A Novel Prognostic Signature Based on Cell-In-Cell-Related Genes for Predicting Survival and Tumor Microenvironment in Pancreatic Cancer

Jianlu Song
Peking Union Medical College Hospital
Rexiati Ruze
Peking Union Medical College Hospital
Yuan Chen
Peking Union Medical College Hospital
Ruiyuan Xu
Peking Union Medical College Hospital
Xinpeng Yin
Peking Union Medical College Hospital
Chengcheng Wang
Peking Union Medical College Hospital
Yupei Zhao (zhao8028@263.net)
Peking Union Medical College Hospital

Research Article

Keywords: Pancreatic cancer, Cell-in-cell, Prognostic model, Tumor microenvironment, Immune infiltration, KRT7

Posted Date: December 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1199936/v1

Abstract

Background: Pancreatic cancer (PC) is a highly malignant tumor featured with high intra-tumoral heterogeneity and poor prognosis. Cell-in-cell (CIC) structures have been reported in multiple tumor types, and their presence is thought to promote clonal selection and tumor evolution. Here, we aimed to establish a CIC-related gene signature for predicting the prognosis and evaluating immune microenvironment in PC.

Methods: In this study, the gene expression data, as well as corresponding clinicopathological data of PC and normal pancreatic tissues were collected from The Cancer Genome Atlas (TCGA), Genotype-Tissue Expression (GTEx), International Cancer Genome Consortium (ICGC) and Gene Expression Omnibus (GEO) databases. Differential gene expression analysis, random forest screening, least absolute shrinkage and selection operator (LASSO) regression and multivariate Cox regression analysis were performed on 101 CIC-related genes to construct a prognostic gene signature. The effectiveness and robustness of the prognostic gene signature were evaluated by receiver operating characteristic (ROC) curves, Kaplan-Meier survival analysis and establishing the nomogram model. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were conducted to annotate the biological functions of the differentially expressed genes (DEGs). Quantitative real-time PCR (qRT-PCR), western blotting and immunohistochemistry (IHC) staining were validated the core gene expression in both mRNA and protein levels.

Results: A 4-gene signature was constructed to stratify patients into the low-risk and high-risk groups with distinct survival outcomes, somatic mutation profiles and immune features. The high-risk group had poorer prognosis than did the low-risk group. This signature was found to be an independent prognostic factor for PC patients with favorable predictive efficiency. Functional enrichment analyses showed that numerous terms and pathways associated with invasion and metastasis were enriched in the high-risk group. Moreover, the high-risk group had a higher tumor mutation burdens and lower immune cell infiltrations. *KRT7*, as the most important risk gene, was significantly associated with the worse prognosis of PC. CIC formation assay performing in PC cell lines indicated that *KRT7* expression was correlated with CIC frequency.

Conclusions: The signature based on four CIC-related genes could be applicable for predicting the prognosis of PC, and targeting CIC processes may be a potential therapeutic option. Further studies are needed to reveal the underlying molecular mechanisms and biological implications of CIC in PC progression.

Introduction

Pancreatic cancer (PC) is a highly lethal malignancy with rising trend in incidence and mortality worldwide, estimating that it will be the second leading cause of cancer related death by 2030 [1, 2]. Currently, curative surgery followed by adjuvant chemotherapy remains a standard therapeutic approach,

however, because of the concealed anatomical location of the pancreas, non-specific symptoms and deficiency of reliable biomarkers, effective screening is not available for PC, and most patients present with locally advanced or metastatic disease at the time of initial diagnosis, leading to missed opportunities for surgical intervention [3]. Therefore, it is essential to clarify the mechanisms of PC progression and to develop more effective therapeutic strategies.

Cell-in-cell (CIC) structures refer to the presence of one or more living cells internalized into another living one with the formation of “bird-eye cells”, which was first reported approximately 120 years ago in tumor tissues [4]. CIC is a long-standing phenomenon that was virtually neglected for decades but is attracting great interest in recent years. CIC structures have been observed in various types of tumors and are associated with the worse prognosis, such as breast cancer, lung cancer and PC [5-7]. From cellular mechanisms, cell cannibalism and entosis are two of the best-characterized CIC processes in cancers [8]. Cell cannibalism generally refers to the engulfment of live or dead cells within cancer cells through a mechanism involving actin, ezrin and caveolin. Previous studies have reported that cancer cells exert potent engulfment activity directed toward homotypic cancer cells, lymphocytes, neutrophils, natural killer cells and mesenchymal stem cells [4, 9-11]. Entosis is a specific form of cell cannibalism mainly induced by the extracellular matrix detachment, aberrant mitosis and glucose starvation [12]. It is thought to occur mostly between homotypic epithelial cells following the establishment of adherens junction via E-cadherin or P-cadherin, formation of mechanical ring enriched in vinculin and actomyosin contraction mediated by the Rho–ROCK–DIAPH1 signaling pathway [13]. Entosis is unlike cell cannibalism engulfed by outside cells, in which the internalized cells actively penetrate into outside cells driving by contractile actomyosin at the opposite cell cortex and subsequently die by lysosome-dependent degradation, leading to a non-apoptotic cell death [14]. In the past decade, numerous studies have provided evidence that the cannibalistic behavior is a hallmark of cancer, conferring cancer cells metabolic advantages under starvation conditions [8, 9, 15]. Moreover, researchers have shown that entosis can promote direct competition between cancer cells in mixed populations and ploidy changes of outside cells, affecting the clonal selection and evolution of cancer cell populations [16, 17]. A recent study has reported that in pancreatic ductal adenocarcinoma (PDAC), the most common type of PC, CIC structures were more prevalent in liver metastasis than primary tumor and poorly differentiated adenocarcinoma or adenosquamous carcinoma than well or moderately differentiated adenocarcinoma, suggesting that CIC phenomenon is associated with aggressive biology in PC [7]. Therefore, evaluating the CIC status is a powerful method for prognosis estimation. However, there are no studies focusing on the prognostic value of CIC-related gene signature and the molecular functions of them in PC.

In the present study, we systematically analyzed the expression profiles and prognostic values of CIC-related genes in PC patients from public datasets. Corresponding prognostic gene signature and a nomogram model were established and validated. To explore the potential mechanisms, we further performed functional enrichment analyses, and compared the somatic mutation profiles and immune features between two risk subgroups. The results of this study may help improve the current plight of PC treatment by designing combination therapeutic strategies based on targeting CIC-related processes.

Materials And Methods

Datasets and Processing

The RNA sequencing (RNA-seq) data of the Genotype-Tissue Expression (GTEx) and The Cancer Genome Atlas (TCGA) datasets were downloaded from the University of California Santa Cruz (UCSC) Xena website (https://xenabrowser.net/datapages/), which included 167 normal samples and 179 tumor samples. The expression data were normalized to transcripts per million (TPM) values and transformed to $\log_2$(TPM+1) for further analyses. The RNA-seq data of the International Cancer Genome Consortium (ICGC) dataset was also downloaded from UCSC Xena website, which included 96 tumor samples. The expression data were normalized to counts per million (CPM) and transformed to $\log_2$(CPM). The normalized expression matrix from microarray datasets of GSE21501 and GSE62452 were downloaded from Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/), which included 132 and 69 tumor samples, respectively. The corresponding clinicopathological information and somatic mutation data of TCGA and ICGC datasets were obtained from TCGA (http://portal.gdc.cancer.gov/) and ICGC (http://dcc.icgc.org/) official website. By combining expression data with corresponding clinicopathological information, samples with absent information of survival status, age, gender, TNM, AJCC stage, grade and patients losing to follow-up or follow-up time less than 30 days were removed. Meanwhile, two samples from GTEx dataset were eliminated because of having low expression levels (less 10,000 genes). Ultimately, GTEx (165 normal samples) and TCGA (167 tumor samples) datasets were chosen as the training cohort. ICGC (87 samples), GSE21501 (98 samples) and GSE62452 (64 samples) datasets were applied to external validation. The clinical characteristics of patients were presented in **Table 1**. The microarray and RNA-seq data of PC cell lines were obtained from GSE21654, GSE40098 and the Cancer Cell Line Encyclopedia (CCLE) database (https://sites.broadinstitute.org/ccle/). The single-cell RNA-seq data of the CRA001160 dataset, including over 50,000 individual pancreatic cells from 24 primary PDAC samples and 11 normal samples, were analyzed and visualized using the Tumor Immune Single-cell Hub (TISCH) website (http://tisch.comp-genomics.org/home/).

Identification of Differentially Expressed CIC-Related Genes

A total of 101 CIC-related genes were extracted from GeneCards website (http://www.genecards.org/), and prior literatures [4, 8, 12, 14], searching by the keywords “cell-in-cell”, “cell cannibalism” and “entosis”. The list of 101 CIC-related genes was shown in **Supplementary Table 1**. The “limma” R package was used to identify the differentially expressed genes (DEGs) between normal and tumor samples. False discovery rate (FDR) < 0.05 and $|\log_2(\text{Fold Change})| \geq 1$ were determined as significance criteria for selecting DEGs. The results of DEGs were visualized by volcano plot and heatmap.

