Development and validation of a four-lipid metabolism gene signature for diagnosis of pancreatic cancer

Yanrong Ye[1,2], Zhe Chen[1], Yun Shen[1], Yan Qin[1] and Hao Wang[3]

1 Department of Pharmacy, Zhongshan Hospital, Fudan University, Shanghai, China
2 Department of Pharmacy, Xiamen Branch, Zhongshan Hospital, Fudan University, Xiamen, China
3 Teaching Center of Experimental Medicine, Shanghai Medical College, Fudan University, Shanghai, China

Keywords
four-gene signature; pancreatic cancer; prognosis; TCGA; WGCNA

Correspondence
Y. Qin, Department of Pharmacy, Zhongshan Hospital, Fudan University, Shanghai 200032, China
E-mail: qin.yan@zs-hospital.sh.cn
H. Wang, Teaching Center of Experimental Medicine, Shanghai Medical College, Fudan University, 138 Yixueyuan Road, Shanghai 200032, China
E-mail: wh91351@126.com

(Received 10 October 2020, revised 17 December 2020, accepted 30 December 2020)

doi:10.1002/2211-5463.13074

Abnormal lipid metabolism is closely related to the malignant biological behavior of tumor cells. Such abnormal lipid metabolism provides energy for rapid proliferation, and certain genes related to lipid metabolism encode important components of tumor signaling pathways. In this study, we analyzed pancreatic cancer datasets from The Cancer Genome Atlas and searched for prognostic genes related to lipid metabolism in the Molecular Signature Database. A risk score model was built and verified using the GSE57495 dataset and International Cancer Genome Consortium dataset. Four molecular subtypes and 4249 differentially expressed genes (DEGs) were identified. The DEGs obtained by Weighted Gene Coexpression Network Construction analysis were intersected with 4249 DEGs to obtain a total of 1340 DEGs. The final prognosis model included *CA8*, *CEP55*, *GNB3* and *SGSM2*, and these had a significant effect on overall survival. The area under the curve at 1, 3 and 5 years was 0.72, 0.79 and 0.87, respectively. These same results were obtained using the validation cohort. Survival analysis data showed that the model could stratify the prognosis of patients with different clinical characteristics, and the model has clinical independence. Functional analysis indicated that the model is associated with multiple cancer-related pathways. Compared with published models, our model has a higher C-index and greater risk value. In summary, this four-gene signature is an independent risk factor for pancreatic cancer survival and may be an effective prognostic indicator.

As a malignancy of the digestive system, pancreatic cancer is one of the most aggressive malignancies in the world. In recent years, pancreatic cancer morbidity and mortality rates have steadily increased, with an overall 5-year survival rate of about 8% for patients [1]. Despite the rapid development of diagnosis and treatment of pancreatic cancer over the past 20 years, the mortality rate of patients remains high [2–4]. The combination of tumor markers and imaging helps to diagnose the disease in a timely and accurate manner [5]; however, due to the insidious nature of the disease and the lack of early clinical signs, it is not easy to diagnose. It has been reported that about 50% of patients have confirmed that the cancer has

Abbreviations
AUC, area under the receiver operating characteristic curve; CI, confidence interval; DEG, differentially expressed gene; FDR, false discovery rate; GEO, Gene Expression Omnibus; GO, Gene Ontology; GSVA, Gene Set Variation Analysis; HR, hazard ratio; ICGC, International Cancer Genome Consortium; KEGG, Kyoto Encyclopedia of Genes and Genomes; KM, Kaplan–Meier; Lasso, least absolute shrinkage and selection operator; NMF, non-negative matrix clustering algorithm; OS, overall survival; RNA-seq, RNA sequencing; ROC, receiver operating characteristic; ssGSEA, single-sample gene set enrichment analysis; TCGA, The Cancer Genome Atlas; TIMER, Tumor Immune Estimation Resource; WGCNA, Weighted Gene Coexpression Network Construction.

metastasized [6,7]. Therefore, revealing the molecular mechanisms of pancreatic cancer progression and developing corresponding targeted therapies are critical to improving pancreatic cancer outcomes.

Lipids play an important function in maintaining normal cell function and homeostasis; they are not only an important part of the cell membrane but also provide precursors for important molecules needed in the growth and differentiation pathways [8,9]. Intracellular lipids come from two sources: one is food intake, and the other is lipid synthesis from scratch by hepatocytes and cells in need. Normal cells in the body acquire lipids mainly from diet and rarely from lipid synthesis from scratch. For most normal cells, lipids that meet cellular needs rarely are synthesized from scratch because of slow cell growth [10,11]. However, Medes *et al.* [12] found in the 1950s that tumor cells synthesize fatty acids primarily by synthesizing them from scratch. In tumor cells, on the one hand, lipids and cholesterol are often activated to meet the needs of tumor cells that are rapidly proliferating, and on the other hand, lipids alter the properties of biofilms and protect cells from oxidative damage from internal and external sources [13]. Lipogenesis is an important feature of rapid malignancy growth [14]. Normal cell lipid synthesis from scratch is rare, and about 90% of fatty acids are synthesized from scratch in tumor cells [12]. Activated lipid scratch synthesis was found to be associated with poorer prognosis and shorter disease-free survival in tumor patients [15,16]. At the molecular level, increased lipid synthesis from scratch in tumors is often accompanied by increased lipid synthase and enhanced activity [17,18]. Thus, aberrant activation of lipid synthesis from scratch is a common feature of tumor cells. In addition, lipid metabolic reprogramming that promotes increased lipogenesis is associated with the abnormal development and progression of pancreatic adenocarcinoma [19]. Mammalian target of rapamycin complex 2 (mTORC2) stimulates the synthesis of sphingomyelin (glucoceramide) and glycerophospholipids (cardiolipin) to promote tumor progression [20]. Studies have shown that mTORC1 stimulates the synthesis of fatty acids and sterols by regulating the expression of SREBP1c, a major adipogenic transcription factor [20–23]. The active form of SREBP1c is sensitive to proteasomal degradation but can enter the nucleus and participate in its transcriptional targets, including its own gene promoter and the promoter of the major enzyme encoding fatty acid synthesis [24]. However, a deeper understanding of lipid metabolism-related genes in the prognosis and treatment of pancreatic cancer is needed.

In this study, lipid metabolism-related gene expression in pancreatic cancer was analyzed to identify key genes that could predict patient prognosis. A differentially expressed gene (DEG) analysis, Weighted Gene Coexpression Network Construction Analysis (WGCNA) and Cox proportional risk model were used to finally construct a signature based on the expression of several key genes as a prognostic signature for pancreatic cancer. This prognostic model can be used as an effective tool to predict the prognosis of patients with pancreatic cancer. These findings will also help identify new therapeutic targets for pancreatic cancer.

Material and methods

Expression spectral data and preprocessing

Human lipid metabolics-related pathways were downloaded from Molecular Signature Database v7.0 [25], and a total of 776 genes related to lipid metabolism were sorted out from the six lipid metabolic pathways in Table 1. Pancreatic cancer RNA sequencing (RNA-seq) expression data and corresponding clinical follow-up data were obtained from the public database The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/) [26], which contained RNA-seq data of 182 patients and clinical information of 171 patients on December 3, 2019. GSE57495 is a microarray dataset from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) [27], containing expression profile data and clinical sample information from 62 patients with early pancreatic cancer. The International Cancer Genome Consortium (ICGC) validation dataset included 257 patients with pancreatic cancer with expression profile data and clinical follow-up information. For TCGA dataset, (a) samples without clinical data and overall survival (OS) <30 days were removed, (b) normal tissue sample data were removed, (c) genes with fragments per kilobase of

Table 1. Pathways involved in lipid metabolism.

Pathways	Database	Gene count
Regulation of lipid metabolism by peroxisome proliferator-activated receptor alpha	Reactome	119
Metabolism of lipids	Reactome	738
Sphingolipid metabolism	Reactome	89
Transcriptional regulation of white adipocyte differentiation	Reactome	84
Glycerophospholipid metabolism	KEGG	77
Fatty acid metabolism	Reactome	177
Sum		1,284 (unique: 776)

Table 2. The clinical information of four datasets.

Characteristics	TCGA set	Training set	P value	GSE57495 set	ICGC set
Age (years)					
<65	78	71	1.00	–	103
≥65	93	83		–	154
Progression-free survival					
Alive	80	69	1.00	21	151
Dead	91	85		41	106
Sex					
Female	78	72	1.00	–	120
Male	93	82		–	137
pathologic_T					
T1	7	7	0.09	–	–
T2	21	19		–	–
T3	138	123		–	–
T4/TX	4	4		–	–
pathologic_N					
N1	119	106	0.21	–	–
N0/NX	51	47		–	–
pathologic_M					
MX	90	81	0.199	–	–
M0/M1	81	72		–	–
Tumor stage					
Stage I	19	18	0.125	–	–
Stage II	142	127		–	–
Stage III	3	3		–	–
Stage IV	3	3		–	–
pathologic_G					
G1	28	33324	0.241	–	–
G2	92	84		–	–
G3	47	43		–	–
G4	2	2		–	–
Total	171	154	–	62	257

exon per million of zero in half of the samples were removed, and (d) the expression profile of genes related to lipid metabolism was preserved. For GEO datasets, (a) normal tissue sample data were removed, (b) OS data from months was converted to days, (c) samples with OS <30 days were removed, (d) Chip probes were mapped to the human gene SYMBOL using Bioconductor package, and (e) the expression profile of genes related to lipid metabolism was preserved. For the ICGC dataset, (a) sample data without survival status were removed, (b) samples with OS <30 days were removed, and (c) the expression profile of genes related to lipid metabolism was preserved. The GSE57495 and ICGC datasets were considered as the validation dataset. The clinical data information is shown in Table 2. The workflow chart is shown in Fig. 1.

Identification of prognostic genes

The expression profile of 776 lipid metabolism genes was extracted from TCGA dataset. However, 15 genes were not found. Furthermore, we keep genes that are not zero in more than half of the samples. As a result, 730 genes were used for subsequent analysis. Next, univariate Cox analysis of coxph function in R package was performed to obtain genes related to prognosis of pancreatic cancer with $P < 0.05$.

Identification of molecular subtypes

Cluster analysis of pancreatic cancer samples was performed by nonnegative matrix clustering algorithm (NMF), and the standard 'brunet' was selected by NMF method for 50 iterations. The clustering number k was set as 2–10, and the average contour width of the common member matrix was determined through the R package 'NMF'. The minimum member of each subclass was set as 10. According to the cophenetic, dispersion and silhouette index were used to determine the optimal clustering number.

