ. Global identification of modular cullin-RING ligase substrates. Cell. 2011; 147:459–74. https://doi.org/10.1016/j.cell.2011.09.019 PMID:21963094

58. Kotian S, Banerjee T, Lockhart A, Huang K, Catalyurek UV, Parvin JD. NUSAP1 influences the DNA damage response by controlling BRCA1 protein levels. Cancer Biol Ther. 2014; 15:533–43. https://doi.org/10.4161/cbt.28019 PMID:24521615

59. Iyer J, Moghe S, Furukawa M, Tsai MY. What's Nu(SAP) in mitosis and cancer? Cell Signal. 2011; 23:991–98. https://doi.org/10.1016/j.cellsig.2010.11.006 PMID:21111812

60. Bogunovic D, O'Neill DW, Belitskaya-Levy I, Vacic V, Yu YL, Adams S, Darvishian F, Berman R, Shapiro R, Pavlick AC, Lonardi S, Zavadil J, Osman I, Bhardwaj N. Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. Proc Natl Acad Sci USA. 2009; 106:20429–34. https://doi.org/10.1073/pnas.0905139106 PMID:19915147

61. Li H, Zhang W, Yan M, Qiu J, Chen J, Sun X, Chen X, Song L, Zhang Y. Nucleolar and spindle associated protein 1 promotes metastasis of cervical carcinoma cells by activating Wnt/β-catenin signaling. J Exp Clin Cancer Res. 2019; 38:33. https://doi.org/10.1186/s13046-019-1037-y PMID:30678687

62. Gulzar ZG, McKenney JK, Brooks JD. Increased expression of NuSAP in recurrent prostate cancer is mediated by E2F1. Oncogene. 2013; 32:70–77. https://doi.org/10.1038/onc.2012.27 PMID:22349817

63. Qian Z, Li Y, Ma J, Xue Y, Xi Y, Hong L, Dai X, Zhang Y, Ji X, Chen Y, Sheng M, Sheng Y, Yang L, et al. Prognostic value of NUSAP1 in progression and expansion of glioblastoma multiforme. J Neurooncol. 2018; 140:199–208. https://doi.org/10.1007/s11060-018-2942-1 PMID:29995176

64. Czapiewski P, Bobowicz M, Pęksa R, Skrzypski M, Gorczyński A, Szczepańska-Michalska K, Korwat A, Jankowski M, Zegarski W, Szulgo-Paczkowska A, Polec T, Piątek M, Skokowski J, et al. Keratin 7 expression in lymph node metastases but not in the primary tumour correlates with distant metastases and poor prognosis in colon carcinoma. Pol J Pathol. 2016; 67:228–34. https://doi.org/10.5114/pjp.2016.63774 PMID:28155971

65. Sano M, Aoyagi K, Takahashi H, Kawamura T, Mabuchi T, Igaki H, Tachimori Y, Kato H, Ochiai A, Honda H, Nimura Y, Nagino M, Yoshida T, Sasaki H. Forkhead box A1 transcriptional pathway in KRT7-expressing esophageal squamous cell carcinomas with extensive lymph node metastasis. Int J Oncol. 2010; 36:321–30. https://doi.org/10.3892/ijo_00000503 PMID:20043065

66. Jiang M, Li H, Zhang Y, Yang Y, Lu R, Liu K, Lin S, Lan X, Wang H, Wu H, Zhu J, Zhou Z, Xu J, et al. Transitional basal cells at the squamous-columnar junction generate Barrett's oesophagus. Nature. 2017; 550:529–33. https://doi.org/10.1038/nature24269 PMID:29019984

67. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 2015; 13:1. https://doi.org/10.1186/s12916-014-0241-z PMID:25563062

68. 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; 9:996–996. https://doi.org/10.3389/fonc.2019.00996 PMID:31612115

69. Yan X, Wan H, Hao X, Lan T, Li W, Xu L, Yuan K, Wu H. Importance of gene expression signatures in pancreatic cancer prognosis and the establishment of a prediction model. Cancer Manag Res. 2018; 11:273–83. https://doi.org/10.2147/CMAR.S185205 PMID:30643453

