Development and clinical validation of A Novel 9-gene prognostic biomarkers based on multi-omics in pancreatic adenocarcinoma

Dafeng Xu (dfx0898@163.com)
Hainan General Hospital https://orcid.org/0000-0002-8839-8433
Yu Wang
Hainan General Hospital
Xiangmei Liu
Hainan General Hospital
Kailun Zhou
Hainan General Hospital
Jincai Wu
Hainan General Hospital
Jiacheng Chen
Hainan General Hospital
Cheng Chen
Hainan General Hospital
Liang Chen
Hainan General Hospital
Jinfang Zheng
Hainan General Hospital

Primary research

Keywords: PAAD, multi-omics, CNV, prognostic signature

Posted Date: August 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-49603/v1

Abstract

Background: The prognosis of patients with pancreatic cancer remains poor due to the lack of biomarkers for early diagnosis and effective prognosis.

Methods: RNA-Seq, SNP, CNV data were downloaded from The Cancer Genome Atlas(TCGA); Univariant cox regression was used to identify prognosis-related genes; GISTIC 2.0 was used to identify significantly amplified or deleted genes; Lasso method was used to construct risk prognosis model, which was then validated in GEO and ICGC cohorts. *rms* package was used to evaluate the overall predictive performance of the model by calculating and comparing the C-index values with other models. Experiments of western blot were performed to evaluate the expression of genes.

Results: 54 candidate genes were obtained after integrating the genomic mutated genes and prognosis-related genes. The Lasso method was conducted to finally ascertain nine characteristic genes, including UNC13B, TSPYL4, MICAL1, KLHDC7B, KLHL32, AIM1, ARHGAP18, DCBLD1, and CACNA2D4. The 9-gene signature model can help with stratifying samples at risk in the training cohort and external validation cohort. In addition, the overall predictive performance of our model is superior to the others. We found that AIM1 and DCBLD1 were highly expressed in tumor tissues, while ARHGAP18, CACNA2D4, and TSPYL4 were lowly expressed in tumor tissues at the protein and transcription levels.

Conclusion: The nine-gene signatures constructed in this study can be used as the novel prognostic markers to predict the survival of PAAD patients.

Background

Currently, pancreatic adenocarcinoma (PAAD) is one of the most deadly tumors, claiming approximately 400,000 lives annually around the world. With a 5-year survival rate after diagnosis lower than 5% [1; 2], PAAD has become the fourth leading cause of death among cancer patients in the United States. The direct cause of such poor prognosis is that the early diagnosis rate of PAAD is extremely low; and most of the patients have shown local infiltration or even distant metastasis at the time of diagnosis [3]. Less than 20% of PAAD patients are fortunate to undergo surgical resection, yet most of these cases will eventually relapse and metastasize [4]. The high morbidity and mortality of PAAD are most attributed to the two clinical dilemmas, one being the lack of effective early diagnostic methods, and the other the lack of effective intervention. Therefore, uncovering the pathogenesis of PAAD and identifying effective prognostic markers and precise molecular targets constitute the core tasks for the treatment of pancreatic cancer.

The pathogenesis of PAAD is a long-term and complicated process that manifests imbalanced biological regulation. Thusly, screening and identifying the precise prognostic factors is of great necessity to help predict the prognosis of pancreatic cancer and devise individualized treatment of patients. Currently, the American Joint Committee on Cancer (AJCC) Cancer Staging is still the most widely used predictive model for pancreatic cancer. The staging system aims to provide guidance for prognostic assessment

and treatment decisions [5]. Nevertheless, the AJCC staging system merely evaluates the basic indicators of pathology, including T staging, N staging, and M staging. PAAD patients of the same AJCC staging may exhibit various clinical prognosis after receiving the same treatment [6], and the current prediction system is far from sufficient to predict the prognosis of PAAD patients accurately, therefore further exploration is necessary.

Recently, most of the reaches on cancer focus on the gene or protein level to explore the biological effects of specific genes or products on the prognosis of pancreatic cancer, thereby identifying numerous potential biomarkers that are implicated in the prognosis of pancreatic cancer. For instance, carbohydrate antigen 199 (CA-199) is the most commonly used tumor marker in clinical PDAC screening, yet its application is restricted in the early diagnosis and treatment of pancreatic cancer due to low sensitivity and specificity [7; 8]. Previous studies have reported that the protein expression levels of CDKN2A / p16, TP53 and SMAD4 / DPC4 could be referred to predict the postoperative survival of PAAD patients [9]. In addition, the family of non-coding RNA has also shown huge potential for predicting prognosis of cancerous patients [10; 11]. Nonetheless, studies on the effects of these indicators on pancreatic cancer mostly focus on the function of a single or several genes, making it difficult to conduct in-depth and systematic research from a dynamic, holistic, and multidimensional perspective. With the continuous improvement of high-throughput sequencing technology, integrated multi-level approach of genome and transcriptome for systematic research [12; 13; 14] provide novel ideas to explore the prognostic markers of pancreatic cancer.

In the current study, we established a 9-gene signature model based on the integrating and screening of genomic and transcriptomic data, and subsequently validated its validity by using internal and external cohorts. In addition, we found that the 9-gene signature is involved in some important biological pathways of pancreatic cancer. In a nutshell, the 9-gene signature established in this study can effectively predict the prognosis of patients with pancreatic cancer and show promising application in clinical practice.

Methods

Data source and preprocessing

TCGA cohort: Normalized RNA-sequencing data as transcripts per million (TPM) and the associated clinical information of the PAAD samples were downloaded from The Cancer Genome Atlas (TCGA) cohort (https://portal.gdc.cancer.gov/.

171 cases with corresponding tumor tissues and clinical information were included in the study. Normalized gene expression data for the TCGA PAAD cohort were log2-transformed for further analysis.

GEO cohort：Gene expression microarray data of 132 normal and primary pancreatic tumor samples from patients were obtained from the Gene Expression Omnibus (GEO) database (accession number

GSE21501). Normalization, quality control, and imputation of array data were performed. Expression data from multiple probes were collapsed by the mean for each sample

Distribution of Clinical information

First, 80% of the samples from the 171 TCGA samples after preprocessing are randomly selected as the training cohort. In order to avoid random allocation bias affecting the stability of subsequent modeling, all samples are pre-repetitively sampled 100 times with replacement, To ensure that the randomly selected samples and all samples are distributed uniformly in clinical characteristics. The TCGA training cohort contains 137 samples, GSE21501 contains 97 samples, and ICGC contains 257 samples. The distribution of clinical information, including age, survival status, gender, and TNM staging, of the three cohorts is elaborated in **Table 1**.

Table 1. Clinical information statistics of three sets of data sets

Characteristic		TCGA training datasets(n=137)	TCGA entire datasets(n=171)	GSE21501(n=97)	ICGC datasets (n=257)
Age(years)	≤65	72	90	-	112
	>65	65	81	-	144
Survival Status	Living	65	80	32	106
	Dead	72	91	65	151
Gender	female	68	78	-	120
	male	69	93	-	137
FAMILY HISTORY	NO	38	45	-	-
	YES	51	62	-	-
SMOKING HISTORY	NO	54	63	-	-
	YES	57	76	-	-
AJCC TUMOR PATHOLOGIC PT	T1	6	7	2	-
	T2	17	21	16	-
	T3	109	138	74	-
	T4	3	3	1	-
AJCC NODES PATHOLOGIC PN	N0	39	47	27	-
	N1/NX	97	119	69	-
AJCC METASTASIS PATHOLOGIC PM	M0	59	77	-	-
	M1/MX	78	94	-	-
AJCC PATHOLOGIC TUMOR STAGE	Stage Ⅰ	15	19	-	-
	Stage Ⅱ	112	142	-	-
	Stage Ⅲ	3	3	-	-
	Stage Ⅳ	4	4	-	-
Grade	G1	25	32	-	-
	G2	72	97	-	-
	G3	39	51	-	-
	G4	1	2	-	-

Prognostic genes obtained after preliminary analysis of multiple omics data

As for the samples of TCGA training cohort, univariate regression analysis was applied to select candidate prognostic genes with p values < 0.01. The most significant top 20 genes among the candidate

prognostic genes were shown in **Table 2**

Table2. Information of top 20 Prognostic Related Gene

Gene	p value	HR	Low 95%CI	High 95%CI
ANLN	2.69E-11	1.051	1.036	1.067
CKAP2L	5.16E-11	1.336	1.225	1.456
ARNTL2	1.73E-10	1.050	1.035	1.066
KIF20A	5.66E-10	1.119	1.080	1.159
MET	7.98E-10	1.010	1.007	1.013
SMCO2	2.65E-09	1.989	1.586	2.494
SLC35F2	4.63E-09	1.027	1.018	1.037
KIF23	5.47E-09	1.186	1.120	1.256
CDCA5	7.63E-09	1.095	1.062	1.129
CEP55	9.44E-09	1.063	1.041	1.086
HJURP	1.49E-08	1.107	1.069	1.147
NT5E	1.68E-08	1.011	1.007	1.015
RARRES3	2.86E-08	1.004	1.003	1.006
TGFBI	2.97E-08	1.004	1.002	1.005
DIAPH3	3.10E-08	1.325	1.199	1.464
BUB1	3.47E-08	1.168	1.106	1.235
DLGAP5	4.15E-08	1.091	1.058	1.126
NUSAP1	4.81E-08	1.044	1.028	1.061
FAM196B	7.29E-08	2.812	1.930	4.097
CENPA	7.50E-08	1.177	1.109	1.249

Construction of Univariate Cox Proportional Hazards Model

Similarly, Guo J, et al[15], applied univariate cox proportional hazards regression analysis on each gene to screen out those that were significantly related to patient's OS in the training cohort, and $p < 0.01$ was selected as the significance level.

Data analysis of copy number variation

GISTIC is widely used and detects both broad and focal (potentially overlapping) recurring events. We used GISTIC 2.0[16] software to identify genes exhibiting significant amplification or deletion. The parameter threshold was defined as the length of amplification or deletion being greater than 0.1 and $p < 0.05$ in the fragments.

Genetic Mutation analysis

To identify significantly mutated genes, we used Mutsig 2.0 software to analyze the maf files of TCGA mutation data, at the significance level of $p < 0.05$.

Construction of prognostic signatures

We selected the genes that are significantly related to patient's OS and have undergone notable amplification, deletion, and mutation. The genes were ranked according to their importance using

RandomSurvivalForest Package in R [17] following the methods described by Meng, J et al. [18]. The number of iterations for Monte Carlo simulation was set to 100, the number of advancing step was 5, and genes with a relative importance of greater than 0.4 were identified as characteristic genes. Furthermore, multivariate Cox regression analysis was performed to construct the following risk scoring model:

$$RiskScore = \sum Expk * eHRk \; n \; k=1$$

Where N is the number of prognostic genes, $Expk$ is the expression value of prognostic genes, and $eHRk$ is the estimated regression coefficient of genes in the multivariate Cox regression analysis.

Functional enrichment analyses

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed using the *clusterprofiler* package in R[19] to identify over-represented GO terms in three categories (biological processes, molecular function and cellular component), and KEGG pathway. For this analyses, a FDR<0.05 was considered to denote statistical significance.

Construction of risk model for the training cohort

Lasso method is a kind of compressed estimation used to obtain a refined model by constructing a penalty function [20]. Meanwhile, it can compress some coefficients and set some coefficients to zero. Therefore, it retains the advantage of subset selection with shrinkage, and is a kind of biased estimation suitable for the processing of complex collinear data. It can help with the selection of variables at the time of parameter estimation, so as to better solve the multicollinearity problem in regression analysis. We use the *glmnet* package in R for lasso cox regression analysis to analyze the trajectory of each independent variable.

Gene Set Enrichment Analysis (GSEA)

GSEA was performed by the JAVA program (http://software.broadinstitute.org/gsea/downloads.jsp) using the MSigDB14 C2 Canonical pathways gene set collection, which contains 1320 gene sets. Gene sets with a false discovery rate (FDR) value less than 0.05 after performing 1000 permutations were considered to be significantly enriched.

External validation of proteins and transcription levels of 9 gene signature

Human Protein Atlas (HPA) provides information on the tissue and cell distribution of 26,000 human proteins. It mainly uses specific antibodies to study the protein expression in cell lines, normal tissues and tumor tissues. Explore the expression of 9 genes (UNC13B, TSPYL4, MICAL1, KLHDC7B, KLHL32, AIM1, ARHGAP18, DCBLD1, and CACNA2D4) in normal and tumor tissues; Explore the expression of 9 genes in pancreatic cancer and normal tissues in GSE62452, GSE107610, and TCGA-PAAD, and draw boxplots on the expression of genes;

Genetic alterations of the 9 predictive genes

cBioportal integrates genomic data including somatic mutations, DNA copy-number, alternations (CNAs), mRNA and microRNA (miRNA) expression, DNA methylation, protein enrichment, and phosphorylated protein enrichment; Pancreas ((UTSW), Pancreas (TCGA PanCan 2018), Pancreas (QCMG 2016), PAAC (JHU), Pancreas (ICGC), Pancreas (TCGA), Pancreas (JHU 2011), ACRNET (Nature, 2017), PANET (Johns Hopkins 2011) , Panet (Shanghai 2013) cohort, analysis of mutation correlation of 9 genes (OncoPrint and bar graph display of gene mutation).