Difference of tumor-infiltrating immune cells in molecular subtypes

Six types (B_cell, CD4_Tcell, CD8_Tcell, neutrophil, macrophage cell and dendritic cell) of tumor-infiltrating immune cell were retrieved from Tumor Immune Estimation Resource (https://cistrome.shinyapps.io/timer/) [28]. Immunity, matrix score, and tumor purity of each sample were calculated in R package estimate. These indicators were compared on molecular subtypes.

Identification of DEGs

R package differentially expressed Seq2 (DESeq2) [29] was applied to calculate the DEGs in molecular subtypes with a false discovery rate (FDR) <0.05 and $|\log_2 FC| > 1$.

WGCNA

Based on expression profiles of DEG, the WGCNA coexpression algorithm was used to mine the coexpression module using the R package WGCNA (http://www.r-project.org/) [30]. First, the appropriate soft threshold is determined by approximate scale-free topology criteria. The adjacency matrix was transformed into a topological matrix, and the genes were clustered using average-linkage hierarchical clustering. Lastly, the dynamic tree cut method was used to determine module eigengenes, at least 30 coexpressed genes.

Functional enrichment and pathway enrichment analysis

Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis

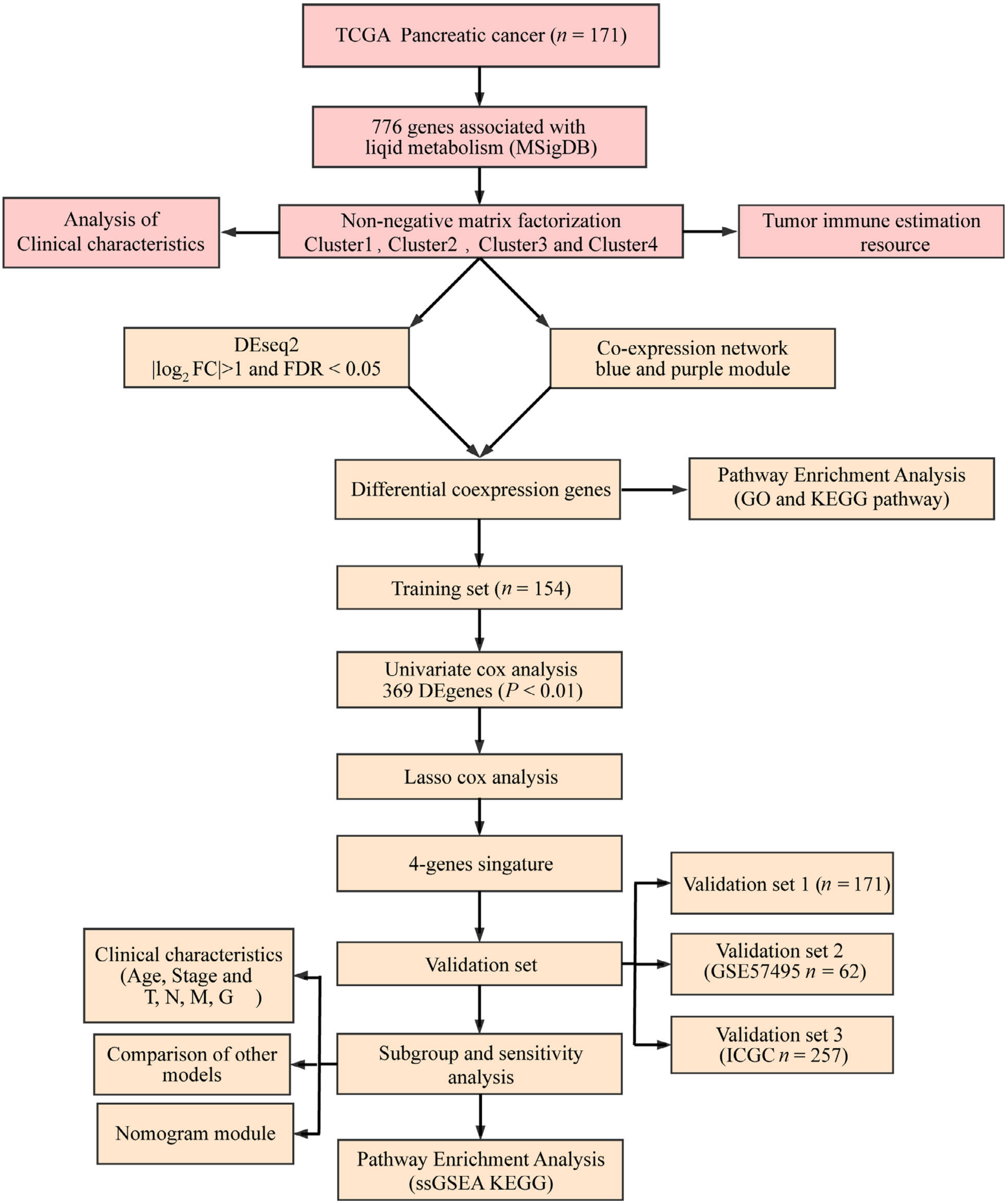

TCGA Pancreatic cancer (*n* = 171)
776 genes associated with liqid metabolism (MSigDB)
Analysis of Clinical characteristics
Non-negative matrix factorization Cluster1 , Cluster2 , Cluster3 and Cluster4
Tumor immune estimation resource
DEseq2 |log2 FC|>1 and FDR < 0.05
Co-expression network blue and purple module
Differential coexpression genes
Pathway Enrichment Analysis (GO and KEGG pathway)
Training set (*n* = 154)
Univariate cox analysis 369 DEgenes (*P* < 0.01)
Lasso cox analysis
4-genes singature
Validation set 1 (*n* = 171)
Validation set 2 (GSE57495 *n* = 62)
Validation set 3 (ICGC *n* = 257)
Validation set
Clinical characteristics (Age, Stage and T, N, M, G)
Comparison of other models
Nomogram module
Subgroup and sensitivity analysis
Pathway Enrichment Analysis (ssGSEA KEGG)

Fig. 1. Workflow chart. FC, fold change.

were performed for DEG based on 'WebGestaltR' [31] in R. FDR < 0.05 was defined as significant.

Construction of a prognostic risk model based on differentially coexpressed genes

First, 90% of samples were randomly selected from the preprocessed 171 TCGA samples as the training set for model construction. To avoid the random allocation bias affecting the stability of subsequent modeling, we repeatedly sampled 100 samples with replacement in advance to ensure that the randomly selected samples were consistent with all samples in age, stage and TNM staging. Univariate Cox regression analysis for OS was performed to identify prognostic DEGs with $P < 0.05$ using survival coxph function in R. Lasso (least absolute shrinkage and selection operator) Cox regression analysis was performed to find characteristic genes using R package glmnet. Subsequently, the multivariate Cox proportional hazards regression model was used to build a prognostic model in the training group. The risk formula was as follows: $\text{RiskScore}_4 = -0.0666 \times \mathit{CA8} + 0.0413 \times \mathit{CEP55} - 0.2189 \times \mathit{GNB3} - 0.0339 \times \mathit{SGSM2}$. Next, the Kaplan–Meier (KM) survival curve was used to compare prognosis between the low- and high-risk groups, which were classified by the median risk score as the cutoff value in all patients. The receiver operating characteristic (ROC) curve was applied to assess diagnostic accuracy through comparing the areas under the ROC curves (AUCs) using timeROC package in the training and validation groups.

Gene set enrichment analysis

The R software package Gene Set Variation Analysis (GSVA) [32] was used for single-sample gene set enrichment analysis (ssGSEA), and the function with correlation >0.45 was selected.

Advantages of genetic signatures

To identify the independence of four gene signatures, we used univariate and multivariate Cox regression to analyze the relationship among age, sex, pathological stage T, stage N, stage M, tumor stage, grade and RiskScore with prognosis. Next, by referring to the literature, we selected four prognostic risk models, 15-gene signature (Chen) [33], 7-gene signature (Cheng) [34], 5-gene signature (Raman) [35] and 9-gene signature (Wu) [36], for comparison with our 4-gene model. ROC curve and KM survival curve of four models were drawn in TCGA dataset. Furthermore, we compared the four models with the restricted mean survival (RMS) using R language RMS [37] and standardized net benefit between four models and four gene signatures.

Results

Identification of four molecular subtypes of pancreatic cancer

Univariate Cox survival analysis of lipid metabolism genes using coxph revealed 189 genes associated with the prognosis of pancreatic cancer. Pancreatic cancer samples were clustered by NMF algorithm, and the optimal number of clusters was determined to be four based on cophenetic, dispersion and silhouette metrics (Fig. 2A,B). The expression of lipid metabolism-related genes showed that the expression of C2 genes was lower than that of C1, C3 and C4 genes (Fig. 2C). Analysis of the prognostic relationship among the four subtypes showed that C1 had the worst prognosis and a significant difference [$P < 0.0001$; hazard ratio (HR), 2.264 (1.582–4.352)], and C2 had the best prognosis (Figs 2D and S1). Subsequently, we compared the clinical characteristics of the four subtypes and found significant differences in T, TNM, cancer grade and age (Fig. S2). The immune evaluation of the four subtypes showed significant differences in immune scores and stroma (Fig. S3).

Identification of differential expression genes and functional analysis

DESeq2 was conducted to calculate the DEGs between C1 and C2, C3 and C4 molecular subtypes; a total of 4249 were obtained (Fig. 3). Next, according to the expression profile of coding genes, the WGCNA coexpression algorithm was used to mine the coexpressed coding genes and coexpression modules, and hierarchical clustering analysis was performed on the samples to show that there were no outlier samples (Fig. 4A). A soft threshold of 10 was selected (Fig. 4B,C). Genes were clustered using the averages-linkage hierarchy clustering method, and 14 modules were obtained by setting height = 0.25, deepSplit = 2 and minModuleSize = 30 (Fig. 4D). The correlation of each module with sex, age, ethnicity and clusters 1, 2, 3 and 4 was further analyzed. The results showed that the modules significantly related to clusters 1, 2, 3 and 4 were magenta, green, brown and turquoise, respectively (Fig. 4E–H). Finally, 4249 DEGs and genes with significant coexpression modules were intersected to obtain 1340 differentially expressed cogenes. To determine whether differential genes are related to pancreatic cancer function, we performed GO molecular function and KEGG function enrichment analysis of 1340 differential genes using R software package WebGestaltR (https://www.r-project.org/help.