70. Raman P, Maddipati R, Lim KH, Tozeren A. Pancreatic cancer survival analysis defines a signature that predicts outcome. PLoS One. 2018; 13:e0201751. https://doi.org/10.1371/journal.pone.0201751 PMID:30092011

71. Birnbaum DJ, Finetti P, Lopresti A, Gilabert M, Poizat F, Raoul JL, Delpero JR, Moutardier V, Birnbaum D, Mamessier E, Bertucci F. A 25-gene classifier predicts overall survival in resectable pancreatic cancer. BMC Med. 2017; 15:170. https://doi.org/10.1186/s12916-017-0936-z PMID:28927421

72. Zhou Z, Cheng Y, Jiang Y, Liu S, Zhang M, Liu J, Zhao Q. Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. Int J Biol Sci. 2018; 14:124–36. https://doi.org/10.7150/ijbs.22619 PMID:29483831

73. Li C, Zeng X, Yu H, Gu Y, Zhang W. Identification of hub genes with diagnostic values in pancreatic cancer by bioinformatics analyses and supervised learning methods. World J Surg Oncol. 2018; 16:223. https://doi.org/10.1186/s12957-018-1519-y PMID:30428899

74. Wu J, Li Z, Zeng K, Wu K, Xu D, Zhou J, Xu L. Key genes associated with pancreatic cancer and their association with outcomes: A bioinformatics analysis. Mol Med Rep. 2019; 20:1343–52. https://doi.org/10.3892/mmr.2019.10321 PMID:31173193

75. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet. 2013; 14:89–99. https://doi.org/10.1038/nrg3394 PMID:23269463

76. Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu; Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell. 2017; 32:185–203.e13. https://doi.org/10.1016/j.ccell.2017.07.007 PMID:28810144

77. Donahue TR, Tran LM, Hill R, Li Y, Kovochich A, Calvopina JH, Patel SG, Wu N, Hindoyan A, Farrell JJ, Li X, Dawson DW, Wu H. Integrative survival-based molecular profiling of human pancreatic cancer. Clin Cancer Res. 2012; 18:1352–63. https://doi.org/10.1158/1078-0432.CCR-11-1539 PMID:22261810

78. Pei H, Li L, Fridley BL, Jenkins GD, Kalari KR, Lingle W, Petersen G, Lou Z, Wang L. FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. Cancer Cell. 2009; 16:259–66. https://doi.org/10.1016/j.ccr.2009.07.016 PMID:19732725

79. Jiang J, Azevedo-Pouly AC, Redis RS, Lee EJ, Gusev Y, Allard D, Sutaria DS, Badawi M, Elgamal OA, Lerner MR, Brackett DJ, Calin GA, Schmittgen TD. Globally increased ultraconserved noncoding RNA expression in pancreatic adenocarcinoma. Oncotarget. 2016; 7:53165–77. https://doi.org/10.18632/oncotarget.10242 PMID:27363020

80. Frampton AE, Castellano L, Colombo T, Giovannetti E, Krell J, Jacob J, Pellegrino L, Roca-Alonso L, Funel N, Gall TM, De Giorgio A, Pinho FG, Fulci V, et al. MicroRNAs cooperatively inhibit a network of tumor suppressor genes to promote pancreatic tumor growth and progression. Gastroenterology. 2014; 146:268–77.e18. https://doi.org/10.1053/j.gastro.2013.10.010 PMID:24120476

81. Badea L, Herlea V, Dima SO, Dumitrascu T, Popescu I. Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. Hepatogastroenterology. 2008; 55:2016–27. PMID:19260470

82. Zhang G, Schetter A, He P, Funamizu N, Gaedcke J, Ghadimi BM, Ried T, Hassan R, Yfantis HG, Lee DH, Lacy C, Maitra A, Hanna N, et al. DPEP1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma. PLoS One. 2012; 7:e31507. https://doi.org/10.1371/journal.pone.0031507 PMID:22363658

83. Zhang G, He P, Tan H, Budhu A, Gaedcke J, Ghadimi BM, Ried T, Yfantis HG, Lee DH, Maitra A, Hanna N, Alexander HR, Hussain SP. Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. Clin Cancer Res. 2013; 19:4983–93. https://doi.org/10.1158/1078-0432.CCR-13-0209 PMID:23918603