Sample Collection

PAAD and adjacent tissues were collected from 4 patients, immediately placed in liquid nitrogen, and preserved at -80°C. Patients and their families in this study have been fully informed and informed consent was obtained from the participants. This study was approved by the Ethics Committee of Hainan General Hospital.

Western blot

The 4 pairs of PAAD and adjacent tissues were separately dissected and pooled to obtain equalized protein content (30 μg). Following the standard Western blot technique, the proteins were respectively electrophoresed through a 10% SDS-PAGE and transferred to immobilon polyvinylidene difluoride membrane (Millipore, Billerica, Massachusetts, USA). The membranes were blocked for 1 h at room temperature in 5% non-fat milk/TBS (10 mM Tris-HCl (pH8.0), 150 mM NaCl, 0.05% tween 20). Proteins were probed with primary antibodies raised against MICAL1, UNC13B, KLHDC7B, KLHL32 (Cell Signaling Technology, MA, USA), ARHGAP18, CACNA2D4, TSPYL4, AIM1, DCBLD1 (Proteintech, China) overnight at 4°C and incubated with GAPDH antibody (1:2000; Sigma, USA) as the internal control. After rinsing three times (10 min each) with TBS, the membrane was incubated with horseradish peroxidase conjugated secondary antibody against rabbit IgG (1:5000, Amersham Bioscience, Piscataway, NJ) for 1 h at room temperature. After washout, the membrane was developed using ECL reagents (Pierce, Rockford, Illinois, USA) and visualized using a chemiluminescence system (PTC-200; Bio-RAD Laboratories, Hercules, California, USA). All western blots were repeated three times.

RNA extraction and real-time polymerase chain reaction(PCR) assay

Total RNA was extracted using TRIzol Reagent (Invitrogen, CA, USA) following the manufacturer's protocol and was reverse- transcribed into complementary DNA (cDNA) using a Superscript Reverse Transcriptase Kit (Transgene, France). Super SYBR Green Kit (Transgen, France) was used to carry out real-time PCR in ABI7300 real-time PCR system (Applied Biosystems). The primers pairs were

ARHGAP18 Forward, 5'-ATCAAGAGGTGGTTGTTGTCAAA-3', ARHGAP18 Reverse, 5'-ACAATGCTTTCCTGTGGATCTC-3', CACNA2D4 Forward, 5'-TGCCTGCAACTCCCAACTTC-3', CACNA2D4 Reverse, 5'-CCAGAGGAATCTTGGCCTGTC-3', DCBLD1 Forward, 5'-ATCACACTGTTTGCGAAAAGACA-3', DCBLD1 Reverse , 5'-GACGGTTACTTCACTTGTGTTCA-3',

KLHDC7B Forward, 5’- GCACCATGCACAACTACCTGT-3’, KLHDC7B Reverse 5’-ATTCGCCACCGATGGCATAG-3’, KLHL32 Forward ,5’-GAACGCTGCCTCAGTATTCAA-3’, KLHL32 Reverse , 5’-AGGGTGATGTCGCAGAGGA-3’, MICAL1 Forward, 5’-GGCACTCGGTGCTAAGAAGTT-3’, MICAL1 Reverse , 5’-CCCCAGTGAATTTCCACCCC-3’, TSPYL4 Forward, 5’-TCTCAGATGATACCGGGGAAG-3’, TSPYL4 Reverse ,5’-GGGCATTTACGTTTGACAACTC-3’, UNC13B Forward, 5’-CTCTGCGTGCGCGTTAAAAG-3’, UNC13B Reverse, 5’-CAGGCGACTAATCTCAAACATGA-3’. The mRNA expression level was normalized to GAPDH and replicated in triplicate according to the manufacturer ’s instructions. The relative expression quantity was calculated using the 2-ΔΔCt method.

Statistics

The median risk score in each cohort was used as cutoff to compare the survival risk between high-risk group and the low-risk group, and subsequently plot the Kaplan-Meier (KM) survival curve. A multivariate Cox regression analysis was performed to test whether the genetic markers were independent prognostic factors. Significance level was set at P <0.05. All these analyses were performed using R 3.4.3.

Results

Analysis of copy number variation

For the data of copy number variation in TCGA, we used GISTIC 2.0 to identify genes with significant amplification or deletion, and the parameter threshold was defined as the fragment with an amplification or deletion length greater than 0.1 at the level of p <0.05. The significantly amplified fragments in the pancreatic cancer genome were listed in **Supplement Table 1**. For example, significantly amplified CCAT1 at 8q24.21 (q value = 7.69E-14), GATA6 at 18q11.2 (q value = 7.69E-14), and RPS16 at 19p13.2 (q value = 3.83E-07) were determined. Gene amplification was found for 82 genes in total.

The significantly deleted fragments in the pancreatic cancer genome were shown in **Supplement Table 2**, such as CDKN2A at 9p21.3 (q value = 1.83E-74), SMAD4 at 18q21.2 (q value = 1.58E-41), and SFN at 1p36.11 (q value = 0.00011). After removing redundancy, a total of 476 deleted genes were deleted. The distribution of copy number variation was shown in **Figure 1A.**

Furthermore, heatmap was plotted to illustrate the top 15 copy number variations (**Figure 1B**). Different copy number variation between different stages and grade samples was determined, yet none reached the significant level.

Analysis of mutation data

We used Mutsig2 to identify the genes with significant mutations. The threshold significance level was set at p <0.05, and a total number of 135 significantly mutated genes were selected. The pronounced mutated 49 genes (p <0.02) in PAAD patients and their distribution were presented in **Figure 2**. The left image shows the total number of synonymous and non-synonymous mutations in each patient, and the right shows the number of samples with mutations of genes. Of the identified significant genes, some

have been reported in previous reach that are closely related to the carcinogenesis and progression, such as KRAS, TP53, CDKN2A, and RNF43.

Pathways and biological processes that involve genes with copy number variation and mutation

By analyzing the amplified and deleted genes which were identified by copy number variant in TCGA, as well as targeted integration of mutant genes, we identified 693 genes in total that are implicated in the corresponding biological processes and pathways. These genes were significantly enriched in 39 KEGG pathways, including MAPK signaling pathway, Hepatocellular carcinoma, TNF signaling pathway and other tumor-related pathways. The Top 20 closed associated KEGG pathways were shown in **Figure 3A**. Furthermore, the gene sets were integrated into a network diagram (**Figure 3B**) based on enrichment condition and overlapped gene sets. The overlapping gene sets tend to cluster, thusly helping with the finding of clusters with similar biological functions.

Construction of risk model based on training cohort

Univariate cox analysis was applied to identify 2625 prognostic genes with p values < 0.01 in training cohort. Through the previous analysis, we have obtained 82 copy amplification, 476 copy deletions and 135 significant mutations genes. After intersection with 2625 prognostic related genes, 54 candidate genes based on multi-omics were obtained.

Subsequently, we need to narrow down the 54 genes range and build a prognostic model while maintaining high accuracy. We used *glmnet* package to perform lasso cox regression analysis. As seen in **Figure 4A**, the gradually decreasing lambda resulted in a higher number of independent variable with coefficients approaching 0. We used 10-fold cross-validation to build the model and to analyze the confidence interval under each lambda. As shown in **Figure 4B**, the model is optimal when lambda equals to 0.03874. Thusly, we selected 15 genes as the target genes under the condition of lambda equals to 0.03874.

Furthermore, we conducted multi-variant cox survival analysis on the 15 genes obtained in the previous step, and retained 9 mRNAs with the lowest AIC value (AIC = 555.89) as the final model. The detailed information of these 9 mRNAs was shown in Table 3.

Table 3: 9- genes significantly associated with the overall survival in the training-set patients

Symbol	coef	HR	Z-score	P value	Low 95%CI	High 95%CI
UNC13B	-0.493	0.6106	-1.945	0.0517	0.3715	1.0037
CACNA2D4	0.510	1.6647	2.053	0.0401	1.0233	2.7082
KLHDC7B	0.131	1.1403	1.747	0.0807	0.9841	1.3212
AIM1	0.627	1.8711	3.072	0.0021	1.2546	2.7905
TSPYL4	-1.133	0.322	-3.691	0.0002	0.1764	0.5877
MICAL1	-1.068	0.3436	-4.41	1.04E-05	0.2137	0.5524
ARHGAP18	0.625	1.8684	2.156	0.0311	1.0585	3.2979
KLHL32	-0.440	0.644	-1.56	0.1187	0.3705	1.1194
DCBLD1	0.606	1.8325	2.345	0.0190	1.1045	3.0403

The prognostic KM curve of the 9 genes is shown in **Fig S1**. Obviously, 7 genes can significantly distinguish the samples of TCGA training cohort into high-risk and low-risk group ($p < 0.05$). One shows marginal significance at the level of $p = 0.062$, whereas the other indicates no significance. Among them, the down-regulated expression of four genes, including UNC13B, TSPYL4, MICAL1 and KLHL32, along with the high-regulated expression of five genes, including KLHDC7B, AIM1, ARHGAP18, DCBLD1 and CACNA2D4, are associated with poor prognosis. Although no significant correlation was detected in CACNA2D4, yet its aberrantly higher expression was determined in high-risk groups, which indicates dismal prognosis than low-risk group. The final version of the 9-mRNA signature formula is as follows:

$$RiskScore_9 = -0.4933*exp^{UNC13B} + 0.5097*exp^{CACNA2D4} + 0.1313*exp^{KLHDC7B}$$

$$+0.6265*exp^{AIM1} - 1.1333*exp^{TSPYL4} - 1.0684*exp^{MICAL1} + 0.6251*exp^{ARHGAP18}$$

$$-0.4401*exp^{KLHL32} + 0.6057*exp^{DCBLD1}$$

ROC and Kaplan-Meier analysis of risk model

Furthermore, we performed ROC analysis on RiskScore for prognostic classification by using *timeROC* package. We analyzed the efficiency of prognostic classification efficiency of the 1st, 3rd and 5th year in both training cohort and testing cohort. As shown in **Figure 5A-B**, the AUC for the 5th years is 0.93 in the training cohort and 0.9 in the testing cohort. We conducted z-score transformation of the Riskscore, dividing the samples with Riskscore > 0 into high-risk group, and those with Riskscore < 0 into lower-risk samples, plotted a KM survival curve. As shown in **Figure 5C-D**, significant differences were determined in the training cohort (log rank $p < 0.0001$, HR = 4.534) and testing cohort (log rank $p < 0.0001$, HR = 3). 86 samples were divided into high-risk group and 51 samples into low-risk group.

Robustness of 9-gene signature in the TCGA -PAAD cohort

In order to ensure the robustness of the model, we used the same model in the TCGA training cohort and the same cutoff value, and verified it in the TCGA-PAAD cohort, and plotted the RiskScore distribution. As shown in **Figure 6A**, the OS of high-risk groups was significantly smaller than that of low-risk groups, which indicates that samples with high RiskScore have worse prognosis. The changed expression of 9 different signature genes with higher risk value was identified. We verified that the high expression of KLHDC7B, AIM1, ARHGAP18, DCBLD1 and CACNA2D4 indicate high risk, demonstrating these 5 genes as risk factors. Meanwhile, it was found that the high expression of UNC13B, TSPYL4, MICAL1 and KLHL32 indicates low risk, demonstrating these 4 genes as protective factors.

Furthermore, we used *timeROC* package in R to perform ROC analysis on the RiskScore to carry out prognostic classification. The efficiency of the prognostic classification in the 1st, 3rd and 5th year was analyzed. As shown in **Figure 6B**, the AUC is high in the model, reaching above 0.90 in the 5th year. Finally, we conducted z-score transformation of the Riskscore, dividing samples with Riskscore greater than zero into high-risk groups, and the other into low-risk group, and plotted a KM curve. As shown in **Figure 6C**,

extremely significant differences could be found, with log rank p <0.0001 and HR = 3.01 (1.819-4.981). 107 samples were divided into high-risk groups and 64 samples were into low-risk groups.

External Verification of the robustness of 9-gene signature

We applied the same model and coefficients of the training cohort in the two external validation cohorts. We also calculated the risk score of each sample based on the expression level of the sample and plotted the RiskScore distribution of the sample.

The results of the GSE21501 cohort are shown in **Figure 7**. We used the *timeROC* package to perform ROC analysis on the RiskScore for prognostic classification. We analyzed the prognosis efficiency of prognostic classification of the 1st and the 3rd year, and found high value of AUC in the model. The AUC of the 1st year is 0.73. We draw a survival curve based on the expression of RiskScore. As shown in **Figure 7B**, extremely significant differences was observed, with log rank p = 0.05, HR = 1.628 (0.997-2.657). Among them, 48 samples were divided into high-risk group and 49 samples were into low-risk group.