Establishment and Verification of the CIC-Related Prognostic Signature

The random forest screening (using “randomForest” R package) and least absolute shrinkage and selection operator (LASSO) regression analysis (using “glmnet” R package) were applied to screen

important DEGs in TCGA cohort. And then, the overlapping CIC-related DEGs between these two methods were further selected to construct the best regression model by stepwise multivariate Cox regression analysis. Finally, four CIC-related genes prognostic signature was established and risk score of each patient was calculated by the regression coefficient of each gene and corresponding expression level using the following formula: Risk score = $(Expr_{gene1} \times Coef_{gene1}) + (Expr_{gene2} \times Coef_{gene2}) + \ldots + (Expr_{genen} \times Coef_{genen})$. For external validation cohorts, the risk score of each patient was calculated by the same formula. The patients were divided into low-risk and high-risk subgroups according to the median of risk score. Principal component analysis (PCA) was applied to visualize the clustering conditions of the prognostic signature. The Kaplan-Meier survival analysis was used to compare the differences in overall survival (OS) probability between two groups, and time-dependent receiver operating characteristic (ROC) analysis was used to evaluate the predictive accuracy of the prognostic gene signature. The univariate and multivariate Cox regression analysis were performed to determine the independent prognostic factors associated with OS.

The nomogram was established to predict the 1-, 2-, 3-year survival probability based on the risk score and other clinicopathological characteristics. The corresponding calibration curves were drawn to assess the efficiency of the nomogram. The differential expression of the four genes signature in normal pancreas and pancreatic cancer samples was verified by the Human Protein Atlas (HPA) online database (http://www.proteinatlas.org) from protein level in forms of immunohistochemical staining images.

Functional Enrichment Analysis for DEGs Between Two Risk Groups

Based on Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) databases, functional pathway enrichment analyses of DEGs were conducted to clarify the functions of genes by applying the “clusterProfiler” R package [18]. GeneMANIA (http://genemania.org/), a platform for gene prioritization and predicting gene function, was used to predict the related networks and functions of CIC-related genes [19].

Somatic Mutation Analysis and Immune Feature Estimation

The landscape of somatic mutation was analyzed and visualized by using the “maftools” R package. Tumor mutation burden (TMB) was calculated and compared of individuals between the low-risk and high-risk groups. The estimate score, stromal score, immune score and tumor purity were calculated by the ESTIMATE algorithm [20]. The CIBERSORT algorithm was performed to quantify the relative abundance of 22 immune cells infiltrated in tumor microenvironment (TME) [21]. The immune subtypes were classified by using the “ImmuneSubtypesClassifier” R package. Representative immune checkpoints were extracted from previous literatures and the expression levels of them were compared. The immunotherapy response was analyzed by ImmuCellAI website (http://bioinfo.life.hust.edu.cn/ImmuCellAI) [22].

Cell Culture

Four PC cell lines BxPC-3, CFPAC-1, PANC-1 and MIA PaCa-2 were purchased from the American Type Culture Collection (ATCC, Manassas, VA, USA). All cell lines were regularly tested for Mycoplasma and identified by Short Tandem Repeat (STR) identification. BxPC-3 cell line was cultured in RPMI-1640 medium (Corning, #10-040-CV), CFPAC-1 cell line was cultured in Iscove's Modified Dulbecco Medium (IMDM; Corning, #15-016-CV), and PANC-1 and MIA PaCa-2 cell lines were cultured in high glucose Dulbecco's Modified Eagle Medium (DMEM; Corning, #10-013-CMR). All medium were supplemented with 10% fetal bovine serum (HyClone, #SH30073.03) and 1% Penicillin-Streptomycin (Life Technologies, #15140-122). Cells were routinely maintained at 37°C with 5% CO_2.

RNA Extraction and Quantitative Real-Time PCR Analysis

Total RNA was extracted from cultured cells by the TRIzol reagent (Life Technologies, #15596-026) and cDNA synthesis was performed using the RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific™, #K1622) following the manufacturer's instructions. Quantitative real-time PCR (qRT-PCR) was performed in triplicate using SYBR Green Master Mix (Applied Biosystems, #A25742) [23]. The expression levels of *GAPDH* were used as the endogenous control and relative expression of *KRT7* was calculated using the $2^{-\Delta\Delta Ct}$ method. The primer sequences were used for qRT-PCR as follows:

KRT7: Forward 5'- CGAGGATATTGCCAACCGCAG-3',

Reverse 5'-CCTCAATCTCAGCCTGGAGCC-3';

GAPDH: Forward 5'-GTCTCCTCTGACTTCAACAGCG-3',

Reverse 5'- ACCACCCTGTTGCTGTAGCCAA-3'.

Western Blot Assay

Protein extracts from cells were prepared using 2% SDS lysis buffer including protein phosphatase inhibitor (Thermo Scientific™, #78440). Total protein (20 μg) was subjected to 10% (v/v) SDS-PAGE gels and transferred to PVDF membrane (Millipore). After blocking with 5% skimmed milk for 2 h at 37 °C, the membrane was incubated by primary antibodies against KRT7 (1:1000; Proteintech, #17513-1-AP), and β-ACTIN (1:2000; Proteintech, #20536-1-AP) at 4°C overnight, followed by incubation with HRP-conjugated secondary antibodies (1:5000; Proteintech, #SA00001-2) at room temperature for 1h. The protein bands were acquired through ECL kit (Beyotime, #P0018AM).

CIC Formation Assay

CIC formation assay was performed as previously described [13]. Briefly, about 2.0×10^5 cells were suspended in a six-well plate precoated with 1 mL solidified 0.5% soft agar for 8 h and then mounted onto glass slides by Cytospin preparation. Cells were fixed by 4% paraformaldehyde solution and immunostained with Phalloidin (Abcam, #ab176753) and DAPI (Abcam, #ab104139). Structures with more than half of cell internalized were counted as CIC structures.

Clinical Specimens and Immunohistochemical Analysis

A total of 24 patients with primary PDAC who underwent surgical resection at the Peking Union Medical College Hospital (PUMCH) were recruited in this study following the guidelines set by the Ethics Committee of Peking Union Medical College Hospital. The tumor and adjacent normal tissues were fixed by 10% formalin and embedded by paraffin. The sections of tissue specimens were used for immunohistochemistry (IHC) incubated with antibody against KRT7 (1:2000; Proteintech, #17513-1-AP). Manual staining and the estimation of IHC score were performed as previously described [23, 24].

Statistical Analysis

Statistical analyses and visualization were performed using R (version 4.1.0) software and GraphPad Prism 9 (version 9.3.0). Kaplan-Meier analysis and log-rank test were used to evaluate associations with survival time. Student's t test, Mann-Whitney test, and chi-square test were utilized for the comparison between two groups. For frequencies of CIC formation performed in cell lines and relative mRNA expression levels confirmed by qRT-PCR, data are means ± standard deviation (SD) of three independent experiments. All *P* values of statistical results were based on two-sided statistical tests, and a *P* value < 0.05 was considered to be statistically significant.

Results

Differential Gene Expression Analysis and Functional Enrichment Analysis of CIC-Related Genes

The expression levels of 101 CIC-related genes were explored in normal pancreas and PC samples using GTEx and TCGA datasets. PCA showed that the distribution difference between normal and tumor samples (**Figure 1A**). A total of 49 DEGs were identified, including 42 upregulated genes and 7 downregulated genes, and visualized by the heatmap (**Figure 1B**) and volcano plot (**Figure 1C**). GO analysis suggested that these DEGs were mainly involved in reactive oxygen species metabolic process, receptor-mediated endocytosis, cell leading edge, endocytic vesicle, tubulin binding and cytokine activity (**Supplementary Figure 1A**). Moreover, KEGG pathway analysis indicated that apoptosis, phagosome, regulation of actin cytoskeleton, ferroptosis, transcriptional misregulation in cancer and focal adhesion were enriched (**Supplementary Figure 1B**).

Construction of the CIC-Related Prognostic Signature in PC

To reduce the number of genes needed for constructing the prognostic signature, we firstly utilized random forest screening to assign an importance factor to each CIC-related DEGs. And then, top 20 genes, sorted by importance, were selected into further analysis (**Figure 1D**). Meanwhile, LASSO regression analysis was performed on 49 DEGs, and 18 candidate genes were retained by the most proper value of lambda (λ) (**Figure 1E**). Subsequently, we combined 10 overlapping candidate genes between these two methods to establish the best regression model by stepwise multivariate Cox regression analysis (**Figure 1F**). Finally, four significantly CIC-related genes contributing to OS in PC

patients were confirmed (**Figure 1G**), and the risk score of each patient was calculated using the following formula: Risk score = (0.362× expression level of *KRT7*) + (0.302 × expression level of *AURKA*) + (−0.114× expression level of *CDKN2A*) + (−0.377 × expression level of *RARB*). The correlation analysis among these four genes was shown in the circle plot (**Figure 1H**). Moreover, the related networks and functions were predicted using geneMANIA website (**Supplementary Figure 1C**). We found that related functions mainly involved in the regulation of cell cycle, mitosis-related process and kinase regulator activity.