84. Yang S, He P, Wang J, Schetter A, Tang W, Funamizu N, Yanaga K, Uwagawa T, Satoskar AR, Gaedcke J,

Bernhardt M, Ghadimi BM, Gaida MM, et al. A novel MIF signaling pathway drives the malignant character of pancreatic cancer by targeting NR3C2. Cancer Res. 2016; 76:3838–50. https://doi.org/10.1158/0008-5472.CAN-15-2841 PMID:27197190

85. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30:207–10. https://doi.org/10.1093/nar/30.1.207 PMID:11752295

86. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47–47. https://doi.org/10.1093/nar/gkv007 PMID:25605792

87. Kolde R. Pheatmap: pretty heatmaps. R package version. 2012; 61: 915.

88. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16:284–87. https://doi.org/10.1089/omi.2011.0118 PMID:22455463

89. Walter W, Sánchez-Cabo F, Ricote M. GOplot: an R package for visually combining expression data with functional analysis. Bioinformatics. 2015; 31:2912–14. https://doi.org/10.1093/bioinformatics/btv300 PMID:25964631

90. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer; 2000. https://doi.org/10.1007/978-1-4757-3294-8

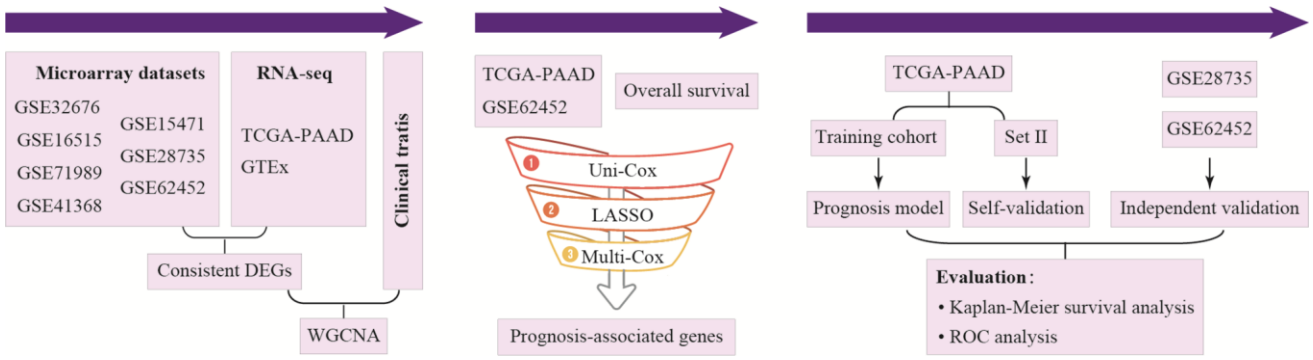
91. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010; 33:1–22. https://doi.org/10.18637/jss.v033.i01 PMID:20808728

92. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008; 28:26. https://doi.org/10.18637/jss.v028.i05