Similarly, for ICGC cohort. The prognostic classification efficiency of 1^{st,} and 3rd year was investigated, respectively. As shown in **Figure 7C**, the model has a high AUC (the AUC for 1st year is 0.75). As shown in **Figure 7D**, significant marginal difference was determined, with log rank p = 0.011 and HR = 1.526 (1.098-2.122). 136 samples were divided into high-risk group and 121 sample were low-risk group.

Risk Model and Prognostic Analysis of Clinical Features

Furthermore, the results of survival analysis determined that only Grade and N stage in the TCGA training cohort samples were significantly related to OS of pancreatic cancer (**Figure 8**), whereas significant difference were not observed in regard to age, gender, family history, T stage, M stage and smoking history.

After performing a RS analysis on the 9-gene markers, we found that the 9-mRNA signatures can effectively separate young, old, male, female, smoking, family history, Grade□+□, T3, Stage □, N0 and N1 stage patients into high risk and low risk groups (p <0.01). It further demonstrated the good predictive ability of this model in regard to different clinical features (**Figure 9**).

Univariate and multivariate analysis of 9-gene signature

To identify the independence of the 9-gene signature model in clinical application, we conducted Univariate and multivariate COX regression analysis into the relevant clinical information carried by the entire TCGA, GSE21501 and ICGC, and calculated 95% CI of HR and p value. We systematically analyzed the clinical information from TCGA, GSE21501, and ICGC (International Cancer Genome Consortium), including age, gender, and grouping information of 9-gene signature. Uni-variant COX regression analysis found that RS, Age, T, N, and Grade in the TCGA cohort were significantly related to survival, yet the corresponding multi-variant COX regression analysis found that RS (HR = 2.43 95% CI = 1.420-4.165, p = 0.001) and age are significantly associated with prognosis and exhibit clinical independence.

As for the external cohort of GSE21501, univariate COX regression analysis found that RS and N were significantly related to survival, yet the corresponding multi-variant COX regression analysis only found N (HR = 2.01, 95% CI = 1.066-3.778, p = 0.031) and Age are significantly associated with prognosis and exhibit clinically independence.

Lastly, as for ICGC, Univariate and multivariate COX regression analysis found that RS was significantly associated with survival. The above findings indicated that our 9-gene signature is a prognostic model independent from other clinical factors and thereby presenting independent predictive capacity in clinical application.

Riskscore-related regulatory pathways

To observe the association between the riskscore and biological function in different samples, we selected the gene expression profiles corresponding to these samples and performed single-sample GSEA analysis using *GSVA* package in R. We calculated the scores of different functions in different samples to obtain the ssGSEA score for each sample. Furthermore, we calculated the correlation between these functions and Riskscore, and then selected the function with a correlation coefficient greater than 0.5. As shown in **Figure 10A**, a small part is negatively correlated with the sample's risk score, whereas most of the part shows a positive correlation with the Riskscore.

We selected the 25 KEGG Pathways that have coefficients greater than 0.5, and then performed cluster analysis based on the enrichment scores (**Figure 10B**). It can be seen that among these 25 pathways, P53 SIGNALING PATHWAY, CELL CYCLE and other tumor progression-related pathways increase with higher RiskScore, whereas the TASTE TRANSDUCTION and CARDIAC MUSCLE CONTRACTION pathways decrease with higher risk score. This also suggests that the dysregulation of these pathways is closely related to tumor development.

Comparison of risk model with other models

After referring to previously published literature, we selected four prognostic risk models, including 15-gene signature (Chen) [21], 7-gene signature (Cheng) [22], 5-gene signature (Raman)[23], and 9-gene signature (Wu) [24] for the comparison with our 9-genes model. To ensure comparability, we calculated the risk score of each PAAD sample in the TCGA using the same method based on the corresponding genes in the 4 models, and evaluated the ROC of each model, and divided the samples into Risk-H and Risk-L groups in accordance with the median risk score. The difference in prognosis between the two groups of samples was calculated.

The ROC and KM curves of the four models are shown in **Fig. 10 A-H**. Apparently, the AUC of the four gene models are all above 0.70, yet the predictive effects of the four models are inferior to our 9-gene model. Significant difference in the prognosis was also determined between Risk-H and Risk-L groups of samples in the four models (log rank p <0.001).

Furthermore, to compare the prediction performance of these models on PAAD samples, we used "*rms*" package in R to calculate the concordance index (C-index) of different models. It can be seen that the C-index of the 9-genes model is the highest (**Fig.10l**), indicating that the overall performance of the 9-gene model outweighs than the other four models.

External validation of proteins and mRNA of 9 gene signature

The protein expressions of 9 genes were analyzed by HPA database. Among them, MICAL1 and UNC13B are negative in tumors and normal tissues, KLHDC7B, KLHL32 are not significantly different in tumors and normal controls. ARHGAP18, CACNA2D4, and TSPYL4 are relatively low expressed in tumors, and AIM1, DCBLD1 are relatively high expressed in tumor tissues (**Figure 11**).

The differential expression of 9 genes in GSE62452, GSE107610 and TCGA-PAAD were analyzed. Among them, AIM1 and DCBLD1 were highly expressed in tumors, while KLHDC7B, ARHGAP18, CACNA2D4, TSPYL4 and KLHL32 were all low expressed in tumors, and the overall expression trends of 9 genes in the three sets of pancreatic cancer cohorts were roughly the same (**Figure 12**).

Genetic alterations of the 9 predictive genes

The mutations of 9 genes was explored in cBioportal database. among which the gene with the highest mutation proportion was UNC13B, accounting for 3%, and the main type of mutation was amplification, followed by KLHDC7B (2.3%) and CACNA2D4 (2.5%). (**Figure 13**)

Clinical validation of 9 genes

Then, we measured the expression levels of UNC13B, TSPYL4, MICAL1, KLHDC7B, AIM1, ARHGAP18, DCBLD1, and CACNA2D4 in four pairs of PADD tumors and normal samples. We found that compared with normal tissues, M1CAL1, KLHDC7B, KLHL32, and UNC13B were not different in human PADD tissues and normal controls. ARHGAP18, CACNA2D4, and TSPYL4 were relatively low expressed in PADD tissues. DCBLD1, A1M1 are relatively highly expressed in tumor tissues, and the experimental results are almost consistent with our data analysis (**Figure 14**)

Table 4. Identification of clinical factors and independence that are related to prognosis using uni- and multi-variant COX regression analysis

Variables	Univariate analysis			Multivariable analysis		
	HR	95%CI of HR	P value	HR	95%CI of HR	P value
Entire TCGA cohort						
9-mRNA risk score						
Risk score(High/Low)	3.01	1.82-4.98	1.79E-05	2.43	1.420-4.165	0.001
Age	1.03	1.01-1.05	6.50E-03	1.02	1.002-1.044	0.033
Gender (Male/Female)	0.8	0.53-1.21	0.294	0.95	0.614-1.458	0.802
FAMILY HISTORY	1.12	0.65-1.92	0.688	0.98	0.435-1.506	0.505
SMOKING HISTORY	1.12	0.71-1.78	0.629	0.81	0.580-1.857	0.901
AJCC PT	2.02	1.07-3.82	0.03	1.30	0.691-2.454	0.414
AJCC PN	2.01	1.20-3.37	0.008	1.46	0.831-2.557	0.189
AJCC PM	0.8	0.53-1.22	0.306	1.00	0.651-1.536	1.000
AJCC STAGE	0.81	0.26-2.57	0.719	1.48	0.662-3.29	0.341
Grade	1.55	1.00-2.40	0.048	1.19	0.782-1.812	0.417
GSE21501 cohort						
9-mRNA risk score						
Risk score(High/Low)	1.63	1.00-2.66	0.051	1.41	0.845-2.347	0.189
T stage	0.93	0.50-1.72	0.817	0.74	0.392-1.407	0.362
N stage	2	1.12-3.56	0.019	2.01	1.066-3.778	0.031
ICGC **cohort**						
9-mRNA risk score						
Risk score(High/Low)	1.53	1.10-2.12	0.012	1.514	1.087-2.110	0.014
Age	1.02	1.00-1.04	0.052	1.017	1.000-1.035	0.056
Gender (Male/Female)	1.25	0.91-1.74	0.172	1.326	0.955-1.840	0.092

Discussion

Human malignant cancer indicates an intricate etiology that involves multiple factors and accumulated lesions after multiple-stage. During such process, genetic factors, epigenetic factors, as well as the numerous related genes under their regulation jointly constitute the complex heterogeneity of tumors, bringing major difficulties to clinical diagnosis and personalized treatment [25; 26; 27]. The continuous development of high-throughput sequencing technology makes it possible to carry out comprehensive and multi-level researches on tumors at genome level and transcriptome level [26]. Meanwhile, integrating multiple omics data and conducting comprehensive research in combination with patients' clinical data prove superior in ascertaining effective therapeutic targets and prognostic markers [28; 29]. In the present study, we analyzed multi-omics data, including transcriptome, copy number variation, and mutation, to screen and construct a 9-gene signature risk prognosis model that is closely related to the prognosis of PAAD patients. The 9-gene signature has strong robustness, and shows stable and consistent prediction performance in both the TCGA internal validation set and the GEO external validation set.

More importantly, we systematically analyzed the clinical information in TCGA, GSE21501, and ICGC cohorts by performing COX regression analysis, and found that the constructed 9-gene signature model maintains consistent and stable clinical independence under the influence of various clinical factors. All

nine genes were demonstrated as independent prognostic factors, further evincing the high stability of the 9-Gene signature. In light of the above results, suffice it to way that the constructed 9-gene signature has huge potential for clinical application by virtue of its stable and consistent predictive power for prognosis on different platforms and in various cohorts.

A plethora of studies have confirmed that KLHDC7B is overexpressed in breast cancer and is involved in regulating the malignant progression of breast cancer [30; 31]. In addition, KLHDC7B has been shown to be closely related to the prognosis of laryngeal cancer [32]. As an actin-binding protein, AIM1is aberrantly expressed in melanoma [33; 34], prostate cancer [35; 36], and bladder cancer [37] and is closely related to prognosis. ARHGAP18, a RhoGTP activating protein, has been shown to correlate with the prognosis of patients with gastric cancer [38] and breast cancer [39], due to its participation in regulating the metastatic capacity of breast cancer cells [40; 41]. TSPYL4 has been confirmed to correlate with the prognosis of patients with head and neck squamous cell carcinoma [42]. KLHL32 is implicated in regulating the production of ROS and plays a vital role in the progression of breast cancer [43; 44] and melanoma [45]. In addition, DCBLD1, CACNA2D4, and UNC13B have not been reported to be associated with the prognosis of cancerous patients; they were determined, for the first time, as prognostic markers of pancreatic cancer in this study.

Finally, the results of GSEA enrichment analysis showed that 9-gene signature is closely related to tumor pathway such as P53 SIGNALING PATHWAY□CELL CYCLE in pancreatic cancer.

It has been corroborated in numerous studies that P53 pathway and aberrant cell cycle regulation exert considerable impact on the biological behavior of pancreatic cancer cells and the prognosis of PAAD patients [45; 46; 47]. These findings are consistent with our results. Therefore, pathway enrichment analysis should further conducted to deepen our understanding about the intricate mechanism of the onset and progression of PAAD.

Researchers have attempted to screen and construct prognostic marker models for pancreatic cancer. Chen et al. constructed a 15-gene signature to predict the prognosis of patients with early pancreatic ductal cancer [21]. Cheng et al. constructed a 7-gene signature for the prognosis of pancreatic cancer based on multiple GEO cohorts [22]. In addition, the 5-gene signature was constructed by by Raman et al. on the basis of GEO and ICGC cohorts [23]. Wu et al., conducted a 9-gene signatures based on TCGA and GEO cohorts [24]. In order to further validate the advantages of our 9-gene signature, we analyzed the above four models simultaneously. The results show that the 9-gene signature constructed in this study is superior to the other 4 models in regard to predicting the prognosis of patients. Further C-index analysis showed that the overall performance of our model is improved than the other four. In view of the above results, our prediction model based on multi-omics data shows stronger advantages. It can help with the prediction of the individual risk of PAAD patients, and provide constructive guidance for patient evaluation and decisions of clinical intervention.

Then we integrated the HPA protein database, GSE62452, GSE107610 and TCGA-PAAD databases to validate the protein and mRNA levels of 9 genes. We found that AIM1 and DCBLD1 were highly expressed

in tumor tissues, while ARHGAP18, CACNA2D4, and TSPYL4 were lowly expressed in tumor tissues. Then we used 4 pairs of PADD tissue and control samples for protein and transcription level analysis, the results were consistent with the analysis

Although we conducted the present experiment to screen and verify the potential prognostic markers for PAAD on the basis of large sample multi-omics data, it is not without limitations. The further verification in vivo and in vitro of these genes is still needed. In addition, the samples used in the study are retrospectively analyzed, more thoroughly and comprehensively investigations are yet to be done prior to clinical applications.