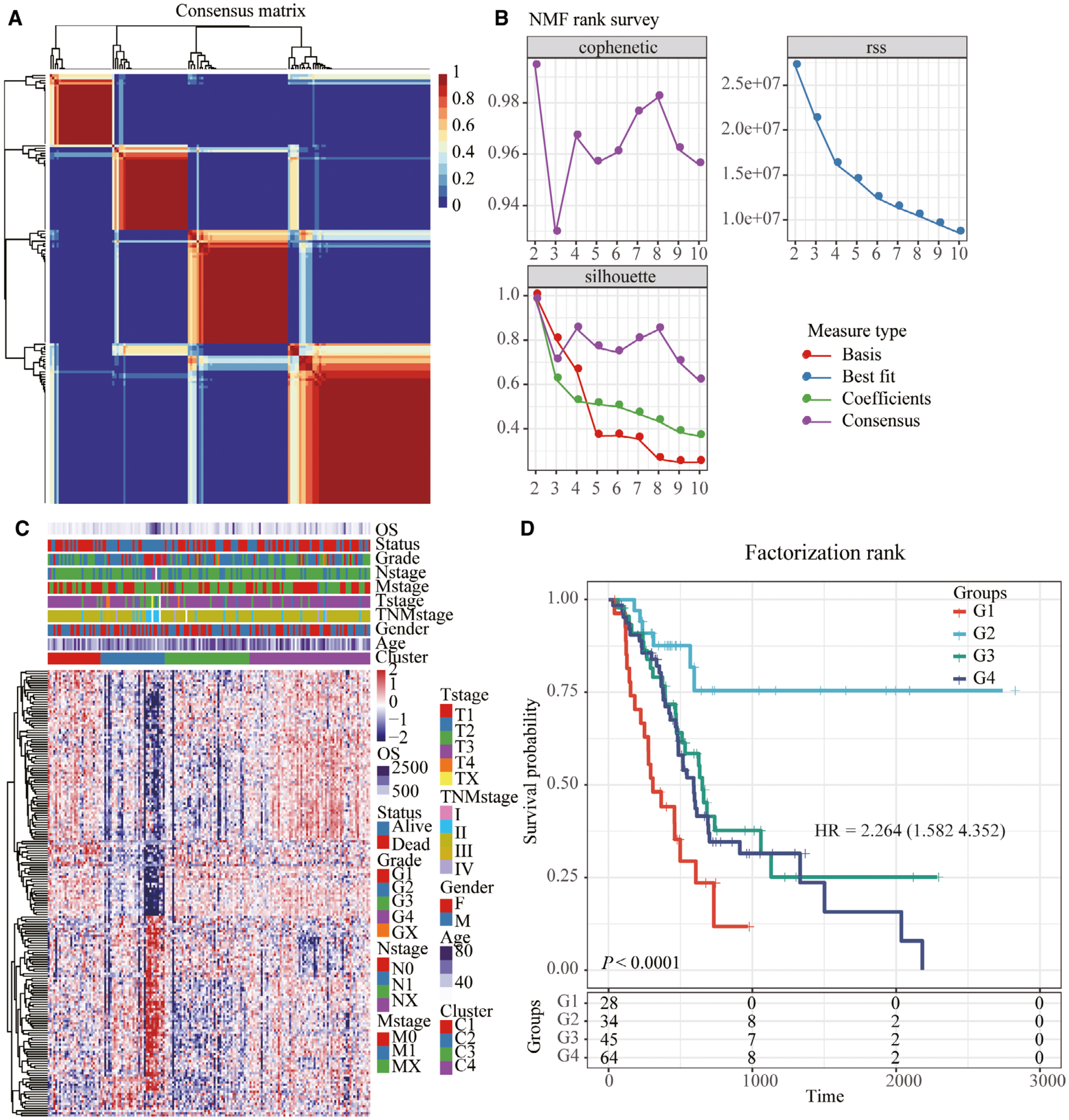


Fig. 2. Identification of molecular subtypes of pancreatic cancer. (A) Consensus map of NMF clustering. (B) The distribution of cophenetic, rss and dispersion with rank = 2–1. (C) Cluster heatmap of 740 lipid metabolism genes. (D) KM prognostic survival curve of molecular subtype.

html). The significant pathways enriched by KEGG are related to insulin secretion and dopaminergic synapse pathways (Fig. 5A). GO enriched 233 GO cellular component (CC) (Fig. 5B), 195 GO molecular function (MF) (Fig. 5C) and 977 GO biological process (BP) (Fig. 5D).

Construction and risk prediction of Gene Signature four-gene signature

Ninety percent of the 171 TCGA samples were randomly selected as the training set for model construction. Univariate Cox proportional hazard analysis was

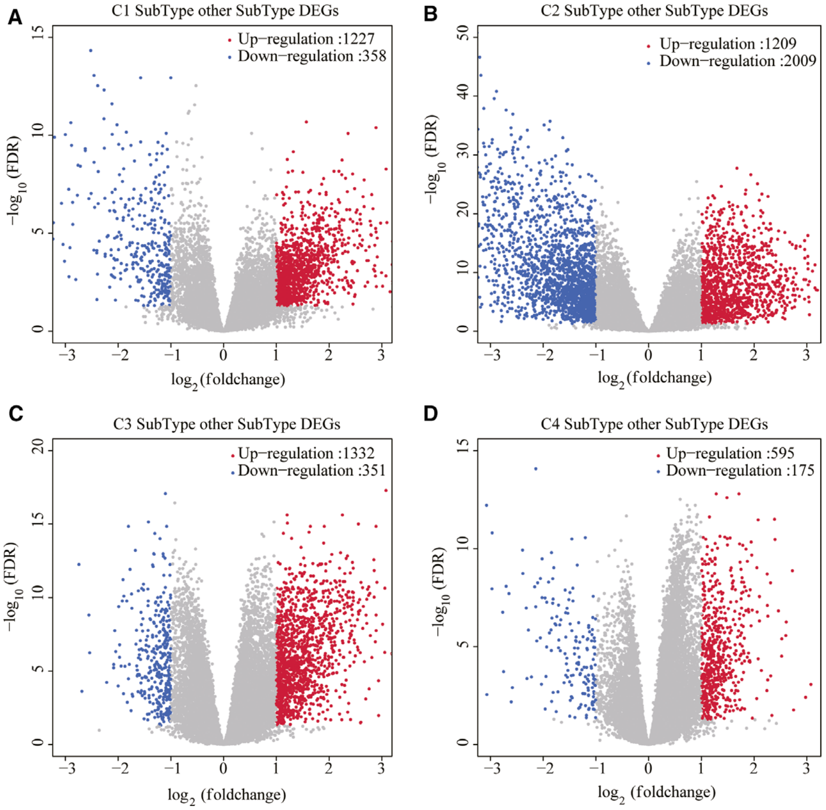


Fig. 3. Identification of DEG. (A) Volcano diagram of DEG in C1 subtype. (B) Volcano diagram of DEG in C2 subtype. (C) Volcano diagram of DEG in C3 subtype. (D) Volcano diagram of DEG in C4 subtype.

conducted for the expression profile of each differentially expressed cogene, and R package survival coxph function was used to obtain 369 genes with significant prognostic differences. To further narrow the range of genes and construct the prognostic model with high accuracy, we used R software package glmnet for Lasso Cox regression analysis. First, the analysis of the change trajectory of each independent variable shows that as the lambda gradually increases, the number of independent variable coefficients tending to zero also gradually increases (Fig. 6A). Then the confidence interval (CI) under each lambda is analyzed, and the model reaches the optimal value when lambda = 0.1753479 (Fig. 6B). Four genes were selected at lambda = 0.1753479 as target genes (Table 3). The four-mRNA signature formula is as follows:\def\mybox{\vrule depth -0.5mm height 4mm width 8mm}

$$\mathrm{RiskScore4} = -0.0666 \times \mathit{CA8} + 0.0413 \times \mathit{CEP55} - 0.2189 \times \mathit{GNB3} - 0.0339 \times \mathit{SGSM2}.$$

The RiskScore of each sample is calculated according to the expression level of the sample, and the RiskScore of the sample is plotted. The survival time of the samples with high RiskScore was significantly lower than that with low RiskScore. The gene expression changes with the increase of risk value showed that *CEP55* was a risk factor, whereas *CA8*, *GNB3* and *SGSM2* were protective factors (Fig. 6C). ROC analysis of RiskScore by R software package timeROC showed that AUC of 1, 3 and 5 years was >0.70 (Fig. 6D). Finally, we carried out *z*-score for RiskScore and divided the samples with *z*-score-based RiskScore greater than zero into the high-risk group and the samples with less than zero into the low-risk group. KM prognostic analysis showed a significant difference between the two groups (Fig. 6E).

Robustness of four-gene signature

To determine the robustness of the model, we used the whole TCGA dataset, GSE57495 dataset and ICGC verification set as the verification dataset, and the same model and coefficient as the training set were adopted. Similarly, the high RiskScore sample had a worse prognostic capability, *CEP55* was a risk factor, and *CA8*, *GNB3* and *SGSM2* were protective factors (Fig. 7A,D,G). ROC analysis showed that the model had high AUC (Fig. 7B,E,H). The results of the KM curve showed that there were significant marginal differences between the two groups (Fig. 7C,F,I). Moreover, we obtained three additional pancreatic cancer datasets from the GEO database, GSE28735,