Evaluation and Validation of the CIC-Related Prognostic Signature

Based on the median of risk scores, patients in TCGA cohort were separated into the low- and high-risk groups. The scatterplots showed that, as the patient's risk score increased, the number of deaths increased and the survival time decreased. Compared with the low-risk group, the expression levels of *KRT7* and *AURKA* in the high-risk group were upregulated, while the expression levels of *CDKN2A* and *RARB* were downregulated (**Figure 2A**). The Kaplan-Meier survival analysis indicated that patients in the high-risk group were significantly associated with shorter OS than those in the low-risk group (**Figure 2B**). The PCA revealed that patients with different risk levels were distributed into two clusters (**Figure 2D**). The area under curves (AUC) were 0.760, 0.766 and 0.807 in the 1-year, 2-year, and 3-year ROC curves, respectively (**Figure 2C**). We also found that, compared with other clinicopathological characteristics, the AUC value of risk score was much higher, suggesting that it was a better prognostic indicator for PC patients (**Figure 2E**). To demonstrate the robustness of the prognostic signature, the predictive efficiency was evaluated in three independent validation cohorts, including ICGC, GSE21501 and GSE62452. The patients from these three validation cohorts were stratified into the low- and the high-risk groups based on the median of risk scores calculated by using the same formula as in TCGA modeling cohort. Consistent with the results of TCGA cohort, patients in the high-risk group demonstrated the worse prognosis than those in the low-risk group. Similarly, the expression levels of *KRT7* and *AURKA* in the high-risk group were increased, while the expression levels of *CDKN2A* and *RARB* were decreased (**Figure 2F and 2G**; **Supplementary Figure 2A, 2B, 2F and 2G**). The PCA confirmed that patients in different subgroups could be divided into two separate directions (**Figure 2I**; **Supplementary Figure 2D and 2I**). Moreover, the AUC value of time-dependent ROC curves analysis reached around 0.700, indicating that the prognostic signature performed well in three validation cohorts (**Figure 2H**; **Supplementary Figure 2C and 2H**), and the risk score had better predictive accuracy compared with other clinicopathological characteristics (**Figure 2J**; **Supplementary Figure 2E and 2J**).

Independent Prognostic Value of the CIC-Related Prognostic Signature and Establishment of the Predictive Nomogram Model

The univariate and multivariate Cox regression analyses were performed to evaluate the prognostic power of the signature. Based on the multivariate Cox regression analysis, the risk score was confirmed to be an independent prognostic factor for OS prediction in all four cohorts (**Supplementary Table 2**). In order to further improve predictive efficiency, the risk score and other clinical characteristics such as age, gender, grade, and stage were together used to establish the predictive nomogram in TCGA and ICGC

cohorts. The C-index for the nomogram was 0.704 (95%CI 0.645–0.763) in TCGA cohort and 0.725 (95%CI 0.644–0.805) in ICGC cohort, indicating that the two nomograms both had well predictive performance (**Figure 3A and 3B**). Subsequently, time-dependent ROC curves, the calibration curves and the decision curve analysis (DCA) were applied to further evaluate the effectiveness of established nomograms. The AUCs of ROC curves for predicting 1-, 2-, and 3-year survival were 0.716, 0.785 and 0.810 in TCGA cohort (**Figure 3C**), 0.762, 0.787 and 0.890 in ICGC cohort (**Figure 3D**). In addition, the calibration curves presented satisfied coherence between observed and predicted 1-year, 2-year and 3-year OS in both cohorts (**Figure 3E and 3F**). DCA was performed to further assess the predictive efficacy between the risk score and nomogram, and the results demonstrated that the nomogram achieved the highest net benefits, suggesting that it was an efficient model to predict the prognosis of PC patients (**Figure 3G and 3H**).

Functional Enrichment Analyses and Somatic Mutation Profiles of the CIC-Related Prognostic Signature

To further explore the biological functions and pathways associated with the established prognostic signature, we firstly analyzed the DEGs between the high-risk and low-risk groups (**Figure 4A and 4B**). A total of 212 DEGs were identified in TCGA cohort, including 189 upregulated genes and 23 downregulated genes. Moreover, in ICGC cohort, 162 DEGs were identified, including 52 upregulated genes and 110 downregulated genes. Notably, *KRT7* was one of the top 10 upregulated genes sorting by adjust *P* value in both TCGA and ICGC cohorts (**Supplementary Figure 3A and 3B**). The GO analysis showed that the DEGs were enriched in several cell differentiation and tumor metastasis-related processes, such as epidermal cell differentiation, extracellular matrix organization, ameboidal-type cell migration, intermediate filament cytoskeleton and cell-cell junction (**Figure 4C and 4D**). And the KEGG pathway analysis demonstrated that the DEGs were mainly enriched in several pathways associated with tumor invasiveness and metastasis, such as PI3K−Akt signaling pathway, Wnt signaling pathway, Hippo signaling pathway, Focal adhesion, ECM−receptor interaction and Regulation of actin cytoskeleton (**Figure 4E and 4F**). To clarify whether the risk score was associated with the mutational landscapes of PC patients, we compared the somatic mutation profiles between the high-risk and low-risk groups in TCGA cohort (**Figure 5A–5D**). Notably, the mutation frequency in the high-risk group was 96.25%, while 78.21% in the low-risk group, indicating that the mutation frequency increased with the risk score elevation. Moreover, *KRAS* and *TP53* were the top two genes with the highest mutation frequencies in both subgroups. We also found that more co-occurrence and mutually exclusive mutations were observed in the high-risk group when compared with the low-risk group (**Figure 5E**). In addition, patients with higher risk scores demonstrated higher TMB levels ($P < 0.001$; **Figure 5F**). We further conducted the same analyses in ICGC cohort and similar results were verified (**Supplementary Figure 4**).

Analysis of Immune Features between Two Groups

PC harbored a highly heterogeneous TME. To further investigate whether the differences in prognosis between two risk groups were associated with immune cell infiltration, we firstly utilized ESTIMATE and CIBERSORT algorithms to explore the immune infiltration levels in TCGA and ICGC cohorts. We found that

the high-risk group was characterized with lower estimate score, stromal score and immune score, while higher tumor purity (**Figure 6A; Supplementary Figure 5A**). The composition and correlation of tumor-infiltrating immune cells showed that CIC-related prognostic signature was positively correlated with T cells regulatory (Tregs) and Macrophages M0, and negatively correlated with T cells CD4 memory resting, natural killer (NK) cells activated, Monocytes, Mast cells activated (**Figure 6B and 6C**) and Macrophages M2 (**Supplementary Figure 5B and 5C**). Patients in the high-risk group exhibited significantly higher infiltrating levels of B cells memory, Macrophages M0 ($P < 0.01$; **Figure 6D**) and Mast cells activated ($P < 0.05$; **Supplementary Figure 5D**), while lower infiltrating levels of B cells naive, Dendritic cells resting, Monocytes ($P < 0.05$; **Figure 6D**) and T cells CD8 ($P < 0.01$; **Supplementary Figure 5D**). Subsequently, the classification of immune subtypes showed that four subtypes and five subtypes were identified in TCGA cohort and ICGC cohort, respectively (**Figure 6E; Supplementary Figure 5E**). Patients in the high-risk group had more C1 (wound healing) and C2 (IFN-γ dominant) subtypes, and less C3 (inflammatory) subtype, suggesting the unfavorable prognosis. Emerging evidence has shown that immune features of TME are associated with the immune checkpoint blockade (ICB) therapeutic responses. The expression levels of some representative immune checkpoints including *CD274* (PD-L1), *CD276* (B7-H3), *CTLA4* (CTLA-4), *HAVCR2* (TIM3), *LAG3* (LAG-3), *PDCD1* (PD-1), *TIGIT* (VSIG9) and *VTCN1* (B7-H4) were compared between two risk groups. We found that *CD276* (**Figure 6F**), *CD274* and *VTCN1* (**Supplementary Figure 5F**) were upregulated in the high-risk group, while *TIGIT* was downregulated (**Figure 6F**). Besides, the results of ICB responses prediction demonstrated that the response rates were higher in the high-risk group, and responders had higher risk scores (**Figure 6G; Supplementary Figure 5G**). Although the rate of response and the risk score between two groups were not statistically significant in TCGA cohort, an increased tendency was observed in the high-risk group and responders, respectively. These findings indicated that patients with higher risk scores might be in an immunosuppressive state.