Conclusion

In summary, we screened and validated a prognostic risk model consisting of 9 genes which exhibited excellent predictive power in both training and validation sets, and all 9 genes are independent prognostic factors. The construction of this multi-omics risk model can improve the ability to predict the prognostic risk of PAAD patients more accurately. Therefore, we recommend that this 9-gene multi-omics model used as a molecular marker to assess the prognostic risk of PAAD patients.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The data used to support the findings of this study are available from the corresponding author on reasonable request.

Competing interests The authors declare that there are no conflicts of interest

Funding

This work is supported by the Natural Science Foundation of Hainan province (20158274).

Authors' contributions

XDF, WY, LXM and ZJF were responsible for the study concept and design. XDF, WY, LXM, ZKL, WJC, CJC, CC and CL were responsible for data acquisition. XDF, WY, LXM ZKL, WJC, CJC CC and CL performed the experiment. XDF, WY and LXM performed data analysis and manuscript writing. ZJF finalized the manuscript. XDF, WY and LXM contributed equally.

Acknowledgements

Not applicable.

References

[1] R.L. Siegel, K.D. Miller, and A. Jemal, Cancer statistics, 2019. CA Cancer J Clin 69 (2019) 7-34.

[2] A. Mohammed, N.B. Janakiram, V. Madka, M. Li, A.S. Asch, and C.V. Rao, Current Challenges and Opportunities for Chemoprevention of Pancreatic Cancer. Curr Med Chem 25 (2018) 2535-2544.

[3] L.C. Chu, M.G. Goggins, and E.K. Fishman, Diagnosis and Detection of Pancreatic Cancer. Cancer J 23 (2017) 333-342.

[4] M. Ilic, and I. Ilic, Epidemiology of pancreatic cancer. World J Gastroenterol 22 (2016) 9694-9705.

[5] B.L. Appel, P. Tolat, D.B. Evans, and S. Tsai, Current staging systems for pancreatic cancer. Cancer J 18 (2012) 539-49.

[6] N.A. Juiz, J. Iovanna, and N. Dusetti, Pancreatic Cancer Heterogeneity Can Be Explained Beyond the Genome. Front Oncol 9 (2019) 246.

[7] M.A. Tempero, E. Uchida, H. Takasaki, D.A. Burnett, Z. Steplewski, and P.M. Pour, Relationship of carbohydrate antigen 19-9 and Lewis antigens in pancreatic cancer. Cancer Res 47 (1987) 5501-3.

[8] M.S. Kim, S.V. Kuppireddy, S. Sakamuri, M. Singal, D. Getnet, H.C. Harsha, R. Goel, L. Balakrishnan, H.K. Jacob, M.K. Kashyap, S.G. Tankala, A. Maitra, C.A. Iacobuzio-Donahue, E. Jaffee, M.G. Goggins, V.E. Velculescu, R.H. Hruban, and A. Pandey, Rapid characterization of candidate biomarkers for pancreatic cancer using cell microarrays (CMAs). J Proteome Res 11 (2012) 5556-63.

[9] M. Oshima, K. Okano, S. Muraki, R. Haba, T. Maeba, Y. Suzuki, and S. Yachida, Immunohistochemically detected expression of 3 major genes (CDKN2A/p16, TP53, and SMAD4/DPC4) strongly predicts survival in patients with resectable pancreatic cancer. Ann Surg 258 (2013) 336-46.

[10] B. Baradaran, R. Shahbazi, and M. Khordadmehr, Dysregulation of key microRNAs in pancreatic cancer development. Biomed Pharmacother 109 (2019) 1008-1015.

[11] A.A. Tesfaye, A.S. Azmi, and P.A. Philip, miRNA and Gene Expression in Pancreatic Ductal Adenocarcinoma. Am J Pathol 189 (2019) 58-70.

[12] N. Rappoport, and R. Shamir, Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 46 (2018) 10546-10562.

[13] Z. Yang, B. Liu, T. Lin, Y. Zhang, L. Zhang, and M. Wang, Multiomics analysis on DNA methylation and the expression of both messenger RNA and microRNA in lung adenocarcinoma. J Cell Physiol 234 (2019) 7579-7586.

[14] M. Zheng, Y. Hu, R. Gou, J. Wang, X. Nie, X. Li, Q. Liu, J. Liu, and B. Lin, Integrated multi-omics analysis of genomics, epigenomics, and transcriptomics in ovarian carcinoma. Aging (Albany NY) 11 (2019) 4198-4215.

[15] J.C. Guo, Y. Wu, Y. Chen, F. Pan, Z.Y. Wu, J.S. Zhang, J.Y. Wu, X.E. Xu, J.M. Zhao, E.M. Li, Y. Zhao, and L.Y. Xu, Protein-coding genes combined with long noncoding RNA as a novel transcriptome molecular staging model to predict the survival of patients with esophageal squamous cell carcinoma. Cancer Commun (Lond) 38 (2018) 4.

[16] C.H. Mermel, S.E. Schumacher, B. Hill, M.L. Meyerson, R. Beroukhim, and G. Getz, GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 12 (2011) R41.

[17] J.M. Taylor, Random Survival Forests. J Thorac Oncol 6 (2011) 1974-5.

[18] J. Meng, P. Li, Q. Zhang, Z. Yang, and S. Fu, A four-long non-coding RNA signature in predicting breast cancer survival. J Exp Clin Cancer Res 33 (2014) 84.

[19] G. Yu, Wang, L. G., Han, Y., He, Q. Y., clusterProfiler: an R package for comparing biological themes among gene clusters. Omics : a journal of integrative biology 16 (2012) 284–287.

[20] S.L. Kukreja, Löfberg, J., & Brenner, M. J. , A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. IFAC proceedings volumes 39 (2006) 814-819.

[21] D.T. Chen, A.H. Davis-Yadley, P.Y. Huang, K. Husain, B.A. Centeno, J. Permuth-Wey, J.M. Pimiento, and M. Malafa, Prognostic Fifteen-Gene Signature for Early Stage Pancreatic Ductal Adenocarcinoma. PLoS One 10 (2015) e0133562.

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

[23] P. Raman, R. Maddipati, K.H. Lim, and A. Tozeren, Pancreatic cancer survival analysis defines a signature that predicts outcome. PLoS One 13 (2018) e0201751.

[24] M. Wu, X. Li, T. Zhang, Z. Liu, and Y. Zhao, Identification of a Nine-Gene Signature and Establishment of a Prognostic Nomogram Predicting Overall Survival of Pancreatic Cancer. Front Oncol 9 (2019) 996.

[25] I. Dagogo-Jack, and A.T. Shaw, Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15 (2018) 81-94.

[26] I.P. Ribeiro, J.B. Melo, and I.M. Carreira, Cytogenetics and Cytogenomics Evaluation in Cancer. Int J Mol Sci 20 (2019).

[27] N. Cancer Genome Atlas, Comprehensive molecular portraits of human breast tumours. Nature 490 (2012) 61-70.

[28] C. Kandoth, M.D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J.F. McMichael, M.A. Wyczalkowski, M.D.M. Leiserson, C.A. Miller, J.S. Welch, M.J. Walter, M.C. Wendl, T.J. Ley, R.K. Wilson, B.J. Raphael, and L. Ding, Mutational landscape and significance across 12 major cancer types. Nature 502 (2013) 333-339.

[29] K.O. Wrzeszczynski, V. Varadan, J. Byrnes, E. Lum, S. Kamalakaran, D.A. Levine, N. Dimitrova, M.Q. Zhang, and R. Lucito, Identification of tumor suppressors and oncogenes from genomic and epigenetic features in ovarian cancer. PLoS One 6 (2011) e28503.

[30] G. Jeong, H. Bae, D. Jeong, J. Ham, S. Park, H.W. Kim, H.S. Kang, and S.J. Kim, A Kelch domain-containing KLHDC7B and a long non-coding RNA ST8SIA6-AS1 act oppositely on breast cancer cell proliferation via the interferon signaling pathway. Sci Rep 8 (2018) 12922.

[31] A. Martin-Pardillos, and S.R.Y. Cajal, Characterization of Kelch domain-containing protein 7B in breast tumours and breast cancer cell lines. Oncol Lett 18 (2019) 2853-2860.

[32] G. Zhang, E. Fan, G. Yue, Q. Zhong, Y. Shuai, M. Wu, G. Feng, Q. Chen, and X. Gou, Five genes as a novel signature for predicting the prognosis of patients with laryngeal cancer. J Cell Biochem (2019).

[33] M.E. Ray, G. Wistow, Y.A. Su, P.S. Meltzer, and J.M. Trent, AIM1, a novel non-lens member of the betagamma-crystallin superfamily, is associated with the control of tumorigenicity in human malignant melanoma. Proc Natl Acad Sci U S A 94 (1997) 3229-34.

[34] S. Hoshimoto, C.T. Kuo, K.K. Chong, T.L. Takeshima, Y. Takei, M.W. Li, S.K. Huang, M.S. Sim, D.L. Morton, and D.S. Hoon, AIM1 and LINE-1 epigenetic aberrations in tumor and serum relate to melanoma progression and disease outcome. J Invest Dermatol 132 (2012) 1689-97.

[35] P. Vainio, J.P. Mpindi, P. Kohonen, V. Fey, T. Mirtti, K.A. Alanen, M. Perala, O. Kallioniemi, and K. Iljin, High-throughput transcriptomic and RNAi analysis identifies AIM1, ERGIC1, TMED3 and TPX2 as potential drug targets in prostate cancer. PLoS One 7 (2012) e39801.

[36] E. Rosenbaum, S. Begum, M. Brait, M. Zahurak, L. Maldonado, L.A. Mangold, M.A. Eisenberger, J.I. Epstein, A.W. Partin, D. Sidransky, and M.O. Hoque, AIM1 promoter hypermethylation as a predictor of decreased risk of recurrence following radical prostatectomy. Prostate 72 (2012) 1133-9.

[37] Z. Yang, A. Liu, Q. Xiong, Y. Xue, F. Liu, S. Zeng, Z. Zhang, Y. Li, Y. Sun, and C. Xu, Prognostic value of differentially methylated gene profiles in bladder cancer. J Cell Physiol 234 (2019) 18763-18772.

[38] Y. Li, S. Ji, L. Fu, T. Jiang, D. Wu, and F. Meng, Over-expression of ARHGAP18 suppressed cell proliferation, migration, invasion, and tumor growth in gastric cancer by restraining over-activation of MAPK signaling pathways. Onco Targets Ther 11 (2018) 279-290.

[39] M.A. Aleskandarany, S. Sonbul, R. Surridge, A. Mukherjee, C. Caldas, M. Diez-Rodriguez, I. Ashankyty, K.I. Albrahim, A.M. Elmouna, R. Aneja, S.G. Martin, I.O. Ellis, A.R. Green, and E.A. Rakha, Rho-GTPase activating-protein 18: a biomarker associated with good prognosis in invasive breast cancer. Br J Cancer 117 (2017) 1176-1184.

[40] M.-N.G. Aguilar-Rojas A, Huerta-Reyes M, Pérez-Solis MA, Silva-García R, Guillén N, Olivo-Marin JC, Activation of human gonadotropin-releasing hormone receptor promotes down regulation of ARHGAP18 and regulates the cell invasion of MDA-MB-231 cells. Molecular and cellular endocrinology 460 (2018) 94-103.

[41] B. Humphries, Z. Wang, Y. Li, J.R. Jhan, Y. Jiang, and C. Yang, ARHGAP18 Downregulation by miR-200b Suppresses Metastasis of Triple-Negative Breast Cancer by Enhancing Activation of RhoA. Cancer Res 77 (2017) 4051-4064.

[42] Y. Pan, Y. Song, L. Cheng, H. Xu, and J. Liu, Analysis of methylation-driven genes for predicting the prognosis of patients with head and neck squamous cell carcinoma. J Cell Biochem 120 (2019) 19482-19495.

[43] W. Deng, Y. Wang, S. Zhao, Y. Zhang, Y. Chen, X. Zhao, L. Liu, S. Sun, L. Zhang, B. Ye, and J. Du, MICAL1 facilitates breast cancer cell proliferation via ROS-sensitive ERK/cyclin D pathway. J Cell Mol Med 22 (2018) 3108-3118.

[44] W. Deng, Y. Wang, L. Gu, B. Duan, J. Cui, Y. Zhang, Y. Chen, S. Sun, J. Dong, and J. Du, MICAL1 controls cell invasive phenotype via regulating oxidative stress in breast cancer cells. BMC Cancer 16 (2016) 489.