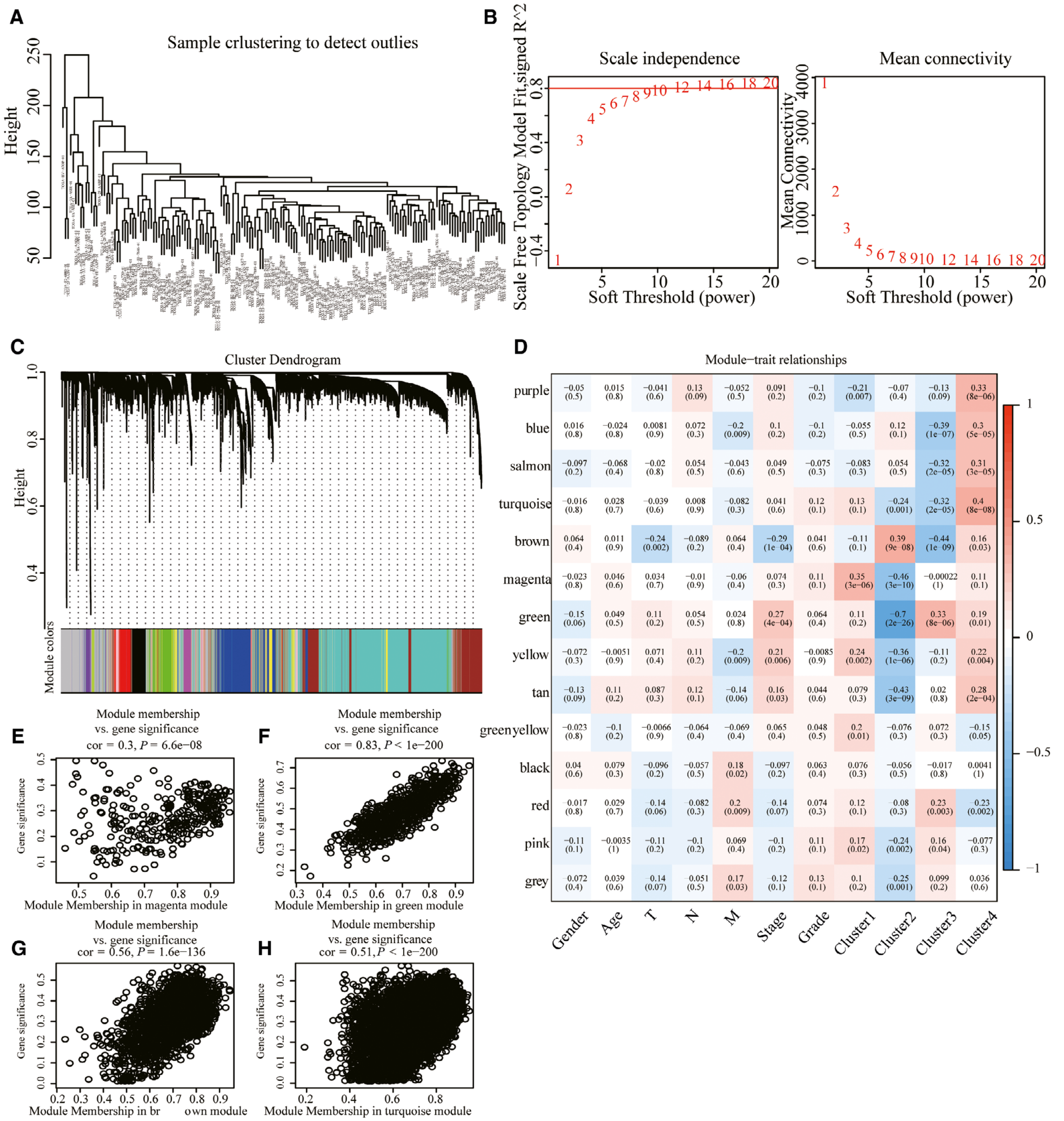


Fig. 4. WGCNA. (A) Cluster analysis of samples. (B) Analysis of network topology for various soft-thresholding powers. (C) Gene dendrogram and module colors. (D) Correlation between 14 modules and clinical phenotype. (E) Gene significance (*y* axis) versus module membership (*x* axis) plotted for magenta module in TCGA dataset. (F) Gene significance (*y* axis) versus module membership (*x* axis) plotted for green module in TCGA dataset. (G) Gene significance (*y* axis) versus module membership (*x* axis) plotted for brown module in TCGA dataset. (H) Gene significance (*y* axis) versus module membership (*x* axis) plotted for turquoise module in TCGA dataset.

GSE62452, and GSE85916, and we used the same method to score the risk for each patient in these three cohorts. First, we performed ROC analysis on the GSE28735 dataset. Because of the short follow-up period, it was not possible to calculate the 5-year AUC, in which the AUC of 1 and 3 years reached more than 0.78 (Fig. S4A), and there was a significant prognostic difference between the high- and low-risk groups (Fig. S4D). ROC analysis in the GSE62452 dataset showed the highest AUC of 0.77 for 3 years

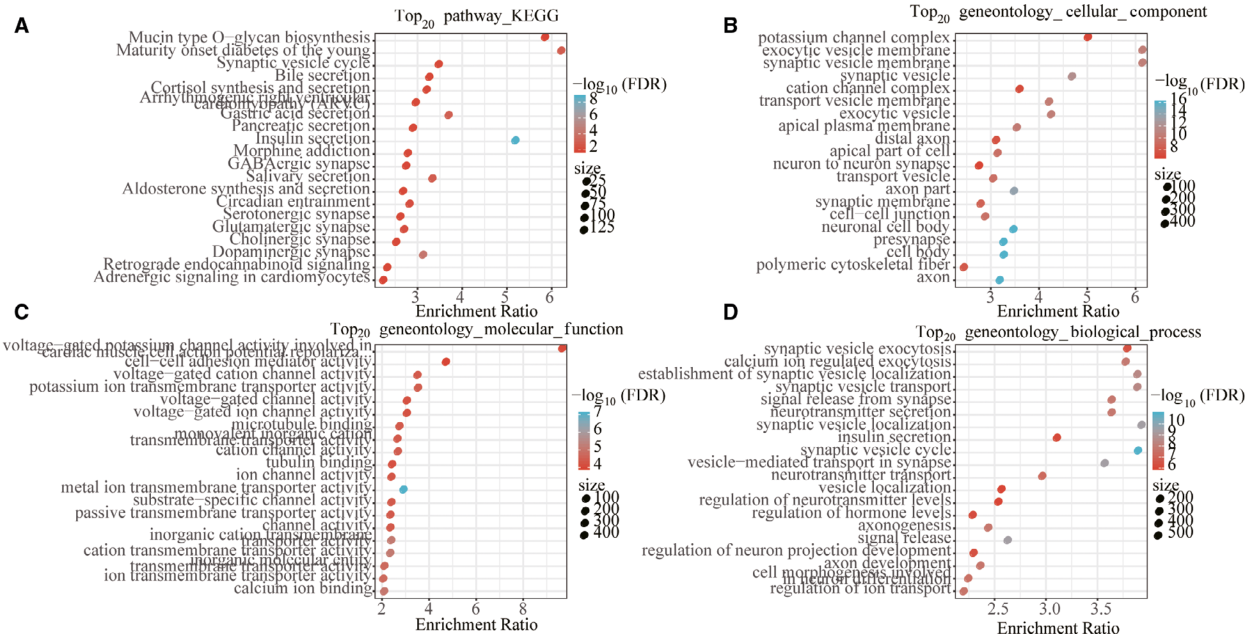


Fig. 5. Functional analysis of differentially coexpressed genes. (A) The DEGs were enriched to the top 20 enriched by KEGG. (B) The DEGs were enriched to the top 20 enriched by GO CC. (C) The DEGs were enriched to the top 20 enriched by GO MF. (D) The DEGs were enriched to the top 20 enriched by GO BP. The color from red to blue represents the significance of *P* value; the darker the red the smaller is the *P* value, and the dot size represents the number of genes enriched into the pathway (the higher the number, the larger the dot).

(Fig. S4B), with significant prognostic differences between the high- and low-risk groups (Fig. S4E). ROC analysis in the GSE85916 dataset showed the highest 5-year AUC of 0.84 (Fig. S4C), with a significant prognostic difference between the high- and low-risk groups (Fig. S4F).

Prognostic analysis of risk models and clinical features

Survival analysis showed that only age, N stage and OS were significantly correlated in the TCGA training set sample ($P < 0.05$), and TNM stage presented significant margin ($P = 0.05464$; Fig. 8). It was further found that four-mRNA signature could distinguish the young and old groups, female, stage I + II, T1 + T2 and T3 + T4 patients from high- and low-risk groups ($P < 0.05$; Fig. 9). These data further illustrate that our model still has good predictive ability in different clinical signs.

Clinical independence and regulatory pathway of four-mRNA signature

To identify the independence of the four-mRNA signature model in clinical applications, we used univariate and multivariate Cox regression analysis to analyze relevant HR, 95% CI of HR and *P* value in the clinical information carried by the whole TCGA data. In TCGA dataset, the univariate COX regression analysis found that sex, T3, T4 versus T1/T2, stage III versus stage I, II and IV, and RiskScore are significantly associated with survival, but the corresponding multivariate Cox regression analysis found that age, stage of N and risk score (HR, 3.606; 95% CI: 1.659–7.839; $P = 0.007$) were significantly associated with survival (Fig. 10A,B). The earlier conditions indicate that our model four-mRNA signature has good predictive performance in clinical application value. To observe the relationship between risk scores of different samples and biological function, we used the R software package GSVA for ssGSEA analysis. The function with a correlation >0.45 was selected, from which it can be seen that most of them are negatively correlated with the risk score of the sample, while a few are positively correlated with the risk score of the sample (Fig. 10C). Cluster analysis results showed that among the 17 pathways, KEGG_P53_SIGNALING_PATHWAY, KEGG_SYSTEMIC_LUPUS_ERYTHE MATOSUS, KEGG_CELL_CYCLE and other metabolic-related pathways increased with the increase of RiskScore, and KEGG_pentose_phosphoate_pathway declined with the increase of RiskScore (Fig. 10D). This also suggests that the dysfunction of these pathways is closely related to tumor development.