KRT7 is correlated with the unfavorable prognosis and CIC formation

As shown in the results of multivariate Cox regression and correlation analysis, *KRT7* was the most important risk gene in the established CIC-related gene signature for significantly predicting prognosis of PC patients and positively correlated with other three genes expression (**Figure 1G and 1H**). We investigated the protein and mRNA expression levels of *KRT7* by HPA database, GTEx and TCGA datasets. We found that the expression of *KRT7* in PC tissue was significantly higher than normal pancreas tissue and the expression level increased with the risk score elevation (**Figure 7A and 7B**). Moreover, the survival analysis showed that, based on the median of *KRT7* expression level, patients with *KRT7* high expression had shorter survival than those with low expression (**Figure 7C**). Given the extensive degree of intra-tumoral heterogeneity in PC, we further examined the expression of *KRT7* among diverse cell types in TME at single-cell level. The results demonstrated that *KRT7* was mainly expressed by malignant ductal cells and expression levels varied among different subpopulations of malignant cells (**Figure 7D**). We also analyzed the gene expression features and correlations with prognosis of *AURKA*, *CDKN2A* and *RARB*, but none of them was as significant as *KRT7* (**Supplementary Figure 6**). The related networks and functions of *KRT7* mainly involved in cytoskeleton remodeling, cell

differentiation and protein localization-related functions (**Figure 7E**). To clarify whether *KRT7* is associated with CIC formation, we compared the expression levels of it among different PC cell lines in three independent datasets (**Figure 7F**). Two cell lines with relatively higher and lower expression of *KRT7* at both mRNA and protein levels were selected respectively for CIC formation assay (**Figure 7G**). We found that BxPC-3 and CFPAC-1 cell lines with relatively higher expression levels of *KRT7* showed higher frequencies of CIC formation than PANC-1 and MIA PaCa-2 with relatively lower expression levels of *KRT7* (**Figure 7H**). Then, we performed IHC staining of KRT7 in 24 PDAC samples from the PUMCH cohort to further validate that high KRT7 expression was associated with unfavorable prognosis in PC (**Table 2; Figure 8A**). According to the median of IHC score, we next divided patients into low expression and high expression subgroups. Combined with KRT7 IHC scores and clinicopathological characteristics, patients from high expression group showed higher proportions of male, IIB-IV stage, lymphatic metastasis and death (**Figure 8B**). Although the results of survival analysis were not statistically significant (P = 0.1026), KRT7 as a risk factor (hazard ratio = 2.901), high expression group had poorer prognosis (**Figure 8C**).

Discussion

The global burden of PC has increased continuously over the past few decades, and as the leading cause of cancer-related death worldwide, its 5-year survival rate approached 10% for the first time in 2020 [1]. Because of no effective screening for PC diagnosis at an early stage, the vast majority of patients are found to be locally advanced or metastatic disease at the time of initial diagnosis. Even after standardized treatment, including surgery and adjuvant chemotherapy, patients still have to face the great risk of recurrence and death [3]. Therefore, it is urgent to confirm reliable biomarkers for early detection screening and predicting OS of PC patients.

Acquiring necessary nutrients from a frequently nutrient-poor environment and utilizing these nutrients to maintain rapid proliferation and progression is a common feature of cancer cell metabolism [25]. In addition to scavenging nutrients from extracellular microenvironment, cancer cells can also engulf and digest whole living cells via two mainly CIC processes, cell cannibalism or entosis, for nutrient recovery [26]. Previous studies have demonstrated that cannibalistic activity is a hallmark of cancer, conferring metabolic advantages on cancer cells under energy stress [8, 15]. Meanwhile, cancer cells with higher aggressiveness can become the “winner” subpopulation by eliminating their less competitive neighbors, which promote the fittest clones expanding within heterogeneous cancer cell populations [16]. CIC structures are prevalent in PC tissues and homotypic CIC (cancer cells internalized other cancer cells) constitutes the main subtype of overall CIC structures in PC [7, 27]. Considering that CIC phenomena in cancer represent a highly aggressive behavior, indicating that CIC-mediated cell competition, by selecting the best competitive clones, may be a potential mechanism to promote PC progression. However, there are no studies to investigate the relationship between CIC-related genes and PC patient’s OS through comprehensive bioinformatics analysis.

In the present study, we developed a risk scoring model based on four CIC-related genes (*KRT7*, *AURKA*, *CDKN2A* and *RARB*) in TCGA cohort, and further performed external validation for its robustness.

According to the value of hazard ratio, *KRT7* and *AURKA* were considered as the risk genes, while *CDKN2A* and *RARB* were considered as the protective genes. KRT7 (Keratin 7) belongs to type II cytokeratin involving in cytoskeleton remodeling, epithelial intermediate filaments formation and motility enhancement of cells [28]. Previous studies have observed KRT7 overexpression in many cancers, including ovarian, gastric, colorectal, and pancreatic cancers, which can facilitate cancer cells migration and metastasis [29-32]. AURKA (Aurora Kinase A) is a cell cycle-regulated kinase involving in microtubule formation, spindle pole stabilization during chromosome segregation, and contributing to tumorigenesis and progression [33]. AURKA-mediated phosphorylation is necessary for CIC-related processes, which promotes entosis in breast cancer cells via regulating microtubule plus-end dynamics and cell rigidity [34]. CDKN2A (Cyclin Dependent Kinase Inhibitor 2A), a well-known tumor suppressor, can induce cell cycle arrest in G1 and G2 phases by inhibiting binding of CDK4 or CDK6 with cyclin D1 and initiating p53-dependent cell cycle arrest. CDKN2A inactivation is found in the vast majority of PC patients, increasing cellular fitness and proliferation, and promotes homotypic CIC formation in breast cancer cells [35, 36]. RARB (Retinoic Acid Receptor Beta), by binding retinoic acid (biologically active vitamin A), can limit cell proliferation and abnormality to inhibiting tumorigenesis. A number of studies have documented that cancer cells show higher levels of RARB promoter methylation when compared with their normal counterparts, leading to functional silencing [37, 38].

Based on the risk score of individuals, patients were divided into the high-risk and low-risk groups. Our results showed that PC patients in the high-risk group had significantly poorer OS than those in the low-risk group, and the risk score was an independent prognostic factor with a credible efficacy evaluating by ROC analysis and nomogram model. To further explore the associations between established signature and differences of prognosis observed, biological functions, mutation profiles and immune features between two risk groups were compared in TCGA and ICGC cohorts. Functional enrichment analyses showed that several cancer-related terms and pathways were enriched, such as extracellular matrix organization, cell-cell junction, ECM-receptor interaction and PI3K-Akt signaling pathway, suggesting that patients in the high-risk group may be at higher degree of cancer-related pathways activation and immunosuppressive status [39]. Furthermore, a higher proportion of patients with *KRAS* and *TP53* somatic mutations were detected in the high-risk group, as well-known driver genes in PC, increasing cancer risk of individuals. TMB, as a numeric index, reflects cancer mutation quantity. Higher TMB results in more tumor neoantigens, which increases chances for immunotherapy and is clinically associated with better ICB responses [40]. In this study, patients in the high-risk group had significantly higher TMB, indicating that these patients may benefit from ICB therapy. Previous studies have shown that the immune microenvironment plays a pivotal role in PC progression [39, 41, 42]. However, the roles of CIC-related genes for PC immune microenvironment are still not clear. In our results, higher infiltration levels of M0 macrophages and lower infiltration levels of $CD8^+$ T cells were observed in the high-risk group, suggesting the presence of an immunosuppressive microenvironment [43]. Besides, it has been reported that tumor-infiltrating B cells contribute to responses to ICB therapy, and tumors of responders showed a higher infiltrating level of memory B cells, while lower infiltrating level of naïve B cells than tumors of non-responders [44]. This finding may be the reason for our results that there are a significantly higher