[45] R. Loria, G. Bon, V. Perotti, E. Gallo, I. Bersani, P. Baldassari, M. Porru, C. Leonetti, S. Di Carlo, P. Visca, M.F. Brizzi, A. Anichini, R. Mortarini, and R. Falcioni, Sema6A and Mical1 control cell growth and survival of BRAFV600E human melanoma cells. Oncotarget 6 (2015) 2779-93.

[46] S.S. Mello, L.J. Valente, N. Raj, J.A. Seoane, B.M. Flowers, J. McClendon, K.T. Bieging-Rolett, J. Lee, D. Ivanochko, M.M. Kozak, D.T. Chang, T.A. Longacre, A.C. Koong, C.H. Arrowsmith, S.K. Kim, H. Vogel, L.D. Wood, R.H. Hruban, C. Curtis, and L.D. Attardi, A p53 Super-tumor Suppressor Reveals a Tumor Suppressive p53-Ptpn14-Yap Axis in Pancreatic Cancer. Cancer Cell 32 (2017) 460-473 e6.

[47] R. Matera, and M.W. Saif, New therapeutic directions for advanced pancreatic cancer: cell cycle inhibitors, stromal modifiers and conjugated therapies. Expert Opin Emerg Drugs 22 (2017) 223-233.

Figures

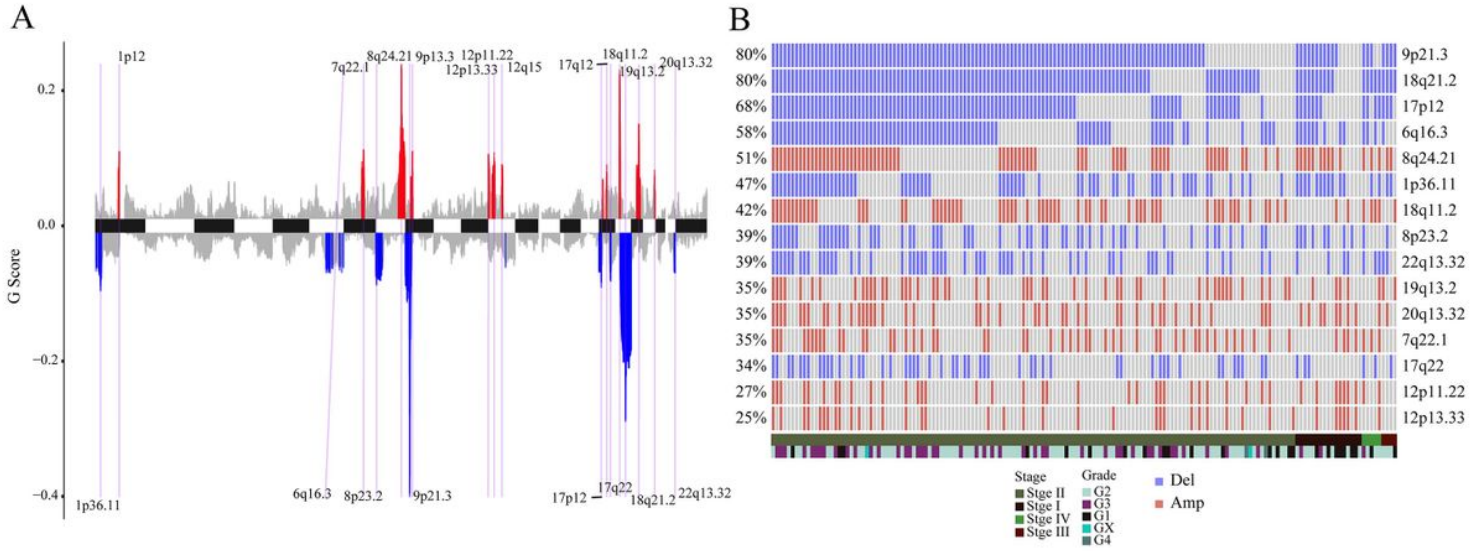

Figure 1

A. Significantly amplified gene fragments and significantly deleted gene fragments in the PAAD genome; The abscissa represents the chromosome number, and the ordinate expresses the score of copy number variation of the gene on each chromosome. A score greater than 0 indicates gene amplification, otherwise gene deletion. Red represents significantly increased copy number, blue represents significantly decreased copy number, and gray indicates insignificant change of copy number. The mutation locations on the chromosome were annotated at the top and the bottom of the Figure. B. CNV heatmap of pancreatic cancer genome. The abscissa represents the sample, the left ordinate represents the percentage of copy number variation, the right ordinate depicts the location of copy number deletion or amplification on the chromosome; red represents copy number amplification, blue represents copy number deletion.

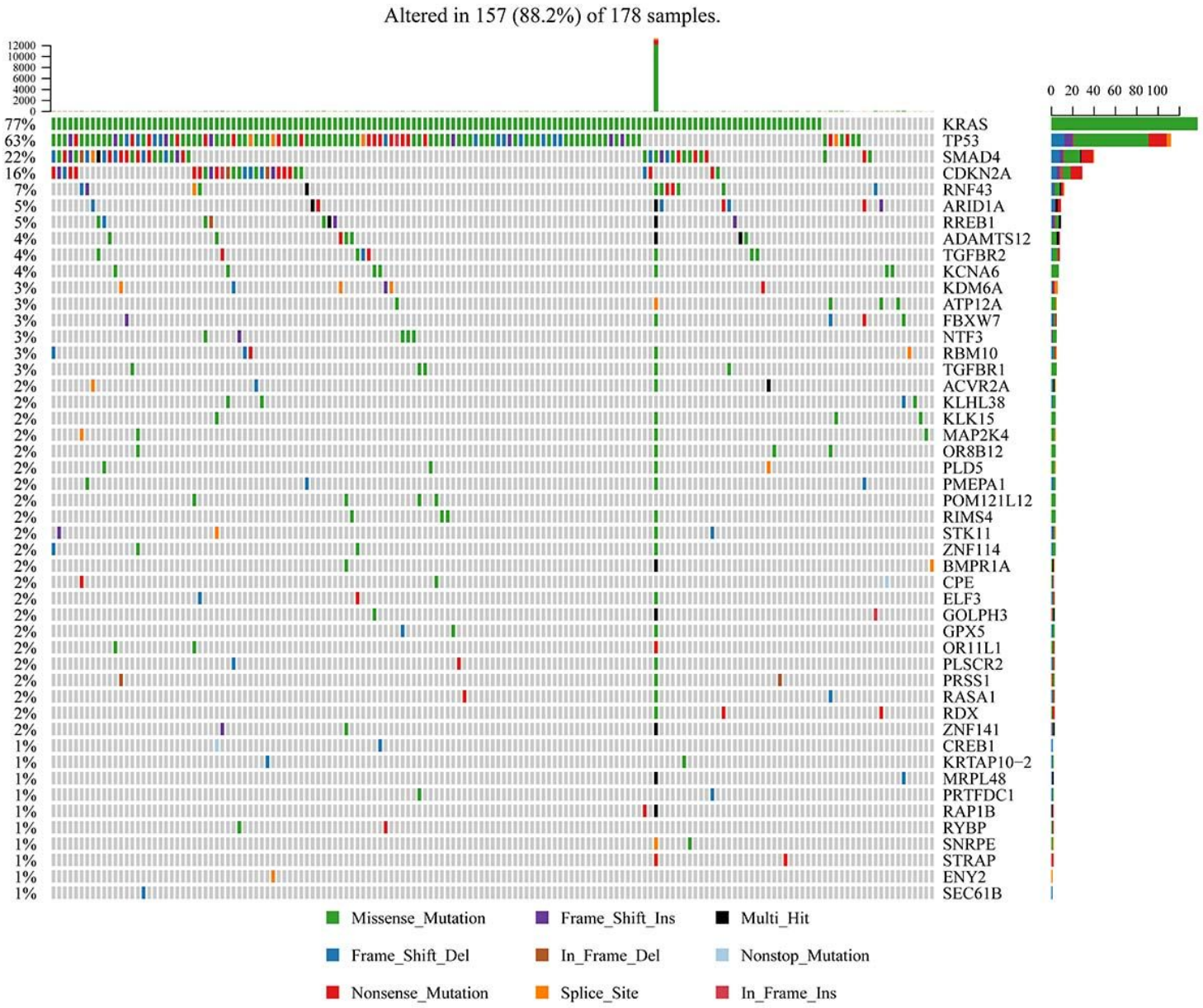

Figure 2

Oncoplot displaying the somatic landscape of PAAD cohort. Genes are ordered by their mutation frequency, and samples are ordered according to disease histology as indicated by the annotation bar (bottom). Side bar plot shows log10 transformed Q-values estimated by MutSigCV.

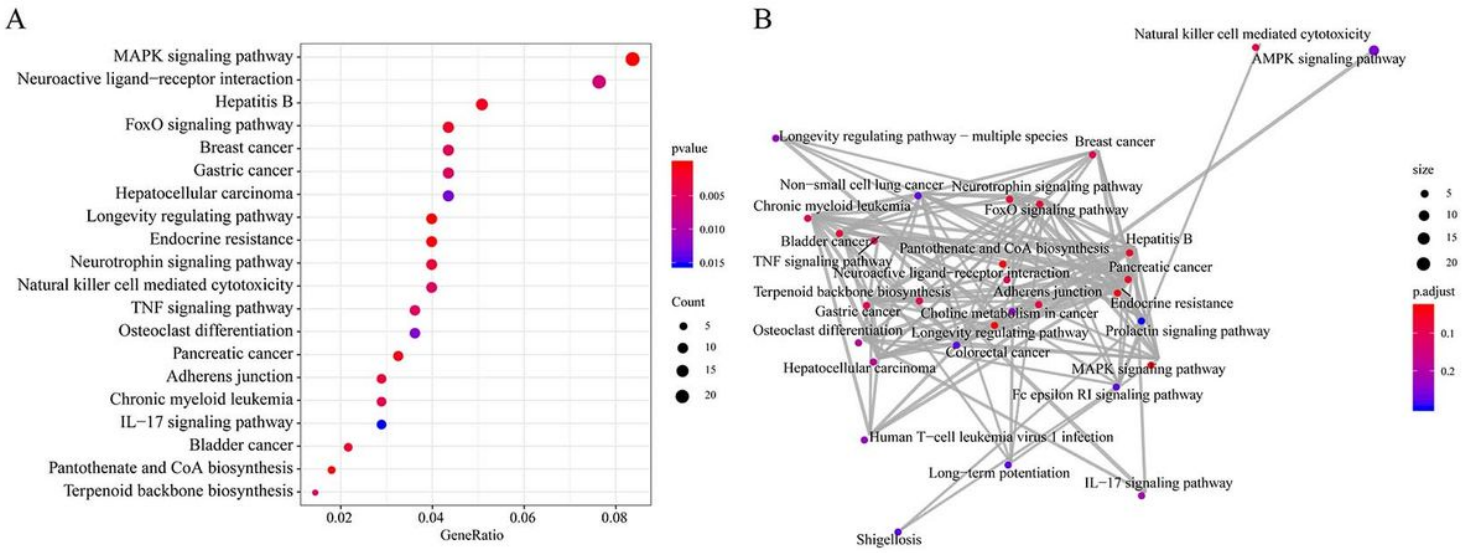

Figure 3

A. KEGG pathways that involved the 693 genes with copy number variation and mutation; B. Integrated functional network of KEGG enrichment. (Color from red to blue represents statistical significance, the bluer the smaller value of P; and the dot size represents the number of genes enriched in the pathway, a bigger dot indicates a larger number of genes.)