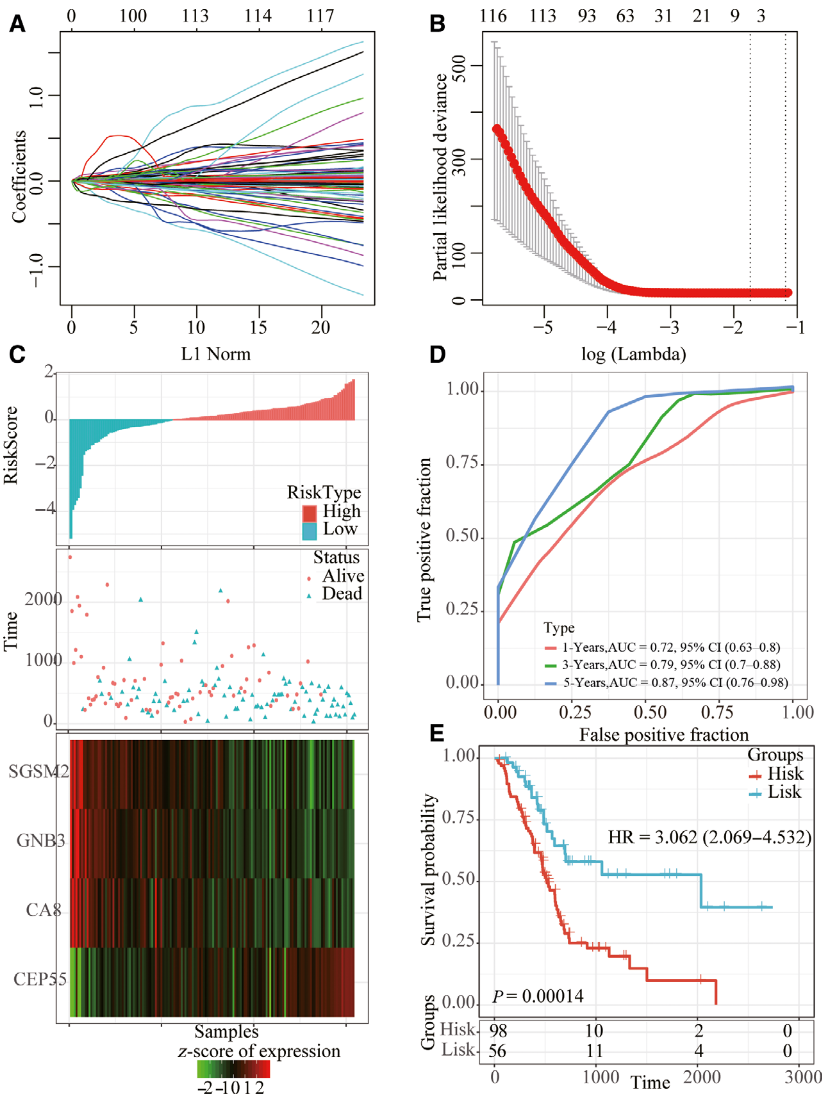


Fig. 6. Construction of risk model. (A) The CI for each lambda. (B) The trajectory of each independent variable, the log of lambda on the horizontal axis and the coefficient on the vertical axis. (C) Risk score, survival time and survival status, and expression of four genes in the training set. (D) ROC curve and AUC of four-gene signature in the training set. (E) KM survival curve of four-gene signature in the training set.

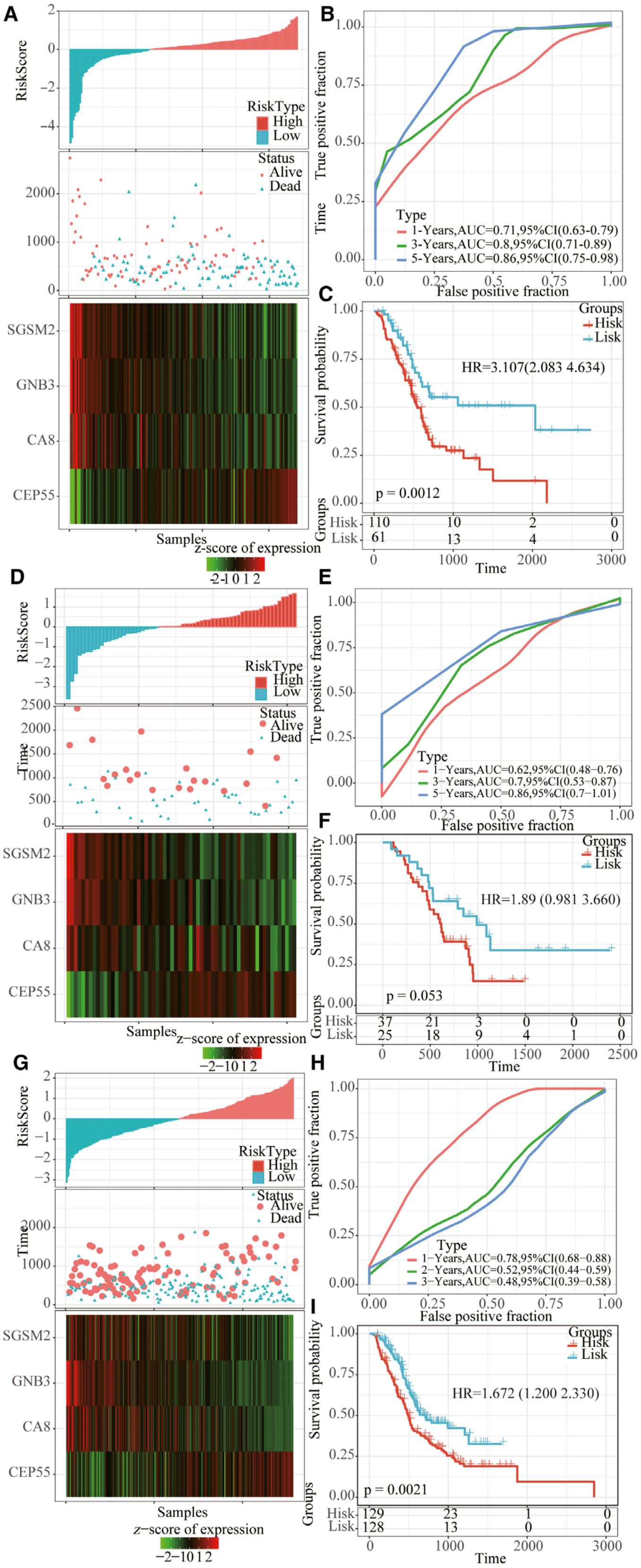
Table 3. Four-mRNA signature.

Symbol	Coefficient	HR	*z*-score	*P* value	Low 95% CI	High 95% CI
CA8	−0.06663	−2	0.045465	0.9355	0.8764	0.9987
CEP55	0.04133	3.639	0.000274	1.0422	1.0193	1.0657
GNB3	−0.21887	−1.397	0.162309	0.8034	0.591	1.0921
SGSM2	−0.03386	−1.6	0.10963	0.9667	0.9274	1.0077

Advantages of risk models

Four prognosis-related risk models, 15-gene signature (Chen), 7-gene signature (Cheng), 5-gene signature (Raman) and 9-gene signature (Wu), were selected and compared with our four-gene model. To make the models comparable, we calculated the risk scores of each pancreatic cancer sample in TCGA using the same

Fig. 7. Robustness of risk model. (A, D, G) Risk score, survival time and survival status, and expression of four genes in whole TCGA dataset, GSE57495 dataset and ICGC verification set. (B, E, H) ROC curve and AUC of four-gene signature in whole TCGA dataset, GSE57495 dataset and ICGC verification set. (C, F, I) KM survival curve of four-gene signature in whole TCGA dataset, GSE57495 dataset and ICGC verification set.


A
RiskScore
RiskType
High
Low
Status
Alive
Dead
Time
SGSM2
GNB3
CA8
CEP55
Samples
z-score of expression
-2-1 0 1 2
B
True positive fraction
False positive fraction
Type
1-Years,AUC=0.71,95%CI(0.63-0.79)
3-Years,AUC=0.8,95%CI(0.71-0.89)
5-Years,AUC=0.86,95%CI(0.75-0.98)
C
Survival probability
Groups
Hisk
Lisk
HR=3.107(2.083 4.634)
p = 0.0012
Hisk 110 10 2 0
Lisk 61 13 4 0
Time
D
E
Type
1−Years,AUC=0.62,95%CI(0.48−0.76)
3−Years,AUC=0.7,95%CI(0.53−0.87)
5−Years,AUC=0.86,95%CI(0.7−1.01)
F
HR=1.89 (0.981 3.660)
p = 0.053
Hisk 37 21 3 0 0 0
Lisk 25 18 9 4 1 0
G
H
Type
1−Years,AUC=0.78,95%CI(0.68−0.88)
2−Years,AUC=0.52,95%CI(0.44−0.59)
3−Years,AUC=0.48,95%CI(0.39−0.58)
I
HR=1.672 (1.200 2.330)
p = 0.0021
Hisk 129 23 1 0
Lisk 128 13 0 0

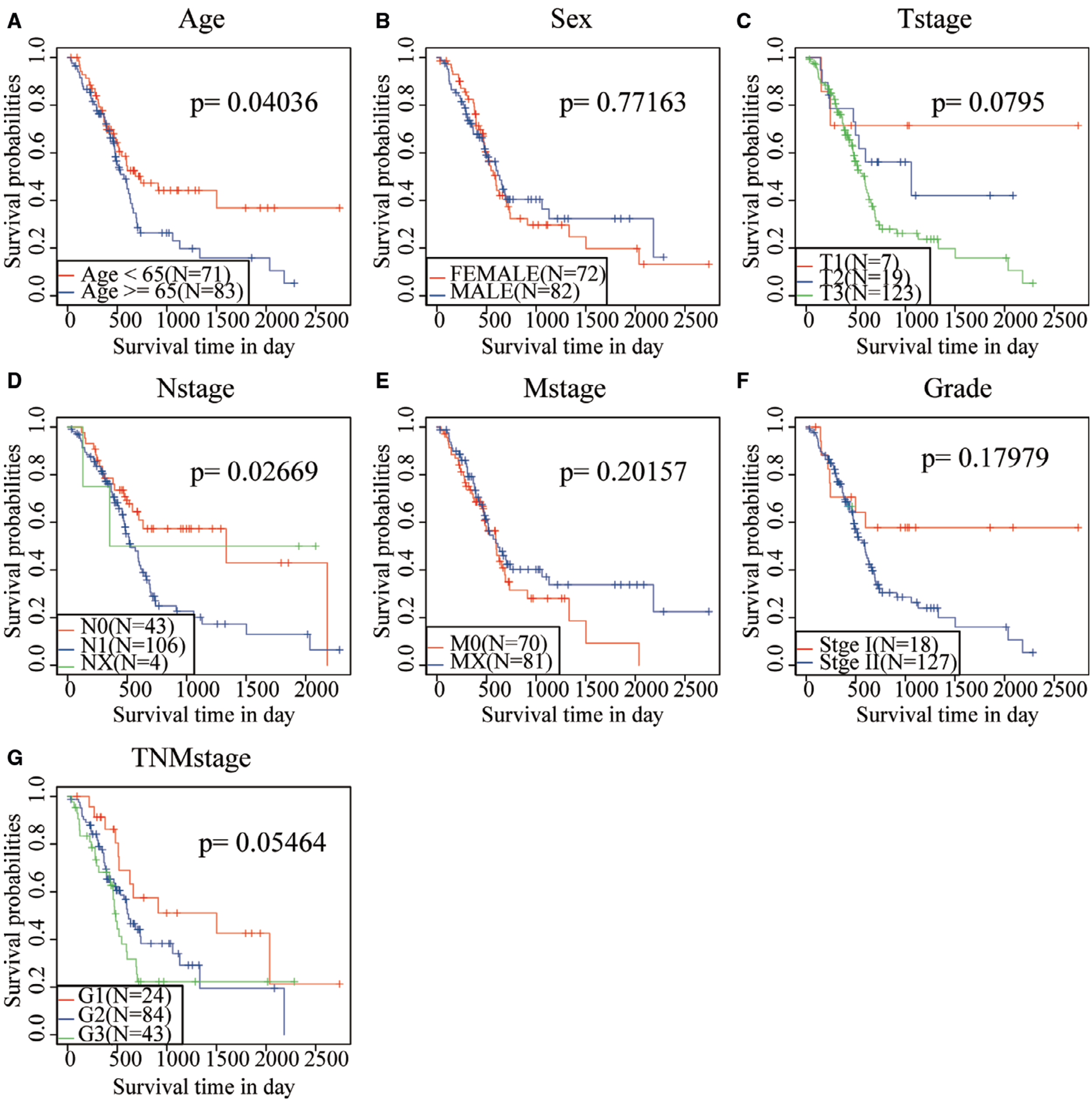


Fig. 8. Prognostic analysis of clinical characteristics and risk models. (A) KM prognosis curve in age samples. (B) KM prognosis curve in sex samples. (C) KM prognosis curve in T stage samples. (D) KM prognosis curve in N stage samples. (E) KM prognosis curve in M stage samples. (F) KM prognosis curve in grade samples. (G) KM prognosis curve in TNM stage samples.