frequency of memory B cells and a significantly lower frequency of naïve B cells in the high-risk group. A previous study, by analyzing TCGA data, has identified six immune subtypes spanning cancer tissue types and molecular subtypes–wound healing (C1), IFN-γ dominant (C2), inflammatory (C3), lymphocyte depleted (C4), immunologically quiet (C5), TGF-β dominant (C6). C3 subtype has the best prognosis, while C1 and C2 subtypes present less favorable outcomes [45]. In our results, nearly half of patients in the low-risk group were C3 subtype, but most patients in the high-risk group were C1 and C2 subtypes, which was consistent with the association of immune subtypes and prognosis. ICB therapy is one of the most successful anti-cancer immunotherapies [46]. However, low response rates limit PC patients to benefit from ICB therapy [3]. We analyzed the expression levels of eight representative immune checkpoints between two risk groups and predicted ICB responses of patients. As shown in our results, the high-risk group exhibited higher expression of *CD274*, *CD276* and *VTCN1* than did the low-risk group. *CD274* encodes PD-L1 protein, the immune inhibitory ligand of PD-1 death receptor, that is observed in various types of tumors, and serves as an immune suppressor by blocking T cells activation and cytokine production [46]. *CD276* is widely expressed on a range of solid tumors, and plays a dual role in anti-tumor immunity, as a co-stimulatory regulator by enhancing the activity of T cells, or as a co-inhibitory regulator by inhibiting T cells and NK cells functions [47]. *VTCN1* is mainly expressed on tumor cells and tumor-associated macrophages, which promotes immune escape by inhibiting the proliferation of T cells and enhancing the function of regulatory T cells [48]. *TIGIT*, as the only downregulated immune checkpoint in the high-risk group, is primarily expressed on T cells and NK cells, which can inhibit anti-tumor immunity by impairing T cell functions, preventing NK cell-mediated lysis, and enhancing the suppressive activity of regulatory T cells [49]. Therefore, these findings may explain why a higher frequency ICB response rate was observed in the high-risk group. Highly intra-tumoral heterogeneity is the main obstacle to effective PC treatment.[50] The results of single-cell transcriptome analysis for PC have found that malignant cells contain distinct subpopulations, including those enriched for either proliferative or migrative features [41, 51]. Meanwhile, this heterogeneity is also found in established cancer cell line such as neuroblastoma, melanoma and breast carcinoma cell lines [52, 53]. Heterogeneous tumor populations can be divided into separate clusters with differently mechanical deformability [53]. Considering that softer tumor cells preferentially internalize stiffer neighboring cells in CIC processes, we speculate that CIC structures observed in PC may be a positive selection to promote the survival of clones with metabolic advantages and higher deformability, which confers a greater degree of the capability to invasion and metastasis on tumor cells [16, 54]. To further verify the correlation of KRT7 expression and unfavorable prognosis for PC patients, we performed IHC scoring on PUMCH cohort. According to the comprehensive analysis of IHC scores and clinicopathological characteristics, KRT7 as the most important risk gene in this prognostic signature, was demonstrated to be significantly associated with male, higher stage, and lymphatic metastasis. Since we only chose a small sample size (n = 24), this may be the reason why Kaplan-Meier survival analysis had no obviously statistical differences. A validation cohort with greater sample size and long-term follow-up was warranted in the future.

To the best of our knowledge, this study is the first attempt to construct a prognostic signature based on CIC-related genes and has shown a favorable efficacy on survival prediction in PC patients. However, our

study has several limitations. First, this study mainly collected retrospective data from public databases, and thus prospective studies are needed to further validate our results. Second, since there are insufficient researches to delineate the dynamic CIC processes in PC, related genes included in this study may not be the core regulators of CIC for PC cells. Third, due to the limitations of bioinformatics analysis, future studies are needed to systematically elucidate the molecular mechanisms and implications of CIC-related genes such as *KRT7* in PC.

In conclusion, our study developed a prognostic signature for PC based on four CIC-related genes to screen patients at high risk and predict survival. Further studies will be able to reveal new insights about CIC phenomena in cancer progression, which may be leveraged for PC therapy.

Abbreviations

PC: Pancreatic cancer; CIC: Cell-in-cell; TCGA: The Cancer Genome Atlas; GTEx: Genotype-Tissue Expression; ICGC: International Cancer Genome Consortium; GEO: Gene Expression Omnibus; LASSO: Least absolute shrinkage and selection operator; ROC: Receiver operating characteristic; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; DEGs: Differentially expressed genes; qRT-PCR: Quantitative real-time PCR; IHC: Immunohistochemistry; PDAC: Pancreatic ductal adenocarcinoma; RNA-seq: RNA sequencing; UCSC: University of California Santa Cruz; TPM: Transcripts per million; CPM: Counts per million; CCLE: Cancer Cell Line Encyclopedia; TISCH: Tumor Immune Single-cell Hub; FDR: False discovery rate; PCA: Principal component analysis; OS: Overall survival; HPA: Human Protein Atlas; TMB: Tumor mutation burden; TME: Tumor microenvironment; ATCC: American Type Culture Collection; STR: Short Tandem Repeat; IMDM: Iscove's Modified Dulbecco Medium; DMEM: Dulbecco's Modified Eagle Medium; PUMCH: Peking Union Medical College Hospital; SD: Standard deviation; AUC: Area under curves; DCA: Decision curve analysis; Tregs: T cells regulatory; Natural killer cells: NK cells; ICB: Immune checkpoint blockade; KRT7: Keratin 7; AURKA: Aurora Kinase A; CDKN2A: Cyclin Dependent Kinase Inhibitor 2A; RARB: Retinoic Acid Receptor Beta;

Declarations

Acknowledgments

We thank the GTEx, TCGA, GEO and ICGC databases for providing valuable datasets.

Availability of data and materials

The public datasets could be download at (https://xenabrowser.net/, https://portal.gdc.cancer.gov/, http://dcc.icgc.org/ and https://www.ncbi.nlm.nih.gov/geo/).

Authors' contributions

JS, RR, and YC collected and analyzed the data. JS and RR wrote the manuscript. YC, RX and XY generated figures and tables. CW and YZ designed the study and revised the manuscript. Obtained funding: CW and YZ. All authors read and approved the final manuscript.

Funding

This study was supported by the CAMS Innovation Fund for Medical Sciences (2021, 2021-I2M-1-002), the National Natural Science Foundation of China (2022, 82102810) and the fellowship of China Postdoctoral Science Foundation (2021, 2021M700501).

Ethics approval and consent to participate

This study conformed to the experimental guidelines of the World Medical

Association and the Ethics Committee of Peking Union Medical College

Hospital. Written informed consent were obtained from all the patients enrolled in this study.

Consent for publication

All of the authors agreed to publish this paper in Cancer Cell International.

Competing interests

The authors declare no competing interests.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A: **Cancer Statistics, 2021**. *CA Cancer J Clin* 2021, **71**(1):7-33.
2. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM: **Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States**. *Cancer Res* 2014, **74**(11):2913-2921.
3. Park W, Chawla A, O'Reilly EM: **Pancreatic Cancer: A Review**. *Jama* 2021, **326**(9):851-862.
4. Overholtzer M, Brugge JS: **The cell biology of cell-in-cell structures**. *Nat Rev Mol Cell Biol* 2008, **9**(10):796-809.
5. Zhang X, Niu Z, Qin H, Fan J, Wang M, Zhang B, Zheng Y, Gao L, Chen Z, Tai Y *et al*: **Subtype-Based Prognostic Analysis of Cell-in-Cell Structures in Early Breast Cancer**. *Front Oncol* 2019, **9**:895.
6. Mackay HL, Moore D, Hall C, Birkbak NJ, Jamal-Hanjani M, Karim SA, Phatak VM, Piñon L, Morton JP, Swanton C *et al*: **Genomic instability in mutant p53 cancer cells upon entotic engulfment**. *Nat Commun* 2018, **9**(1):3070.
7. Hayashi A, Yavas A, McIntyre CA, Ho YJ, Erakky A, Wong W, Varghese AM, Melchor JP, Overholtzer M, O'Reilly EM *et al*: **Genetic and clinical correlates of entosis in pancreatic ductal adenocarcinoma**.

Mod Pathol 2020, **33**(9):1822-1831.

8. Fais S, Overholtzer M: **Cell-in-cell phenomena in cancer**. *Nat Rev Cancer* 2018, **18**(12):758-766.
9. Lugini L, Matarrese P, Tinari A, Lozupone F, Federici C, Iessi E, Gentile M, Luciani F, Parmiani G, Rivoltini L *et al*: **Cannibalism of live lymphocytes by human metastatic but not primary melanoma cells**. *Cancer Res* 2006, **66**(7):3629-3638.
10. Fan J, Fang Q, Yang Y, Cui M, Zhao M, Qi J, Luo R, Du W, Liu S, Sun Q: **Role of Heterotypic Neutrophil-in-Tumor Structure in the Prognosis of Patients With Buccal Mucosa Squamous Cell Carcinoma**. *Front Oncol* 2020, **10**:541878.
11. Chen YC, Gonzalez ME, Burman B, Zhao X, Anwar T, Tran M, Medhora N, Hiziroglu AB, Lee W, Cheng YH *et al*: **Mesenchymal Stem/Stromal Cell Engulfment Reveals Metastatic Advantage in Breast Cancer**. *Cell Rep* 2019, **27**(13):3916-3926.e3915.
12. Galluzzi L, Vitale I, Aaronson SA, Abrams JM, Adam D, Agostinis P, Alnemri ES, Altucci L, Amelio I, Andrews DW *et al*: **Molecular mechanisms of cell death: recommendations of the Nomenclature Committee on Cell Death 2018**. *Cell Death Differ* 2018, **25**(3):486-541.
13. Wang M, Niu Z, Qin H, Ruan B, Zheng Y, Ning X, Gu S, Gao L, Chen Z, Wang X *et al*: **Mechanical Ring Interfaces between Adherens Junction and Contractile Actomyosin to Coordinate Entotic Cell-in-Cell Formation**. *Cell Rep* 2020, **32**(8):108071.
14. Overholtzer M, Mailleux AA, Mouneimne G, Normand G, Schnitt SJ, King RW, Cibas ES, Brugge JS: **A nonapoptotic cell death process, entosis, that occurs by cell-in-cell invasion**. *Cell* 2007, **131**(5):966-979.
15. Hamann JC, Surcel A, Chen R, Teragawa C, Albeck JG, Robinson DN, Overholtzer M: **Entosis Is Induced by Glucose Starvation**. *Cell Rep* 2017, **20**(1):201-210.
16. Sun Q, Luo T, Ren Y, Florey O, Shirasawa S, Sasazuki T, Robinson DN, Overholtzer M: **Competition between human cells by entosis**. *Cell Res* 2014, **24**(11):1299-1310.
17. Krajcovic M, Johnson NB, Sun Q, Normand G, Hoover N, Yao E, Richardson AL, King RW, Cibas ES, Schnitt SJ *et al*: **A non-genetic route to aneuploidy in human cancers**. *Nat Cell Biol* 2011, **13**(3):324-330.
18. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L *et al*: **clusterProfiler 4.0: A universal enrichment tool for interpreting omics data**. *Innovation (N Y)* 2021, **2**(3):100141.
19. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT *et al*: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W214-220.
20. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA *et al*: **Inferring tumour purity and stromal and immune cell admixture from expression data**. *Nat Commun* 2013, **4**:2612.
21. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D *et al*: **Determining cell type abundance and expression from bulk tissues with digital cytometry**. *Nat Biotechnol* 2019, **37**(7):773-782.