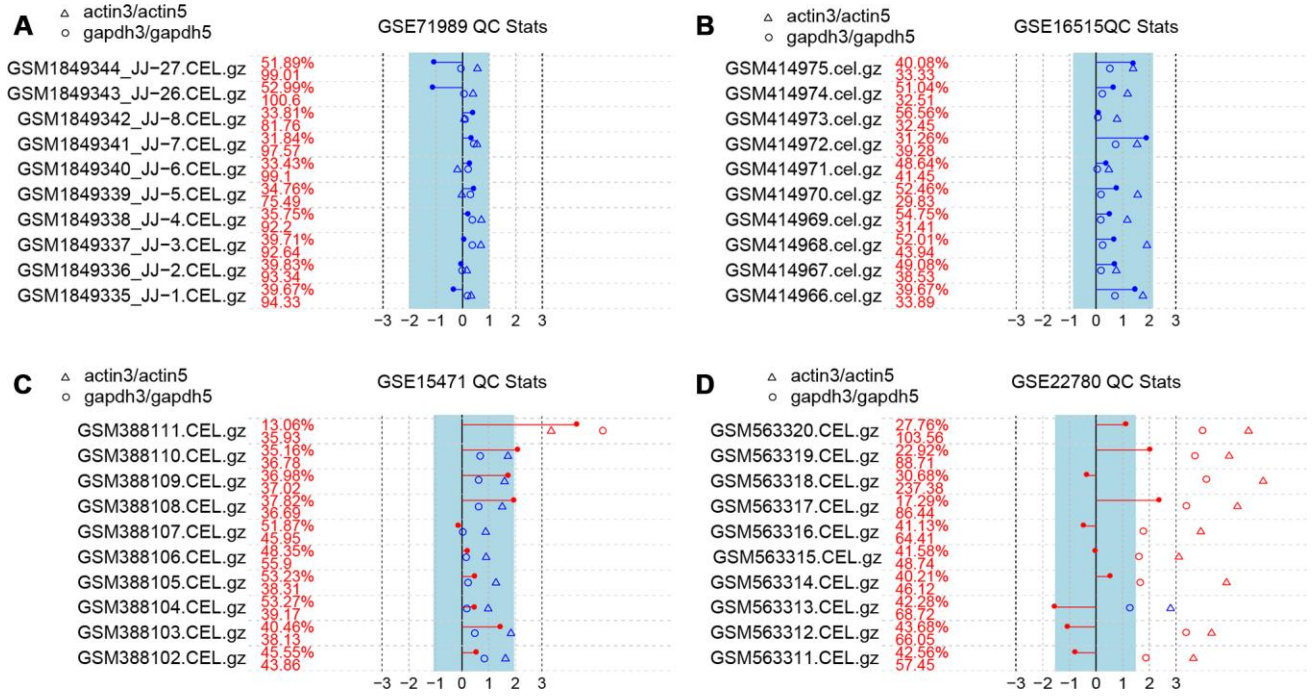
93. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000; 56:337–44. https://doi.org/10.1111/j.0006-341X.2000.00337.x PMID:10877287