method. Samples were divided into the risk-H and risk-L groups according to the median risk score, and the KM prognosis curve showed that there were significant differences in OS prognosis of samples from the four models in the risk-H and risk-L groups ($P < 0.05$; Fig. 11A–D). The ROC analysis results of the model showed that the prediction effect of the four models was worse than that of the four-gene signature models (Fig. 11E–H). The RMS curve was further drawn using the R language RMS package, indicating that the AUC of the four gene models was higher than that of the four models (Fig. 11I). Curves of risk coefficients also show that the model is relatively high (Fig. 11J).

Discussion

Pancreatic cancer is by default the king of all cancers, placing a heavy burden on families and the world [38].

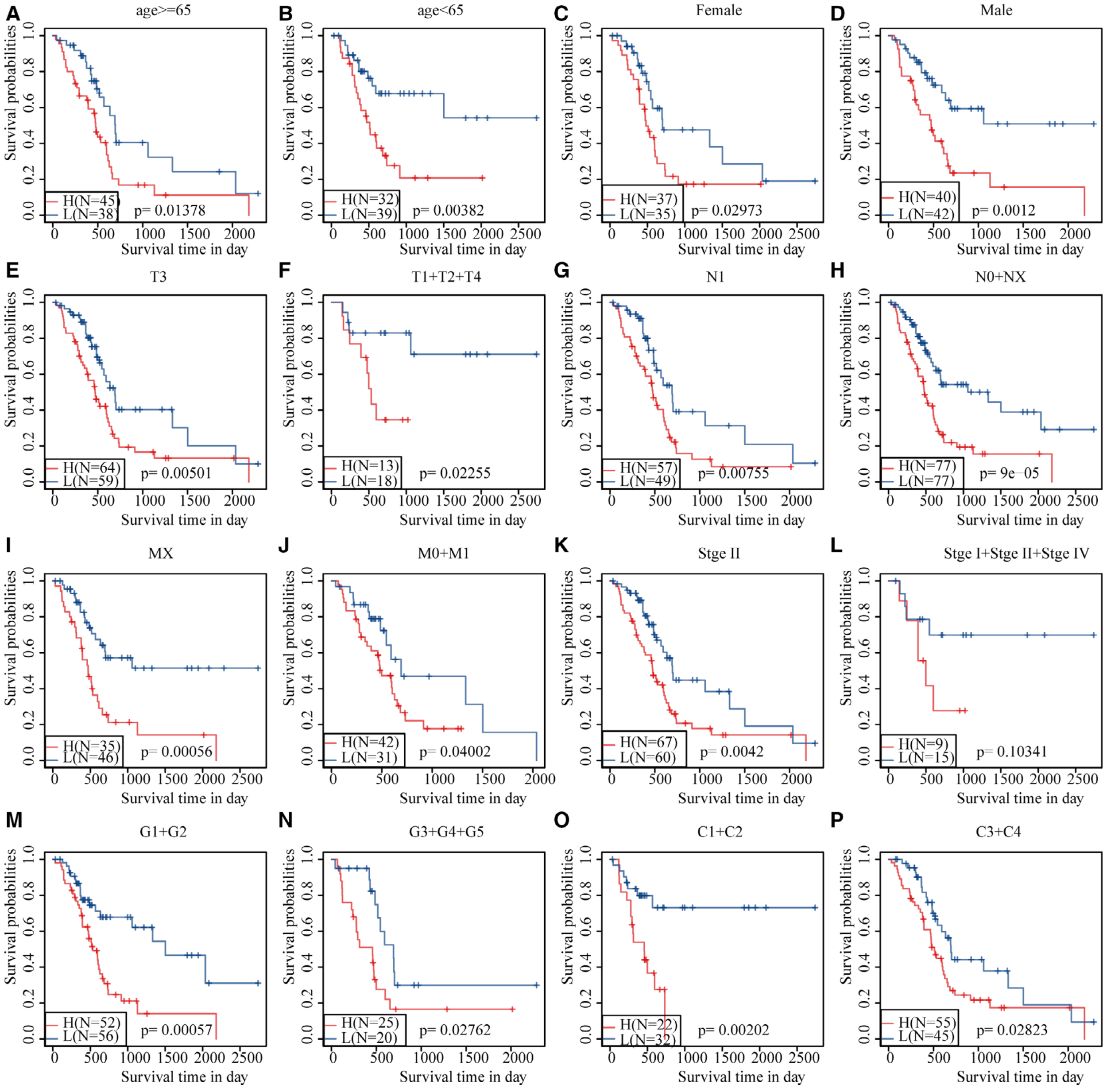


Fig. 9. Prognostic analysis of clinical characteristics and risk models. (A) KM prognosis curve in young samples (age ≤ 65 years). (B) KM prognosis curve in old samples (age > 65 years). (C) KM prognosis curve in female samples. (D) KM prognosis curve in male samples. (E) KM prognosis curve in T3 stage samples. (F) KM prognosis curve in T1 + T2 + T4 stage samples. (G) KM prognosis curve in N1 stage samples. (H) KM prognosis curve in N0 + NX stage samples. (I) KM prognosis curve in MX stage samples. (J) KM prognosis curve in M0 + M1 stage samples. (K) KM prognosis curve in stage II samples. (L) KM prognosis curve in stage I + II + IV samples. (M) KM prognosis curve in G1 + G2 stage samples. (N) KM prognosis curve in G3 + G4 + G5 stage samples. (O) KM prognosis curve in C1 + C2 samples. (P) KM prognosis curve in C3 + C4 samples.

Accurate prediction of the prognosis of pancreatic cancer is important for the choice of treatment and the improvement of the prognosis. In this study, four independent datasets were used to identify DEGs associated with lipid metabolism between pancreatic cancer and normal cervical tissue. A total of 189 DEGs were identified, and four molecular isoforms were identified based on lipid metabolism genes combined with WGCNA to finally identify four gene signatures.

The prognostic value of various gene expression profiles has been studied in pancreatic cancer over the past decade. For example, Demirkol Canli *et al.* [39]

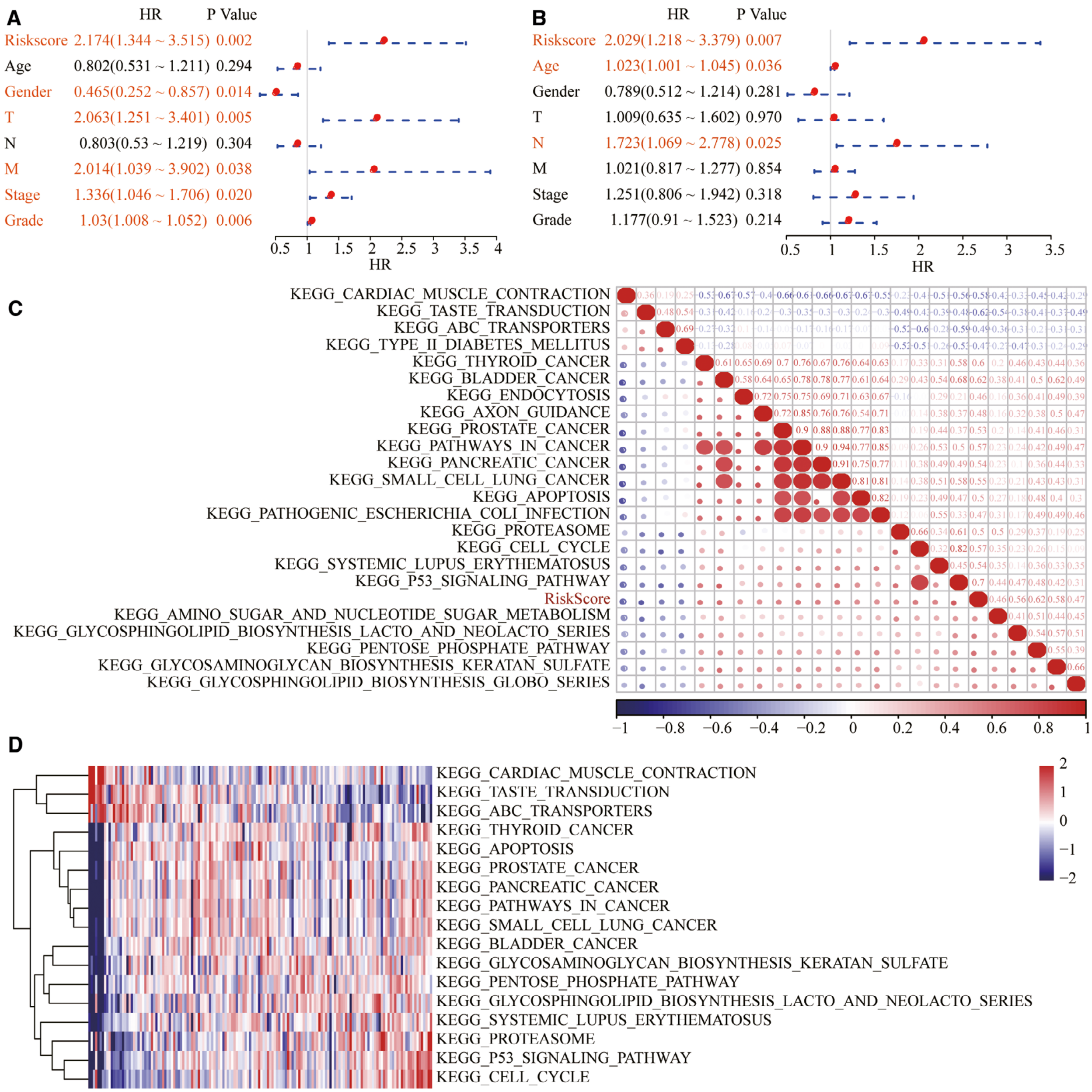


Fig. 10. Clinical independence and regulatory pathway of four-mRNA signature. (A) Univariate Cox regression analysis of four-mRNA signature. (B) Multivariate Cox regression analysis of four-mRNA signature. (C) Clustering of correlation coefficients between KEGG pathways and RiskScore with a correlation >0.45. (D) The change of ssGSEA score in each sample with the increase of risk score in the KEGG pathway. The horizontal axis represents the sample, and the risk score increases from left to right.