22. Miao YR, Zhang Q, Lei Q, Luo M, Xie GY, Wang H, Guo AY: **ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy**. *Adv Sci (Weinh)* 2020, **7**(7):1902880.
23. Wang C, Zhang T, Liao Q, Dai M, Guo J, Yang X, Tan W, Lin D, Wu C, Zhao Y: **Metformin inhibits pancreatic cancer metastasis caused by SMAD4 deficiency and consequent HNF4G upregulation**. *Protein Cell* 2021, **12**(2):128-144.
24. Chen Y, Wang C, Song J, Xu R, Ruze R, Zhao Y: **S100A2 Is a Prognostic Biomarker Involved in Immune Infiltration and Predict Immunotherapy Response in Pancreatic Cancer**. *Frontiers in Immunology* 2021, **12**(4907).
25. Pavlova NN, Thompson CB: **The Emerging Hallmarks of Cancer Metabolism**. *Cell Metab* 2016, **23**(1):27-47.
26. Krajcovic M, Krishna S, Akkari L, Joyce JA, Overholtzer M: **mTOR regulates phagosome and entotic vacuole fission**. *Mol Biol Cell* 2013, **24**(23):3736-3745.
27. Huang H, He M, Zhang Y, Zhang B, Niu Z, Zheng Y, Li W, Cui P, Wang X, Sun Q: **Identification and validation of heterotypic cell-in-cell structure as an adverse prognostic predictor for young patients of resectable pancreatic ductal adenocarcinoma**. *Signal Transduct Target Ther* 2020, **5**(1):246.
28. Owens DW, Lane EB: **The quest for the function of simple epithelial keratins**. *Bioessays* 2003, **25**(8):748-758.
29. Communal L, Roy N, Cahuzac M, Rahimi K, Köbel M, Provencher DM, Mes-Masson AM: **A Keratin 7 and E-Cadherin Signature Is Highly Predictive of Tubo-Ovarian High-Grade Serous Carcinoma Prognosis**. *Int J Mol Sci* 2021, **22**(10).
30. Huang B, Song JH, Cheng Y, Abraham JM, Ibrahim S, Sun Z, Ke X, Meltzer SJ: **Long non-coding antisense RNA KRT7-AS is activated in gastric cancers and supports cancer cell progression by increasing KRT7 expression**. *Oncogene* 2016, **35**(37):4927-4936.
31. Chen S, Su T, Zhang Y, Lee A, He J, Ge Q, Wang L, Si J, Zhuo W, Wang L: **Fusobacterium nucleatum promotes colorectal cancer metastasis by modulating KRT7-AS/KRT7**. *Gut Microbes* 2020, **11**(3):511-525.
32. Wang W, Wang J, Yang C, Wang J: **MicroRNA-216a targets WT1 expression and regulates KRT7 transcription to mediate the progression of pancreatic cancer-A transcriptome analysis**. *IUBMB Life* 2021, **73**(6):866-882.
33. Xie Y, Zhu S, Zhong M, Yang M, Sun X, Liu J, Kroemer G, Lotze M, Zeh HJ, 3rd, Kang R *et al*: **Inhibition of Aurora Kinase A Induces Necroptosis in Pancreatic Carcinoma**. *Gastroenterology* 2017, **153**(5):1429-1443.e1425.
34. Xia P, Zhou J, Song X, Wu B, Liu X, Li D, Zhang S, Wang Z, Yu H, Ward T *et al*: **Aurora A orchestrates entosis by regulating a dynamic MCAK-TIP150 interaction**. *J Mol Cell Biol* 2014, **6**(3):240-254.
35. Hayashi A, Hong J, Iacobuzio-Donahue CA: **The pancreatic cancer genome revisited**. *Nat Rev Gastroenterol Hepatol* 2021, **18**(7):469-481.

36. Liang J, Fan J, Wang M, Niu Z, Zhang Z, Yuan L, Tai Y, Chen Z, Song S, Wang X *et al*: **CDKN2A inhibits formation of homotypic cell-in-cell structures**. *Oncogenesis* 2018, **7**(6):50.
37. Niles RM: **Biomarker and animal models for assessment of retinoid efficacy in cancer chemoprevention**. *Acta Pharmacol Sin* 2007, **28**(9):1383-1391.
38. Orlando FA, Brown KD: **Unraveling breast cancer heterogeneity through transcriptomic and epigenomic analysis**. *Ann Surg Oncol* 2009, **16**(8):2270-2279.
39. Mao X, Xu J, Wang W, Liang C, Hua J, Liu J, Zhang B, Meng Q, Yu X, Shi S: **Crosstalk between cancer-associated fibroblasts and immune cells in the tumor microenvironment: new findings and future perspectives**. *Mol Cancer* 2021, **20**(1):131.
40. Jardim DL, Goodman A, de Melo Gagliato D, Kurzrock R: **The Challenges of Tumor Mutational Burden as an Immunotherapy Biomarker**. *Cancer Cell* 2021, **39**(2):154-173.
41. Ligorio M, Sil S, Malagon-Lopez J, Nieman LT, Misale S, Di Pilato M, Ebright RY, Karabacak MN, Kulkarni AS, Liu A *et al*: **Stromal Microenvironment Shapes the Intratumoral Architecture of Pancreatic Cancer**. *Cell* 2019, **178**(1):160-175.e127.
42. Grünwald BT, Devisme A, Andrieux G, Vyas F, Aliar K, McCloskey CW, Macklin A, Jang GH, Denroche R, Romero JM *et al*: **Spatially confined sub-tumor microenvironments in pancreatic cancer**. *Cell* 2021.
43. DeNardo DG, Ruffell B: **Macrophages as regulators of tumour immunity and immunotherapy**. *Nat Rev Immunol* 2019, **19**(6):369-382.
44. Helmink BA, Reddy SM, Gao J, Zhang S, Basar R, Thakur R, Yizhak K, Sade-Feldman M, Blando J, Han G *et al*: **B cells and tertiary lymphoid structures promote immunotherapy response**. *Nature* 2020, **577**(7791):549-555.
45. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA *et al*: **The Immune Landscape of Cancer**. *Immunity* 2018, **48**(4):812-830.e814.
46. Doroshow DB, Bhalla S, Beasley MB, Sholl LM, Kerr KM, Gnjatic S, Wistuba, II, Rimm DL, Tsao MS, Hirsch FR: **PD-L1 as a biomarker of response to immune-checkpoint inhibitors**. *Nat Rev Clin Oncol* 2021, **18**(6):345-362.
47. Flem-Karlsen K, Fodstad Ø, Tan M, Nunes-Xavier CE: **B7-H3 in Cancer - Beyond Immune Regulation**. *Trends Cancer* 2018, **4**(6):401-404.
48. Podojil JR, Miller SD: **Potential targeting of B7-H4 for the treatment of cancer**. *Immunol Rev* 2017, **276**(1):40-51.
49. Chauvin JM, Zarour HM: **TIGIT in cancer immunotherapy**. *J Immunother Cancer* 2020, **8**(2).
50. Wang S, Zheng Y, Yang F, Zhu L, Zhu XQ, Wang ZF, Wu XL, Zhou CH, Yan JY, Hu BY *et al*: **The molecular biology of pancreatic adenocarcinoma: translational challenges and clinical perspectives**. *Signal Transduct Target Ther* 2021, **6**(1):249.
51. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, Liu L, Huang D, Jiang J, Cui GS *et al*: **Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma**. *Cell Res* 2019, **29**(9):725-738.