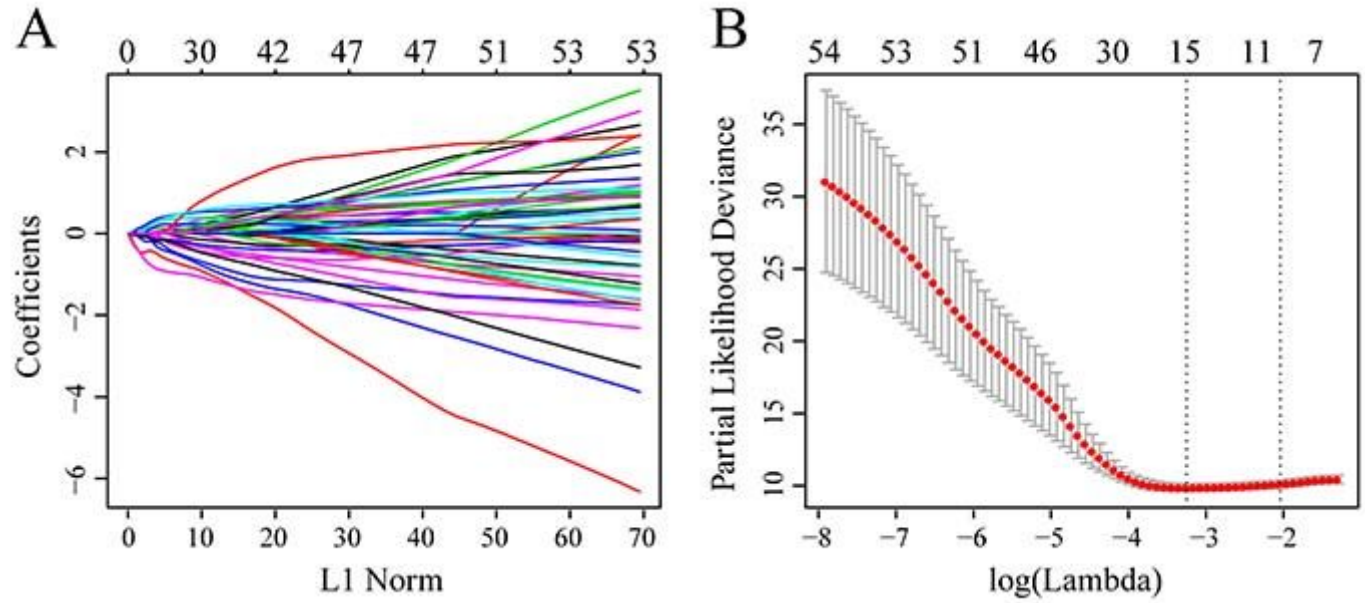

Figure 4

A. The changing trajectory of each independent variable. The abscissa represents the log value of the independent variable lambda, and the ordinate indicates the coefficient of such independent variable.

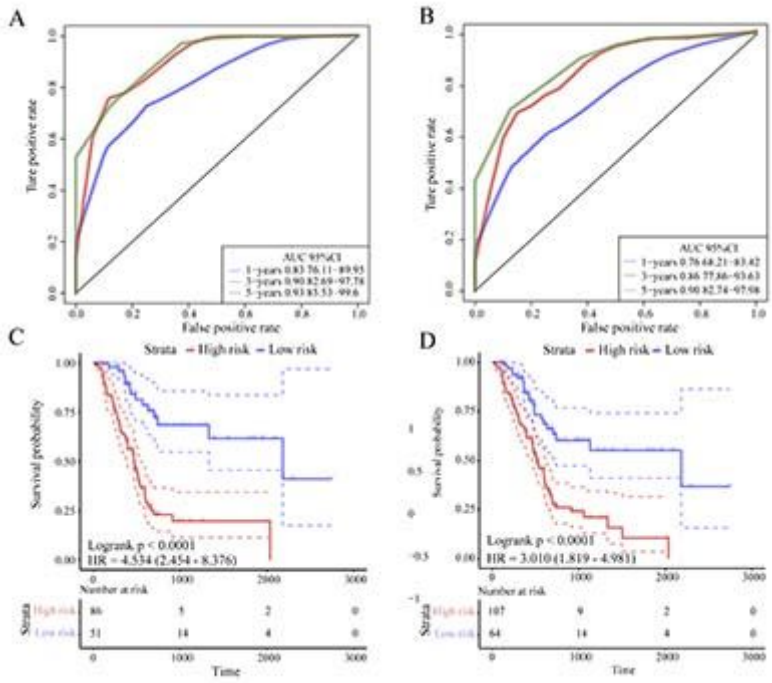

Figure 5

A. ROC curve based on 9-gene signature classification in the training cohort; B. ROC curve based on 9-gene signature classification in the TCGA-PADD cohorts ; C. KM curve based on 9-gene signature classification in the training cohort; D. KM curve based on 9-gene signature classification in the TCGA-PADD cohorts

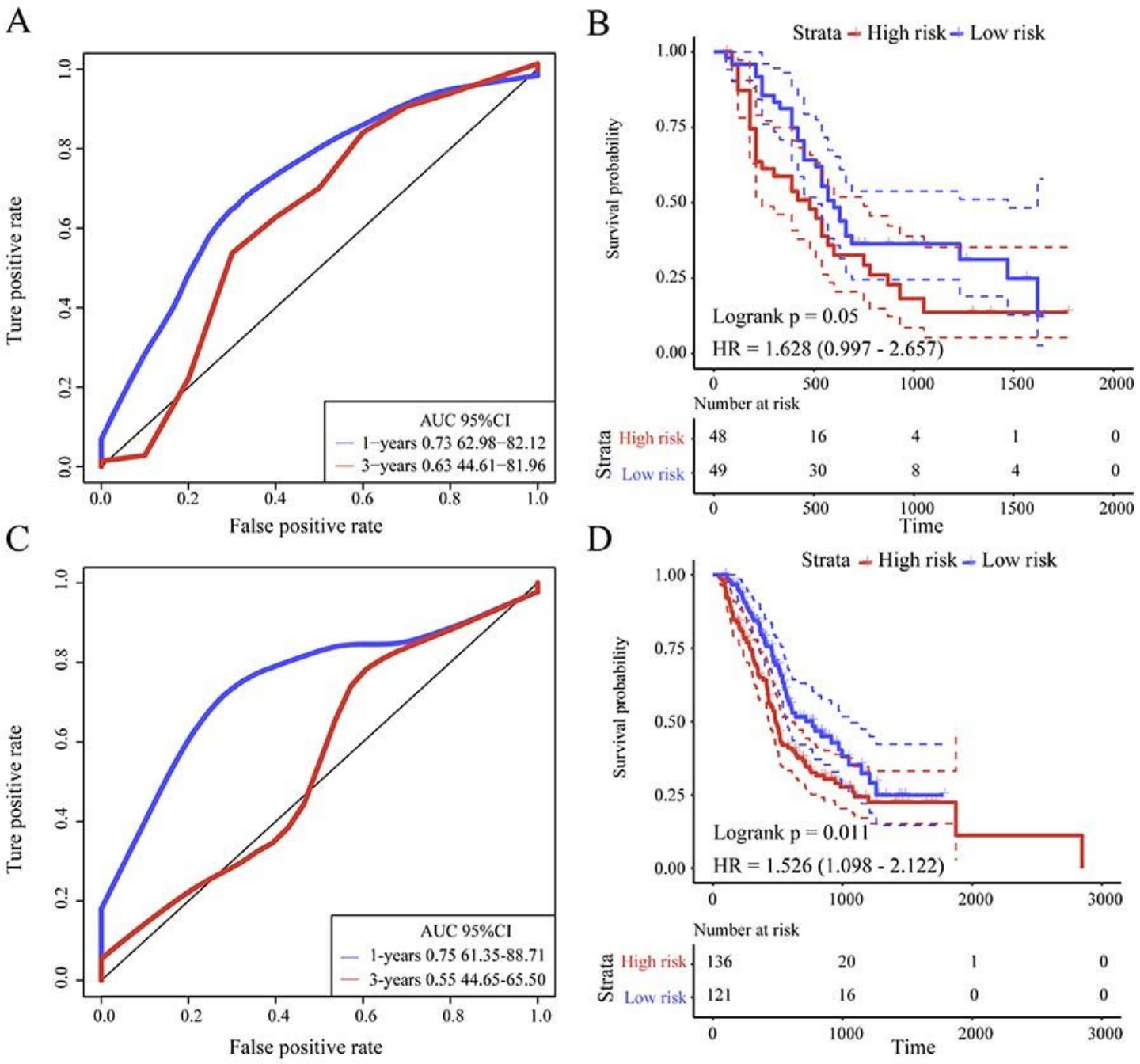

Figure 6

A. ROC curve of 9-gene signature classification in GSE21501 cohort; B. Distribution of 9-gene signature in the KM survival curve of GSE21501 cohort; C. ROC curve of 9-gene signature classification in ICGC cohort; D. distribution of 9-gene signature in the KM survival curve of ICGC cohort.