identified a gene signature composed of 20 prognostic genes (PPS20), indicating OS and event-free survival of pancreatic cancer. Wolfe *et al.* [40] developed a four-miRNA molecular signature that is associated with risk for local-regional recurrence and OS after pancreatic cancer resection. Meng *et al.* [41] constructed a novel eight-mRNA signature to predict the prognosis of PAAD patients by applying ESTIMATE scoring to RNA-seq-based transcriptome data. Chen *et al.* [33] developed a prognostic 15-gene signature to know OS by analyzing microarray data from 63 patients with early pancreatic ductal adenocarcinoma (PDAC) (stages IB, IIA and IIB) in the Moffitt cohort. Although these reports are promising, the proposed genome is either too large or traditionally genetically composed. Therefore, focusing on key biological

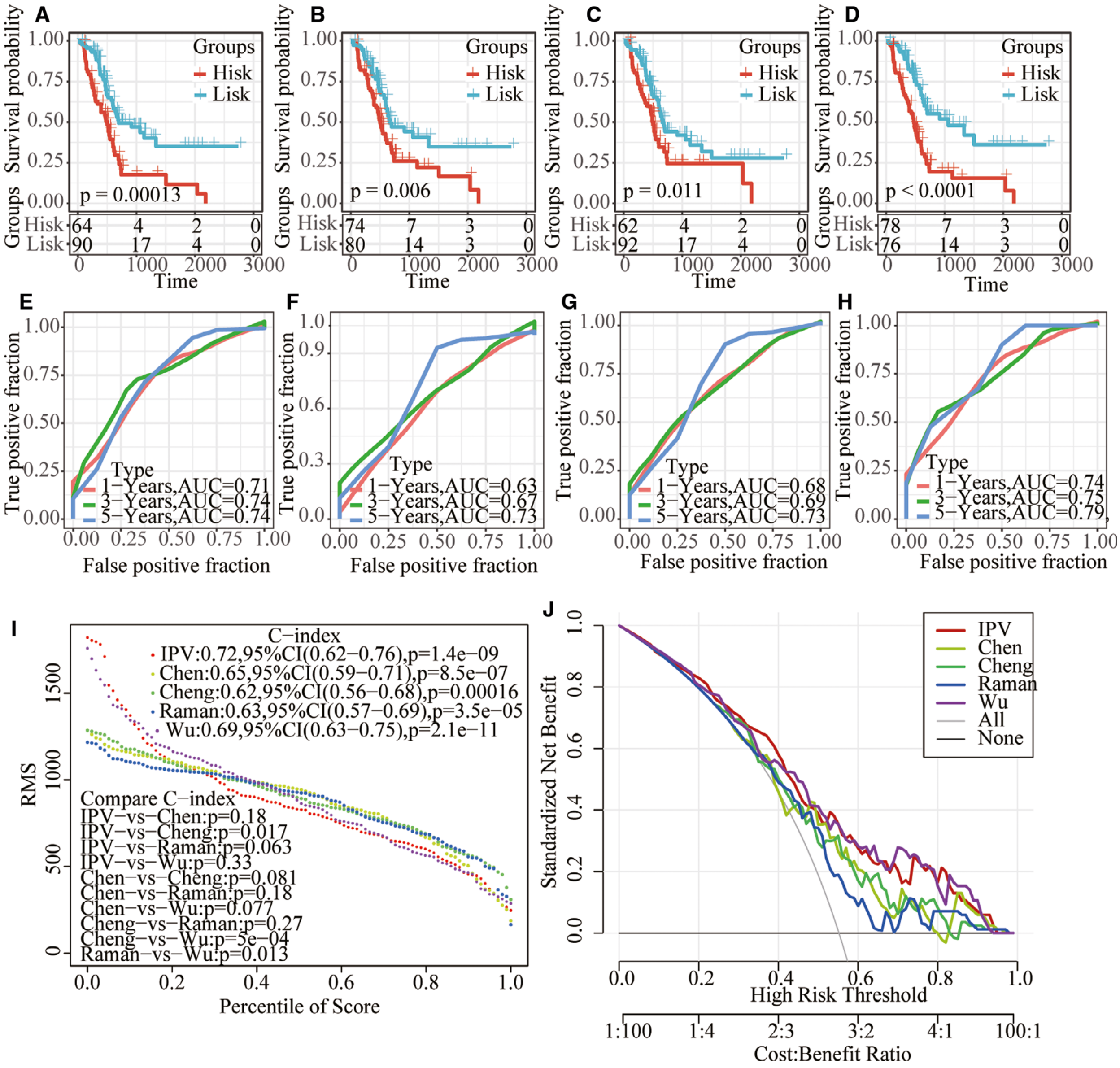


Fig. 11. The superiority of the risk model. (A–D) KM prognosis curve of four models. (E–H) AUC of 1, 3, 5 and 3 years in four models. (I) RMS curves of five models. (J) Risk coefficient curves of five models.

processes, such as lipid metabolism, which have a significant impact on cancer occurrence and progression, may introduce potential biomarkers for pancreatic cancer screening, as well as new treatment strategies and targets.

Abnormalities in signaling pathways are one of the important advances in cancer. With this in mind, abnormal activity of energy metabolism, such as lipid metabolism, has a unique role in cancer development, and its expression and active state have become of interest to researchers for screening and therapeutic inventions [42,43]. Based on this, we used a bioinformatics approach to mine the correlation between lipid gene status and prognostic prediction in patients with pancreatic cancer. We found significant aberrant expression of *CEP55*, *CA8*, *GNB3* and *SGSM2* as the key dysregulated metabolic factors within the study population. Overexpression of *CEP55* activates p21 and enhances the cell-cycle transition. Also, *CEP55* upregulation promotes PANC cell aggressiveness via activating pancreatic cancer [44]. Many studies have shown that *CA8* is associated with poor prognosis in a variety of tumors, [45–47], but it has not been reported in pancreatic cancer. *GNB3* was reported to influence development of metastasis in low-grade tumors [48]. *SGSM2* downregulation promoted estrogen receptor-positive breast cancer cell migration via modulating cell adhesion and cytoskeleton dynamics [49]. These published reports reinforce the potential of these genes as a comprehensive prognosis. We screened prognostic genes from lipid metabolism-related genes and divided four molecular subtypes to select four genes that are likely to be involved in lipid metabolic processes, although these genes have not been studied in depth in lipid metabolism studies. On this basis, we suggest that four gene signatures are likely to serve as prognostic biological indicators of pancreatic cancer, and that these genes may be involved in important lipid metabolism processes.

Inevitably, there are some limitations in the research, and we hope to address these in future work. First, although three study cohorts were included in this study, our findings should be confirmed in a separate cohort. Second, the prognostic value of lipid metabolism genes was studied using gene microarrays, and this single assay should also be validated by other methods, such as real-time quantitative RT-PCR. Third, the majority of genes in our prognostic model have not been reported in studies related to lipid metabolism. Their specific clinical significance, biological function and potential mechanism of action should be studied in further experiments. In conclusion, more experimental evidence is needed to determine the function of these genes in pancreatic cancer.

Conclusions

Our study found that four lipid metabolism-related genes were significantly associated with prognosis in patients with pancreatic cancer; therefore, the four-gene signature with some clinicopathological characteristics could be a useful biomarker for the prognosis of pancreatic cancer.

Conflict of interest

The authors declare no conflict of interest.

Data Accessibility

The analyzed datasets generated during the study are available from the corresponding author on reasonable request.

Author contributions

YY and HW conceived and designed the research. YQ drafted the manuscript and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. ZC contributed to date acquisition. YS analyzed data. YY interpreted data. HW revised the manuscript for important intellectual content. All authors approved the final version to be published.

References

1 Murakami T, Hiroshima Y, Matsuyama R, Homma Y, Hoffman RM and Endo I (2019) Role of the tumor microenvironment in pancreatic cancer. *Ann Gastroenterol Surg* **3**, 130–137.

2 Petrushnko W, Gundara JS, De Reuver PR, O'Grady G, Samra JS and Mittal A (2016) Systematic review of peri-operative prognostic biomarkers in pancreatic ductal adenocarcinoma. *HPB* **18**, 652–663.

3 Yadav D and Lowenfels AB (2013) The epidemiology of pancreatitis and pancreatic cancer. *Gastroenterology* **144**, 1252–1261.

4 Ji BL, Xia LP, Zhou FX, Mao GZ and Xu LX (2016) Aconitine induces cell apoptosis in human pancreatic cancer via NF-kappaB signaling pathway. *Eur Rev Med Pharmacol Sci* **20**, 4955–4964.

5 Hidalgo M (2010) Pancreatic cancer. *New Engl J Med* **362**, 1605–1617.

6 Zhang SH, Liu GF, Li XF, Liu L and Yu SN (2018) Efficacy of different chemotherapy regimens in treatment of advanced or metastatic pancreatic cancer: a network meta-analysis. *J Cell Physiol* **233**, 3352–3374.

7 Jemal A, Bray F, Center MM, Ferlay J, Ward E and Forman D (2011) Global cancer statistics. *CA Cancer J Clin* **61**, 69–90.

8 Pompeia C, Lopes LR, Miyasaka CK, Procopio J, Sannomiya P and Curi R (2000) Effect of fatty acids on leukocyte function. *Braz J Med Biol Res* **33**, 1255–1268.