52. Boeva V, Louis-Brennetot C, Peltier A, Durand S, Pierre-Eugène C, Raynal V, Etchevers HC, Thomas S, Lermine A, Daudigeos-Dubus E *et al*: **Heterogeneity of neuroblastoma cell identity defined by transcriptional circuitries**. *Nat Genet* 2017, **49**(9):1408-1413.
53. Lv J, Liu Y, Cheng F, Li J, Zhou Y, Zhang T, Zhou N, Li C, Wang Z, Ma L *et al*: **Cell softness regulates tumorigenicity and stemness of cancer cells**. *Embo j* 2021, **40**(2):e106123.
54. Gensbittel V, Kräter M, Harlepp S, Busnelli I, Guck J, Goetz JG: **Mechanical Adaptability of Tumor Cells in Metastasis**. *Dev Cell* 2021, **56**(2):164-179.

Tables

Table 1 Clinical characteristics of pancreatic cancer patients in the multiple datasets.

Variables	TCGA	ICGC	GSE21501	GSE62452
	n = 167 (%)	n = 87 (%)	n = 98 (%)	n = 64 (%)
Age				
<65	77 (46.1)	35 (40.2)	NA	NA
≥65	90 (53.9)	52 (59.8)	NA	NA
Gender				
Female	76 (45.5)	41 (47.1)	NA	NA
Male	91 (54.5)	46 (52.9)	NA	NA
Grade				
G1–G2	118 (70.7)	54 (62.1)	NA	33 (51.6)
G3–G4	49 (29.3)	33 (37.9)	NA	31 (48.4)
Stage				
I–II	160 (95.8)	NA	NA	48 (75.0)
III–IV	7 (4.2)	NA	NA	16 (25.0)
T				
T1–T2	27 (16.2)	13 (14.9)	18 (18.4)	NA
T3–T4	140 (83.8)	74 (85.1)	80 (81.6)	NA
N				
N0/NX	49 (29.3)	28 (32.2)	12 (12.2)	NA
N1	118 (70.7)	59 (67.8)	86 (87.8)	NA
Status				
Alive	76 (45.5)	32 (36.8)	35 (35.7)	15 (23.4)
Dead	91 (54.5)	55 (63.2)	63 (64.3)	49 (76.6)

TCGA, The Cancer Genome Atlas; ICGC, International Cancer Genome Consortium; NA, not available.

Table 2 Clinical characteristics of pancreatic cancer patients in the PUMCH cohort.

Variables	Total	KRT7 High expression	KRT7 Low expression
	n = 24 (%)	n = 11 (%)	n = 13 (%)
Age (year)			
<65	16 (66.7)	8 (72.7)	8 (61.5)
≥65	8 (33.3)	3 (27.3)	5 (38.5)
Gender			
Female	9 (37.5)	3 (27.3)	6 (46.2)
Male	15 (62.5)	8 (72.7)	7 (53.8)
Tumor grade			
Moderately	17 (70.8)	8 (72.7)	9 (69.2)
Poorly	7 (29.2)	3 (27.3)	4 (30.8)
Stage			
I-IIA	11 (45.8)	3 (27.3)	8 (61.5)
IIB-IV	13 (54.2)	8 (72.7)	5 (38.5)
T			
T1-T2	14 (58.3)	7 (63.6)	7 (53.8)
T3-T4	10 (41.7)	4 (36.4)	6 (46.2)
N			
N0	12 (50.0)	3 (27.3)	9 (69.2)
N1	12 (50.0)	8 (72.7)	4 (30.8)
M			
M0	22 (91.7)	10 (90.9)	12 (92.3)
M1	2 (0.3)	1 (9.1)	1 (7.7)
Status			
Alive	15 (62.5)	5 (45.5)	10 (76.9)
Dead	9 (37.5)	6 (54.5)	3 (23.1)
OS time (months)			
Median (range)	18 (2-38)	15 (2-38)	18 (6-36)
KRT7 IHC score			

Median (range)	6 (1-12)	9 (8-12)	6 (1-6)

OS, overall survival.

Figures

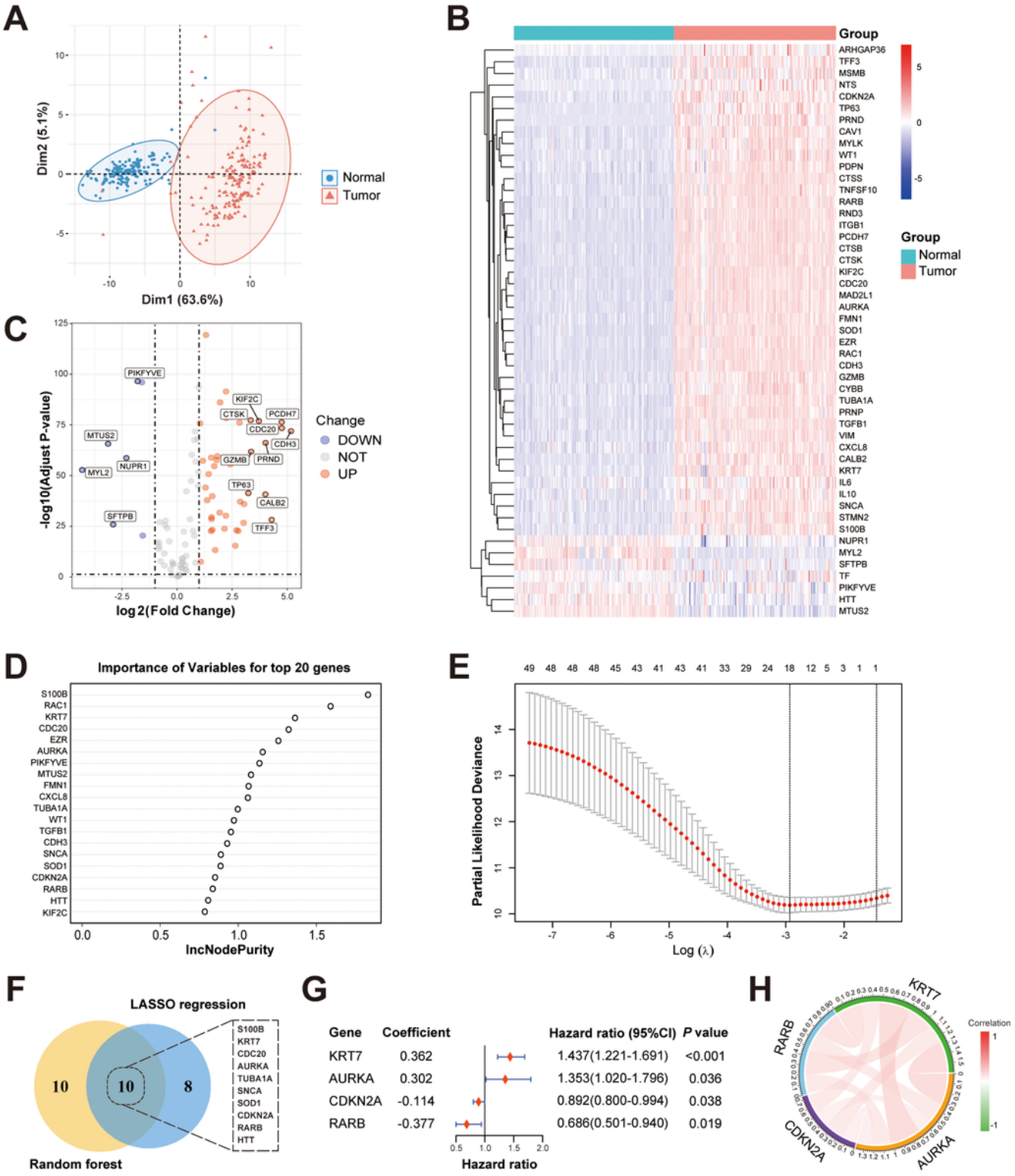

Figure 1

Identification of DEGs and construction of the CIC-related prognostic signature.

(A) PCA based on CIC-related genes of tumor and normal samples from the TCGA and GTEx datasets. **(B and C)** Heatmap and volcano plot of CIC-related DEGs between normal and tumor samples. **(D)** Top 20 genes sorted by importance of variables using random forest screening. **(E)** The most proper log (λ) value in LASSO regression analysis. **(F)** Ten overlapping genes based on the results of random forest screening and LASSO regression analysis. **(G)** The results of multivariate Cox regression analysis for 4 significantly CIC-related genes contributing to OS in PC. **(H)** The correlation analysis of the 4 genes.