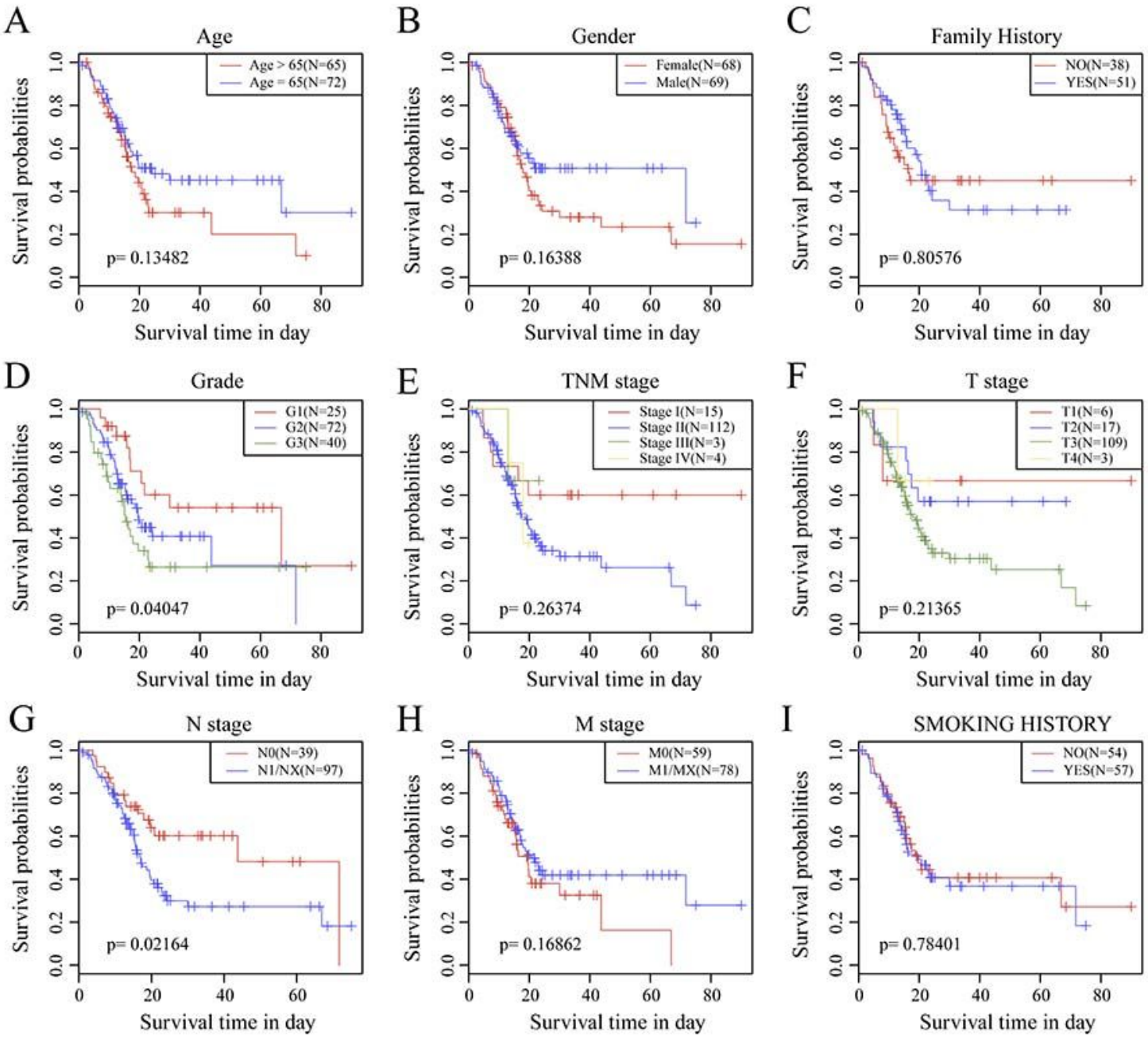

Figure 7

KM curves of different clinical features and OS prognosis

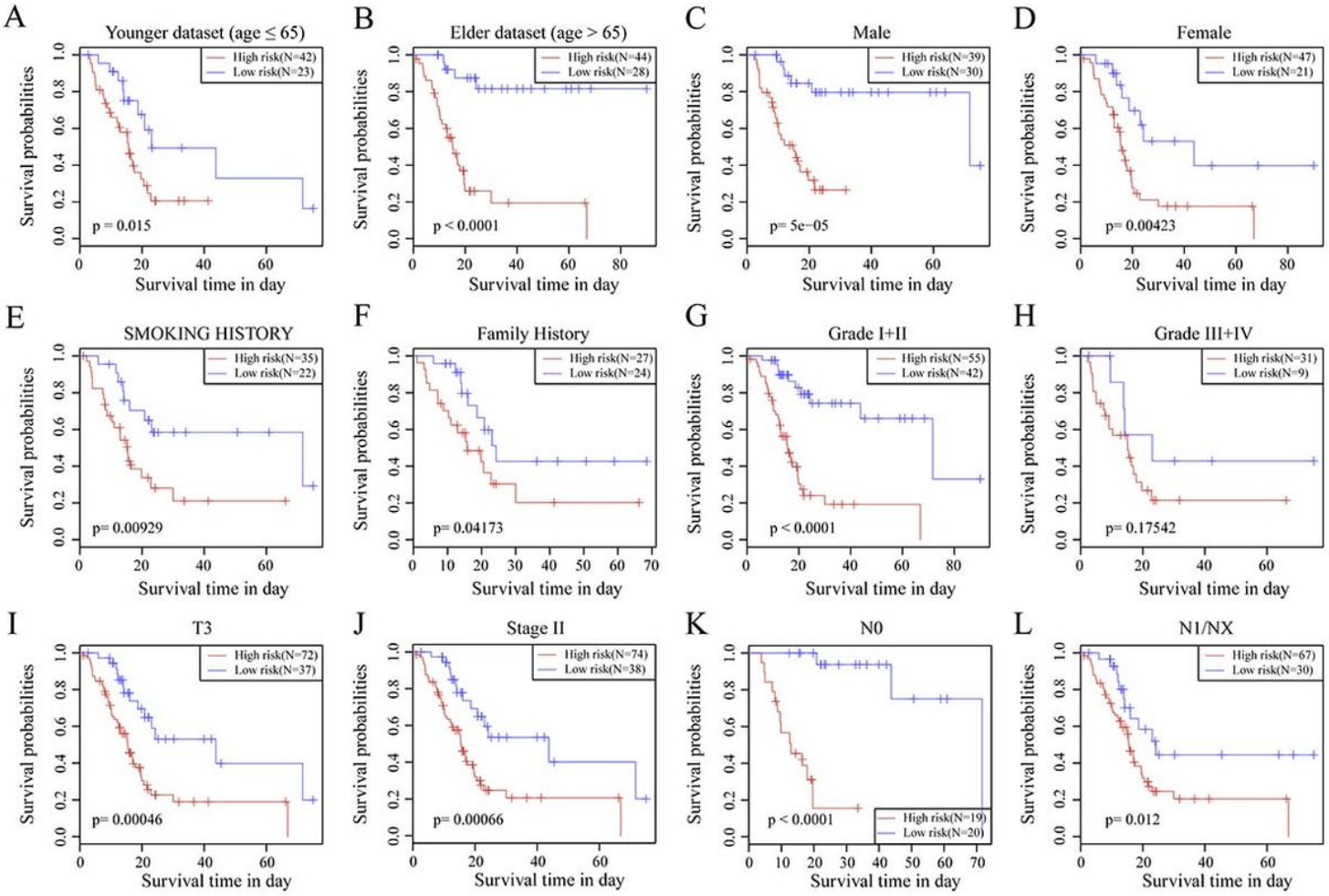

Figure 8

A. KM curve of samples with Age≤65; B. KM curve of samples with Age> 65; C. KM curve of male samples; D. KM curve of female samples; E. KM curve of samples with smoking history; F. KM curve of samples with family history; G. KM curve of Grade + samples; H. KM curve of Grade + samples; I. KM curve of T3-stage samples; J. KM curve of stage samples; K: KM curve of stage N0 samples; L: KM curve of N1 / NX stage sample.

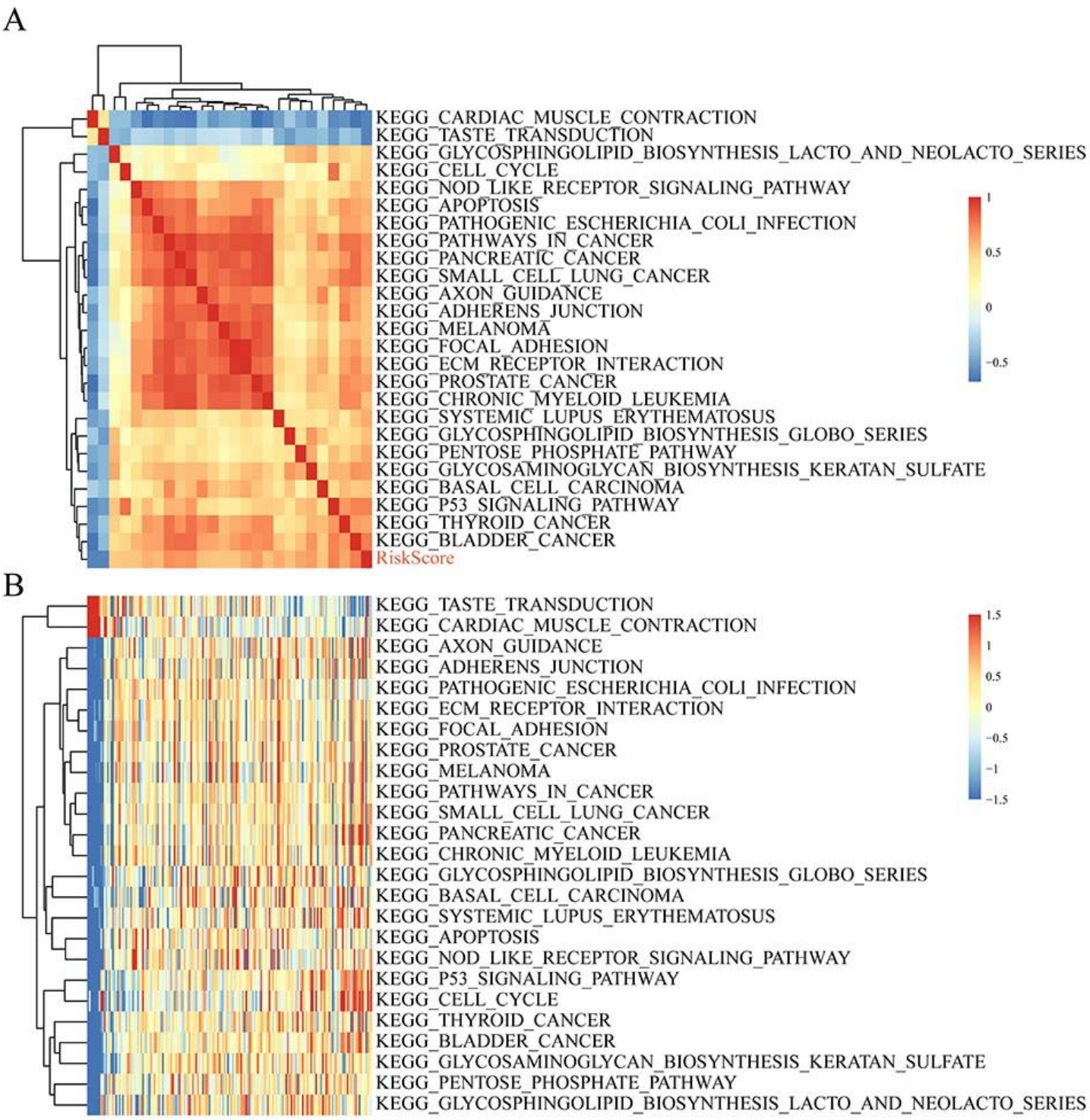

Figure 9

A. Clustering of KEGG pathways that are significantly correlated with riskscore (coefficient greater than 0.5); B: Change of relationship between KEGG pathways that are significantly correlated with riskscore (coefficient greater than 0.5) and ssGSEA score (the abscissa represents the sample, and the risk score increases from left to right)

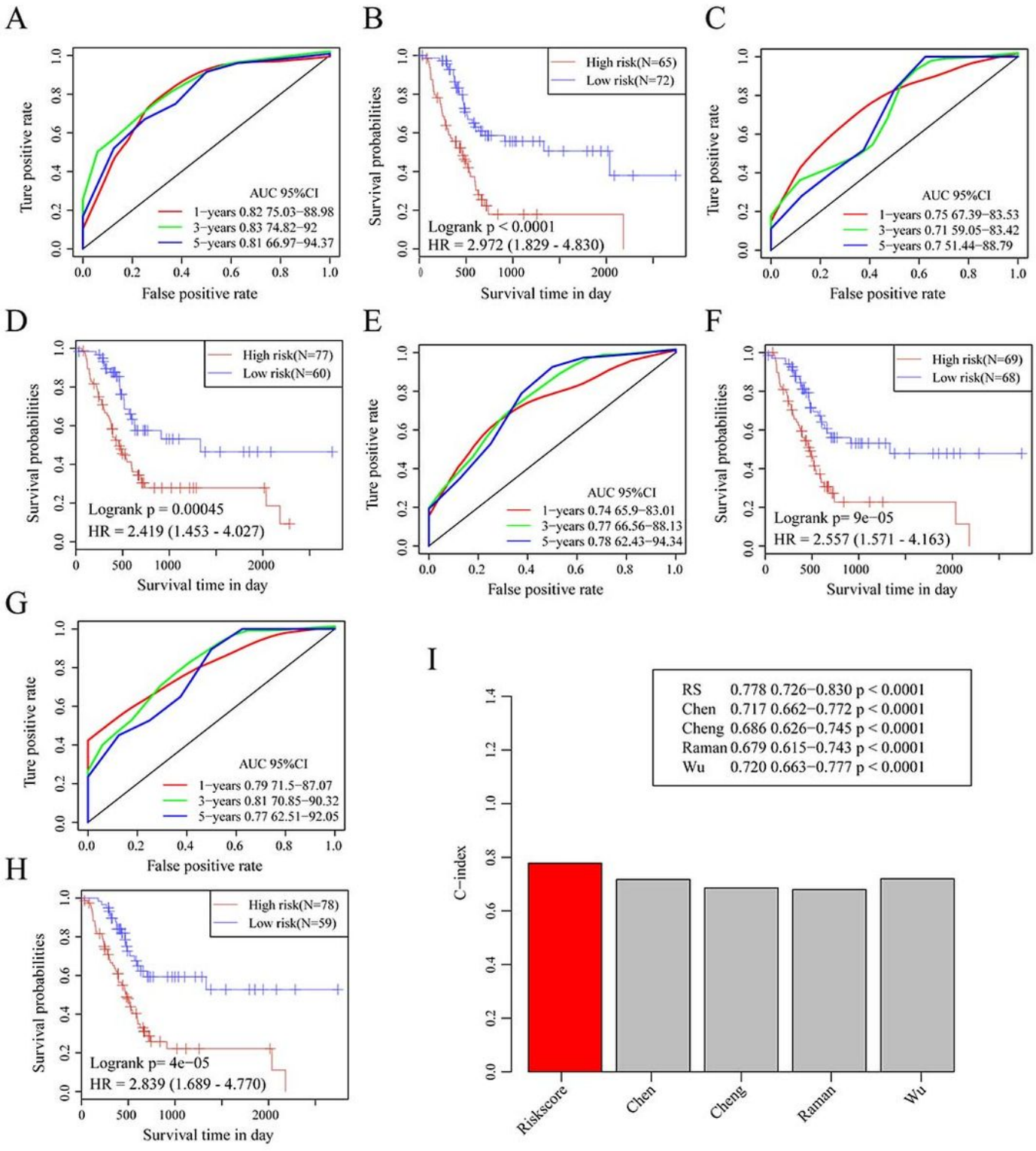

Figure 10

A. ROC curve of 15-gene signature in TCGA training cohort; B. KM curve of 15-gene signature in TCGA training cohort; C. ROC curve of 7-gene signature in TCGA training cohort; D. KM curve of 7-gene signature in TCGA training cohort; E. ROC curve of 5-gene signature in TCGA training cohort; F. KM curve of 5-gene signature in TCGA training cohort; G. ROC curve of 9-gene signature in TCGA training cohort; H. KM curve of 9-gene signature in TCGA training cohort; I. C-indexes of 5 prognostic risk models.