# SUPPLEMENTARY MATERIALS

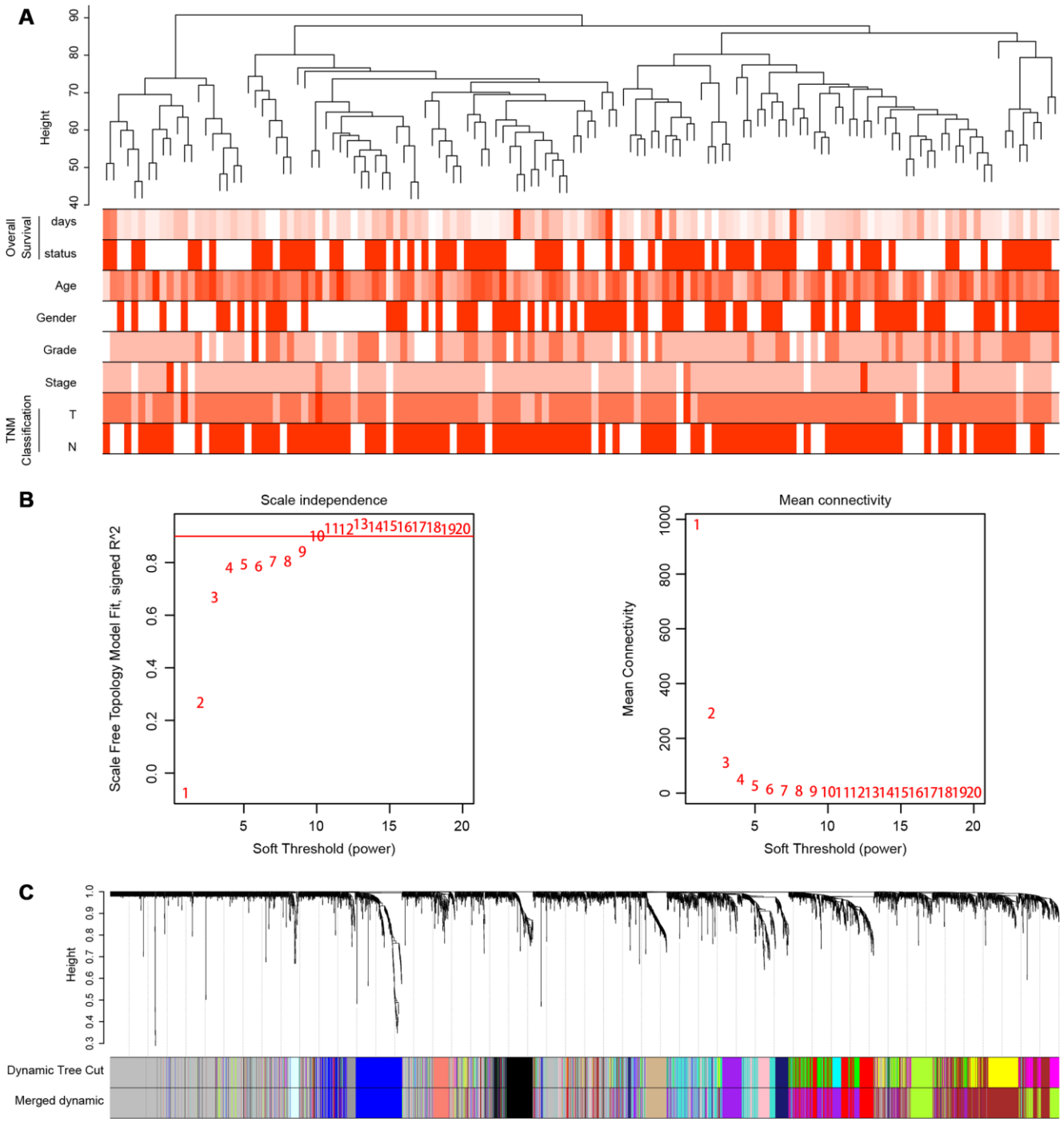
## Supplementary Figures




**Supplementary Figure 1. Flowchart presenting the study design of this retrospective analysis.** "Uni-Cox" refers to univariate Cox regression, and "Multi-Cox" refers to multivariate Cox model construction.




**Supplementary Figure 2. Quality control results from four pancreatic cancer microarrays.** For each panel, the raw data names of each case are presented in the first column, the % present calls (top) and average background (bottom) are displayed in the second column, and the range within which the scale factor should be located (a horizontal line with a dot at the top) is indicated by the blue stripe. The 3'-to-5' ratios of *β-actin* and *GAPDH* are plotted as triangles and circles, respectively. Outliers of the ratio are shown in red, and otherwise are shown in blue. The representative results of 10 samples in each array demonstrate the all-passed (**A** and **B**), one outlier case (**C**) and all-failed (**D**) samples in our quality control recheck.




**Supplementary Figure 3. Modules of TCGA pancreatic cancer expression profiles (N = 135).** (**A**) Sample dendrogram and clinical trait heatmap. With the exception of four outliers, all the samples could be hierarchically clustered. Binary variables (overall survival status, gender) are presented as red or white blocks, and continuous variables (overall survival days, age, grade, stage, T, N) are presented as color-coded blocks, with the color saturation corresponding to the value of the variable. (**B**) Determination of the soft-thresholding power for the optimal scale-free topology fit index (scale-free $R^2$) (left) and mean connectivity (right). The red horizontal line represents $R^2 = 0.9$. (**C**) Module identification. The dendrogram represents the gene clustering based on the TOM dissimilarity measure. Genes with relative interrelatedness are located on the same or neighboring branches. A dynamic tree cut at module size 30 resulted in 17 color-coded modules. By merging the modules after calculating the dissimilarity of the module eigengenes at a cut-off of 0.75, we identified 14 modules.

## Supplementary Tables

Please browse Full Text version to see the data of Supplementary Tables 1 and 2.

**Supplementary Table 1. The 542 differentially expressed genes in both the sequencing and multi-microarray datasets.**

**Supplementary Table 2. The correlation of a gene with its Gene Significance (GS) for a specific clinical trait or Module Membership (MM) in a specific module.**