9 Jarc E and Petan T (2019) Lipid droplets and the management of cellular stress. *Yale J Biol Med* **92**, 435–452.

10 Cheng C, Geng F, Cheng X and Guo D (2018) Lipid metabolism reprogramming and its potential targets in cancer. *Cancer Commun* **38**, 27.

11 Baenke F, Peck B, Miess H and Schulze A (2013) Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development. *Dis Models Mech* **6**, 1353–1363.

12 Medes G, Thomas A and Weinhouse S (1953) Metabolism of neoplastic tissue. IV. A study of lipid synthesis in neoplastic tissue slices *in vitro*. *Cancer Res* **13**, 27–29.

13 Rysman E, Brusselmans K, Scheys K, Timmermans L, Derua R, Munck S, Van Veldhoven PP, Waltregny D, Daniels VW, Machiels J *et al.* (2010) De novo

lipogenesis protects cancer cells from free radicals and chemotherapeutics by promoting membrane lipid saturation. *Cancer Res* **70**, 8117–8126.

14 Hanahan D and Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674.

15 Calvisi DF, Wang C, Ho C, Ladu S, Lee SA, Mattu S, Destefanis G, Delogu S, Zimmermann A, Ericsson J *et al.* (2011) Increased lipogenesis, induced by AKT-mTORC1-RPS6 signaling, promotes development of human hepatocellular carcinoma. *Gastroenterology* **140**, 1071–1083.

16 Kuhajda FP (2006) Fatty acid synthase and cancer: new application of an old pathway. *Cancer Res* **66**, 5977–5980.

17 Menendez JA and Lupu R (2007) Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nat Rev Cancer* **7**, 763–777.

18 Swinnen JV, Brusselmans K and Verhoeven G (2006) Increased lipogenesis in cancer cells: new players, novel targets. *Curr Opin Clin Nutr Metab Care* **9**, 358–365.

19 Sunami Y, Rebelo A and Kleeff J (2017) Lipid metabolism and lipid droplets in pancreatic cancer and stellate cells. *Cancers* **10**, 3.

20 Guri Y, Colombi M, Dazert E, Hindupur SK, Roszik J, Moes S, Jenoe P, Heim MH, Riezman I, Riezman H *et al.* (2017) mTORC2 promotes tumorigenesis via lipid synthesis. *Cancer Cell* **32**, 807–823.e12.

21 Owen JL, Zhang Y, Bae SH, Farooqi MS, Liang G, Hammer RE, Goldstein JL and Brown MS (2012) Insulin stimulation of SREBP-1c processing in transgenic rat hepatocytes requires p70 S6-kinase. *Proc Natl Acad Sci USA* **109**, 16184–16189.

22 Yecies JL, Zhang HH, Menon S, Liu S, Yecies D, Lipovsky AI, Gorgun C, Kwiatkowski DJ, Hotamisligil GS, Lee CH *et al.* (2011) Akt stimulates hepatic SREBP1c and lipogenesis through parallel mTORC1-dependent and independent pathways. *Cell Metab* **14**, 21–32.

23 Düvel K, Yecies JL, Menon S, Raman P, Lipovsky AI, Souza AL, Triantafellow E, Ma Q, Gorski R, Cleaver S *et al.* (2010) Activation of a metabolic gene regulatory network downstream of mTOR complex 1. *Mol Cell* **39**, 171–183.

24 Horton JD, Goldstein JL and Brown MS (2002) SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. *J Clin Investig* **109**, 1125–1131.

25 Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P and Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740.

26 (2018) The TCGA Legacy. *Cell* **173**, 281–282.

27 Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP *et al.* (2010) Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* **20**, 440–446.

28 Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, Li B and Liu XS (2017) TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res* **77**, e108–e110.

29 Li Z, Jiang C and Yuan Y (2019) TCGA based integrated genomic analyses of ceRNA network and novel subtypes revealing potential biomarkers for the prognosis and target therapy of tongue squamous cell carcinoma. *PLoS One* **14**, e0216834.

30 Langfelder P and Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559.

31 Liao Y, Wang J, Jaehnig EJ, Shi Z and Zhang B (2019) WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* **47**, W199–W205.

32 Hanzelmann S, Castelo R and Guinney J (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7.

33 Chen DT, Davis-Yadley AH, Huang PY, Husain K, Centeno BA, Permuth-Wey J, Pimiento JM and Malafa M (2015) Prognostic fifteen-gene signature for early stage pancreatic ductal adenocarcinoma. *PLoS One* **10**, e0133562.

34 Cheng Y, Wang K, Geng L, Sun J, Xu W, Liu D, Gong S and Zhu Y (2019) Identification of candidate diagnostic and prognostic biomarkers for pancreatic carcinoma. *EBioMedicine* **40**, 382–393.

35 Raman P, Maddipati R, Lim KH and Tozeren A (2018) Pancreatic cancer survival analysis defines a signature that predicts outcome. *PLoS One* **13**, e0201751.

36 Wu M, Li X, Zhang T, Liu Z and Zhao Y (2019) Identification of a nine-gene signature and establishment of a prognostic nomogram predicting overall survival of pancreatic cancer. *Front Oncol* **9**, 996.

37 Eng KH, Schiller E and Morrell K (2015) On representing the prognostic value of continuous gene expression biomarkers with the restricted mean survival curve. *Oncotarget* **6**, 36308–36318.

38 Makar M, Worple E, Dove J, Hunsinger M, Arora T, Oxenberg J and Blansfield JA (2019) Disparities in care: impact of socioeconomic factors on pancreatic surgery: exploring the national cancer database. *Am Surg* **85**, 327–334.

39 Demirkol Canli S, Dedeoglu E, Akbar MW, Kucukkaraduman B, Isbilen M, Erdogan OS, Erciyas SK, Yazici H, Vural B and Gure AO (2020) A novel 20-gene prognostic score in pancreatic adenocarcinoma. *PLoS One* **15**, e0231835.

40 Wolfe AR, Wald P, Webb A, Sebastian N, Walston S, Robb R, Chen W, Vedaie M, Dillhoff M, Frankel WL *et al.* (2020) A microRNA-based signature predicts

local-regional failure and overall survival after pancreatic cancer resection. *Oncotarget* **11**, 913–923.

41 Meng Z, Ren D, Zhang K, Zhao J, Jin X and Wu H (2020) Using ESTIMATE algorithm to establish an 8-mRNA signature prognosis prediction system and identify immunocyte infiltration-related genes in Pancreatic adenocarcinoma. *Aging* **12**, 5048–5070.

42 Qin C, Yang G, Yang J, Ren B, Wang H, Chen G, Zhao F, You L, Wang W and Zhao Y (2020) Metabolism of pancreatic cancer: paving the way to better anticancer strategies. *Mol Cancer* **19**, 50.

43 Huang KC, Li J, Zhang C, Tan Y and Cheng JX (2020) Multiplex stimulated raman scattering imaging cytometry reveals lipid-rich protrusions in cancer cells under stress condition. *iScience* **23**, 100953.

44 Peng T, Zhou W, Guo F, Wu HS, Wang CY, Wang L and Yang ZY (2017) Centrosomal protein 55 activates NF-kappaB signalling and promotes pancreatic cancer cells aggressiveness. *Sci Rep* **7**, 5925.

45 Nishikata M, Nishimori I, Taniuchi K, Takeuchi T, Minakuchi T, Kohsaki T, Adachi Y, Ohtsuki Y and Onishi S (2007) Carbonic anhydrase-related protein VIII promotes colon cancer cell growth. *Mol Carcinog* **46**, 208–214.

46 Akisawa Y, Nishimori I, Taniuchi K, Okamoto N, Takeuchi T, Sonobe H, Ohtsuki Y and Onishi S (2003) Expression of carbonic anhydrase-related protein CA-RP VIII in non-small cell lung cancer. *Virchows Arch* **442**, 66–70.

47 Lu SH, Takeuchi T, Fujita J, Ishida T, Akisawa Y, Nishimori I, Kohsaki T, Onishi S, Sonobe H and Ohtsuki Y (2004) Effect of carbonic anhydrase-related protein VIII expression on lung adenocarcinoma cell growth. *Lung Cancer* **44**, 273–280.

48 Krippl P, Langsenlehner U, Renner W, Yazdani-Biuki B, Wolf G, Wascher TC, Paulweber B and Samonigg H (2004) The 825C>T polymorphism of the G-protein beta-3 subunit gene (GNB3) and breast cancer. *Cancer Lett* **206**, 59–62.

49 Lin JH, Lee WJ, Wu HC, Wu CH, Chen LC, Huang CC, Chang HL, Cheng TC, Chang HW, Ho CT *et al.* (2019) Small G protein signalling modulator 2 (SGSM2) is involved in oestrogen receptor-positive breast cancer metastasis through enhancement of migratory cell adhesion via interaction with E-cadherin. *Cell Adh Migr* **13**, 120–137.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. KM prognosis curves of four molecular subtypes. (A) KM curve between C1 and C2 molecular subtypes. (B) KM curve between C1 and C3 molecular subtypes. (C) KM curve between C1 and C4 molecular subtypes. (D) KM curve between C2 and C3 molecular subtypes. (E) KM curve between C2 and C4 molecular subtypes. (F) KM curve between C3 and C4 molecular subtypes.

Fig. S2. Comparison of clinical characteristics of molecular subtypes. (A) Sample distribution of T stages in four subtypes. (B) Sample distribution of N stages in four subtypes. (C) Sample distribution of M stages in four subtypes. (D) Sample distribution of TNM stages in four subtypes. (E) Sample distribution of tumor stages in four subtypes. (F) Sample distribution of sex stages in four subtypes. (G) Sample distribution of age stages in four subtypes.

Fig. S3. Comparison of immune characteristics in molecular subtypes. (A) B cell score between molecular subtypes. (B) CD4 cell score between molecular subtypes. (C) CD8 cell score between molecular subtypes. (D) Neutrophil cell score between molecular subtypes. (E) Macrophage cell score between molecular subtypes. (F) Dendritic cell score between molecular subtypes. (G) Stromal score between molecular subtypes. (H) Est_Immune score between molecular subtypes. (I) ESTIMATE score between molecular subtypes.

Fig. S4. Prognostic ability of four-gene signature. (A, D) ROC curve and KM curve of RiskScore in GSE28735 dataset. (B, E) ROC curve and KM curve of RiskScore in GSE62452 dataset. (C, F) ROC curve and KM curve of RiskScore in GSE85916 dataset.