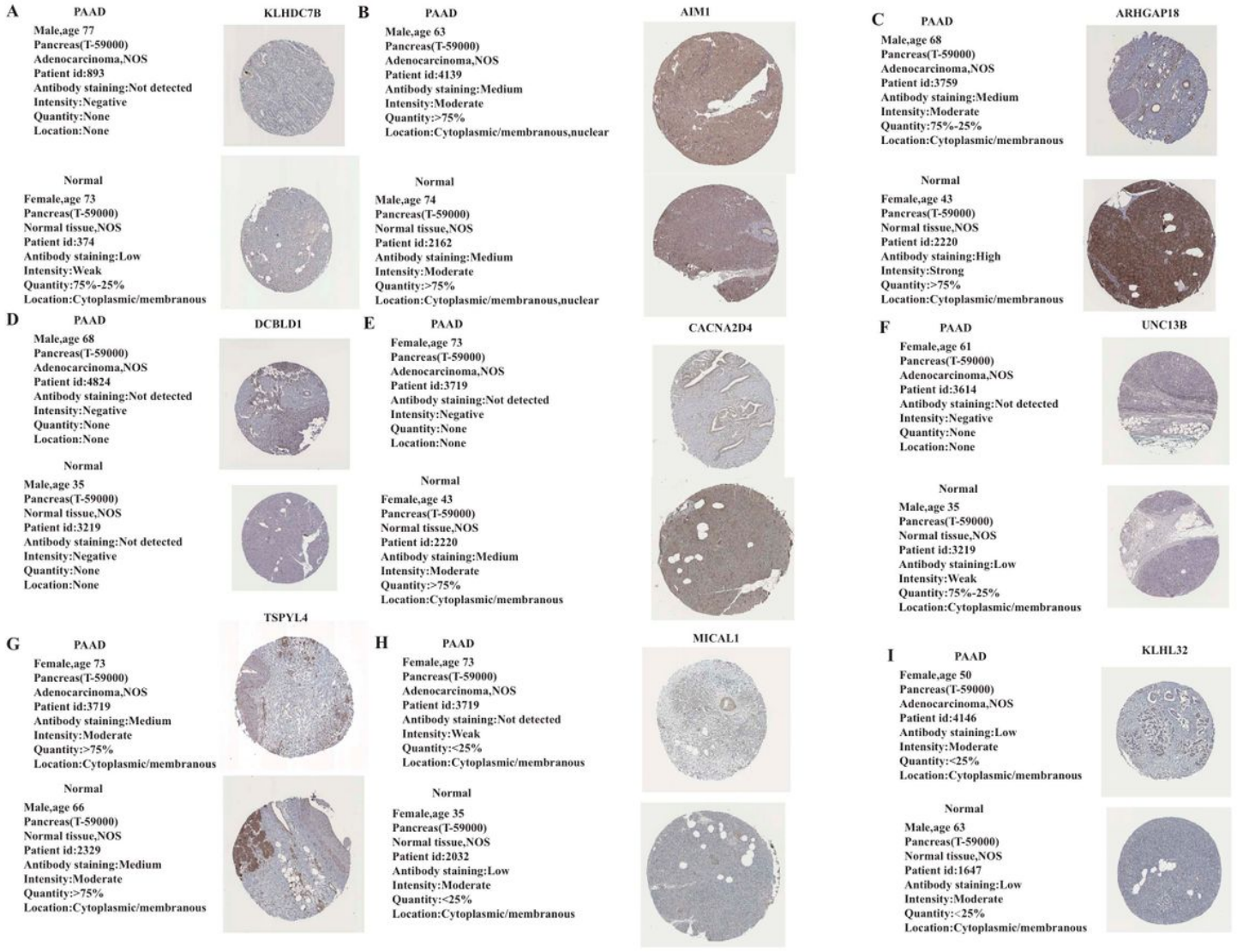

Figure 11

Analysis of protein expression of 9 genes.

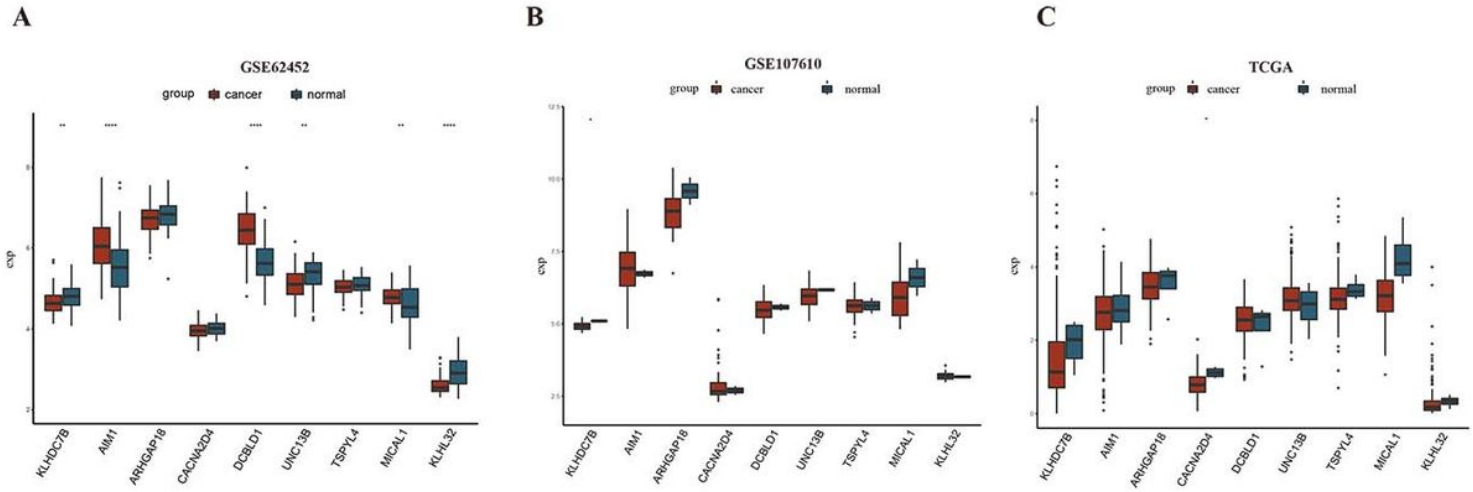

Figure 12

Analysis of mRNA expression of 9 genes. A: Differential expression analysis of 9 genes in GSE62452; B Differential expression analysis of 8 genes in GSE107610.C.Differential expression analysis of 9 genes in TCGA cohort.

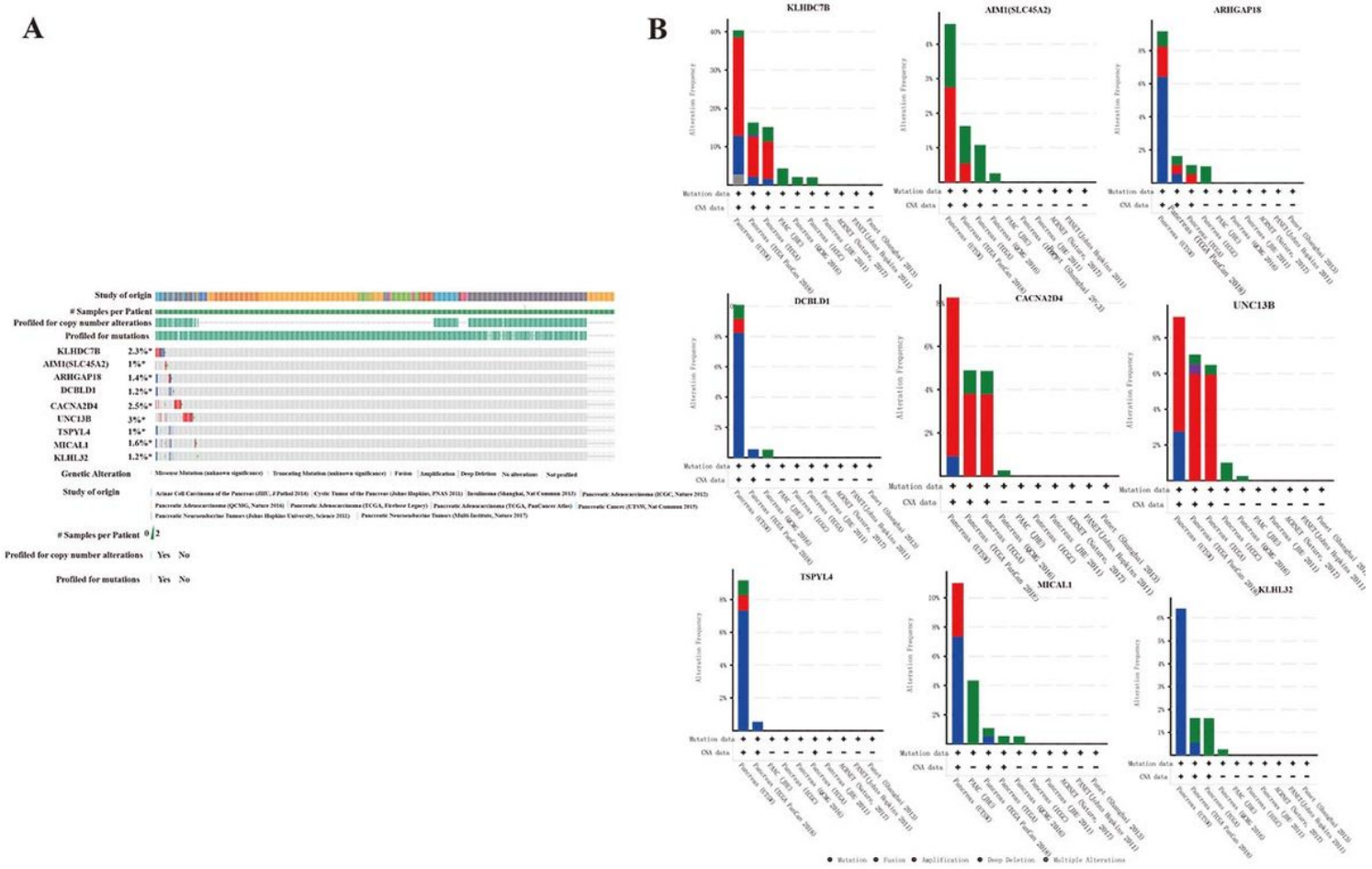

Figure 13

Analysis of mutations of 9 genes in the cBioportal (A) OncoPrint display of gene mutation distribution; (B) Histogram display of gene mutation distribution

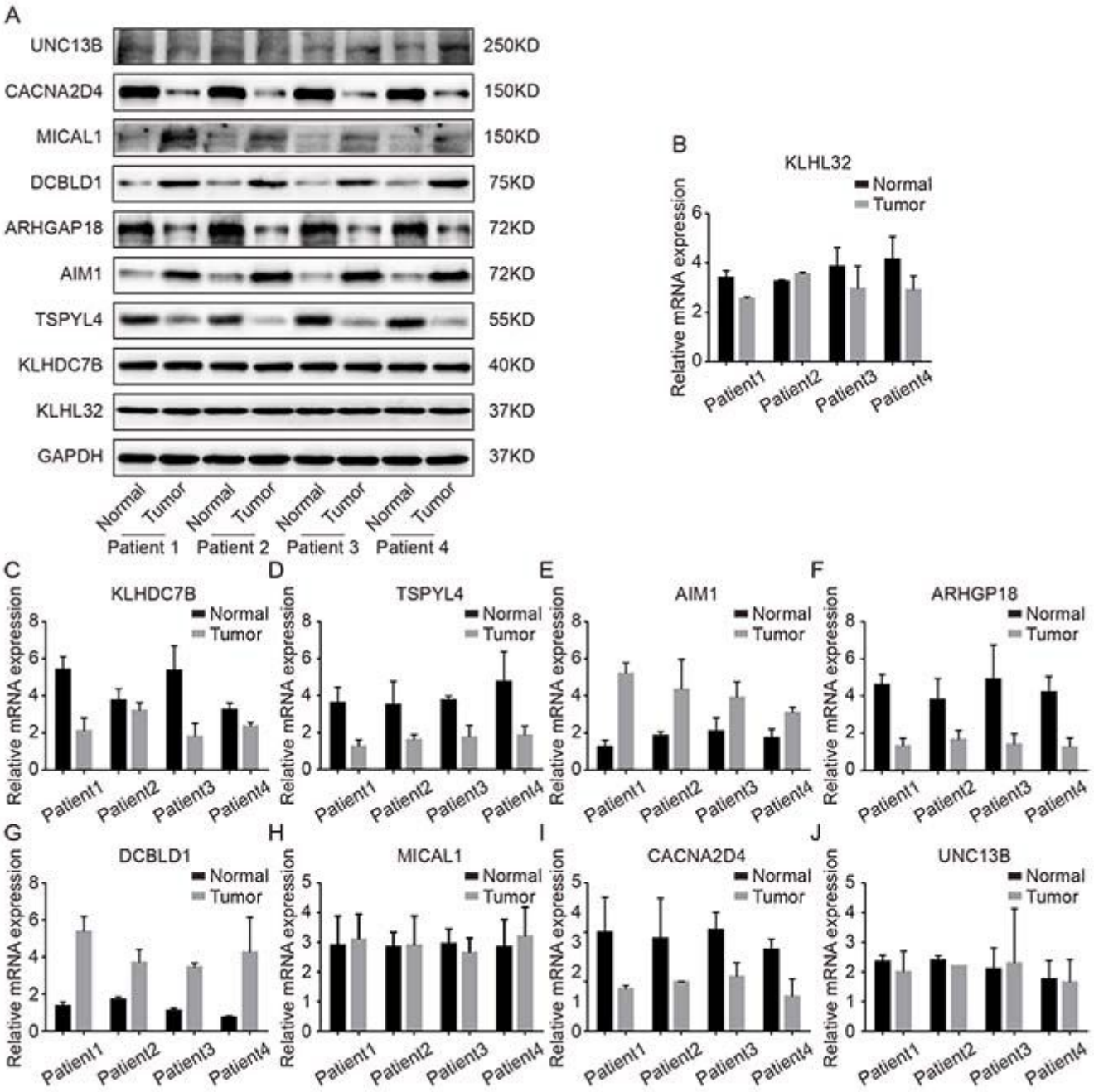

Figure 14

Clinical validation of 9 genes. A. The protein expression of 9 genes; B-J: The mRNA expression of 9 genes by qRT-PCR

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementTable1.pdf
- SupplementTable2.pdf