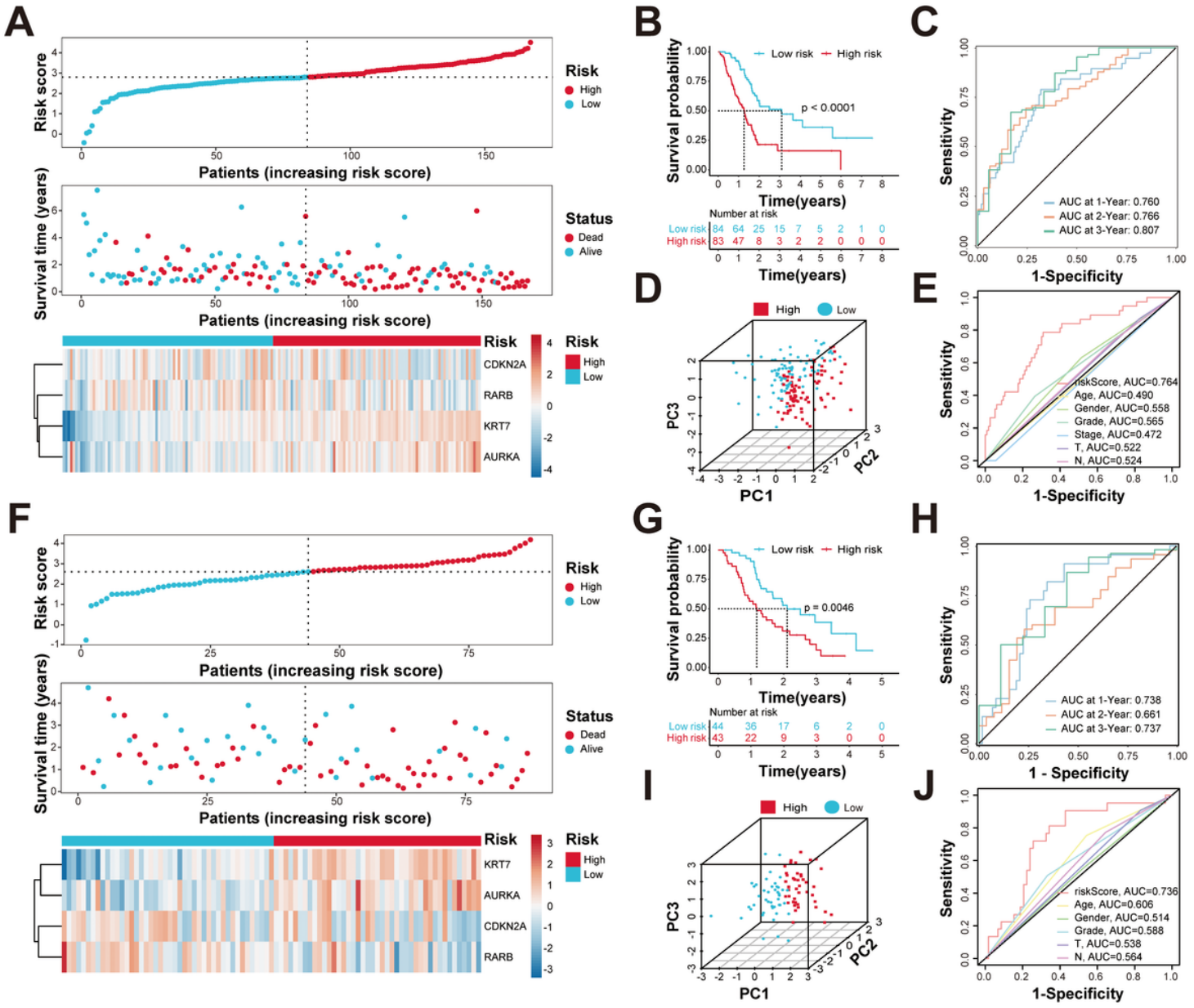

Figure 2

Evaluation and validation of CIC-related prognostic signature in TCGA and ICGC cohorts.

(A and F) Distribution of risk scores, OS status overview, and 4 genes expression in TCGA (A) and ICGC (F) cohorts. **(B and G)** Kaplan-Meier curves for OS of patients between the low- and high-risk groups in TCGA (B) and ICGC (G) cohorts. **(C and H)** ROC curves for 1-, 2- and 3-year OS prediction of the prognostic signature in TCGA (C) and ICGC (H) cohorts. **(D and I)** PCA based on the prognostic signature in TCGA (D) and ICGC(I) cohorts. **(E and J)** ROC curves of the risk score and other clinicopathological characteristics in TCGA (E) and ICGC (J) cohorts.

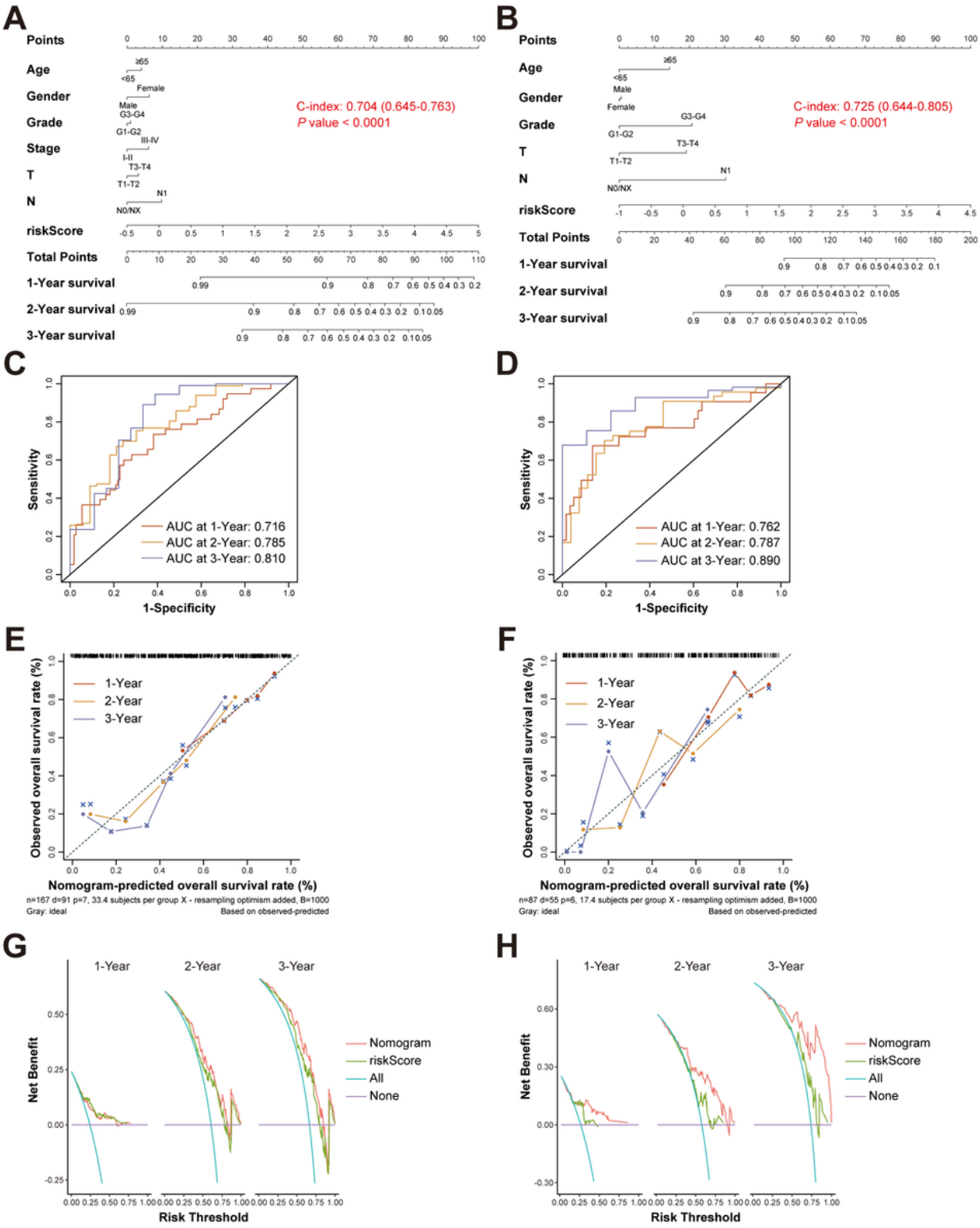

Figure 3

Establishment and evaluation of the predictive nomogram model.

(A and B) Nomograms based on the risk score and clinicopathological variables for predicting the probability of 1-, 2-, 3-year OS in TCGA (A) and ICGC (B) cohorts. **(C and D)** Time-dependent ROC analysis of the nomogram in TCGA (C) and ICGC (D) cohorts. **(E and F)** Calibration curves of the nomogram in terms of agreement between observed and predicted 1-, 2- and 3-year survival probability in TCGA (E) and ICGC (F) cohorts. **(G and H)** The 1-, 2- and 3-year DCA curves of the nomogram in TCGA (E) and ICGC (F) cohorts.

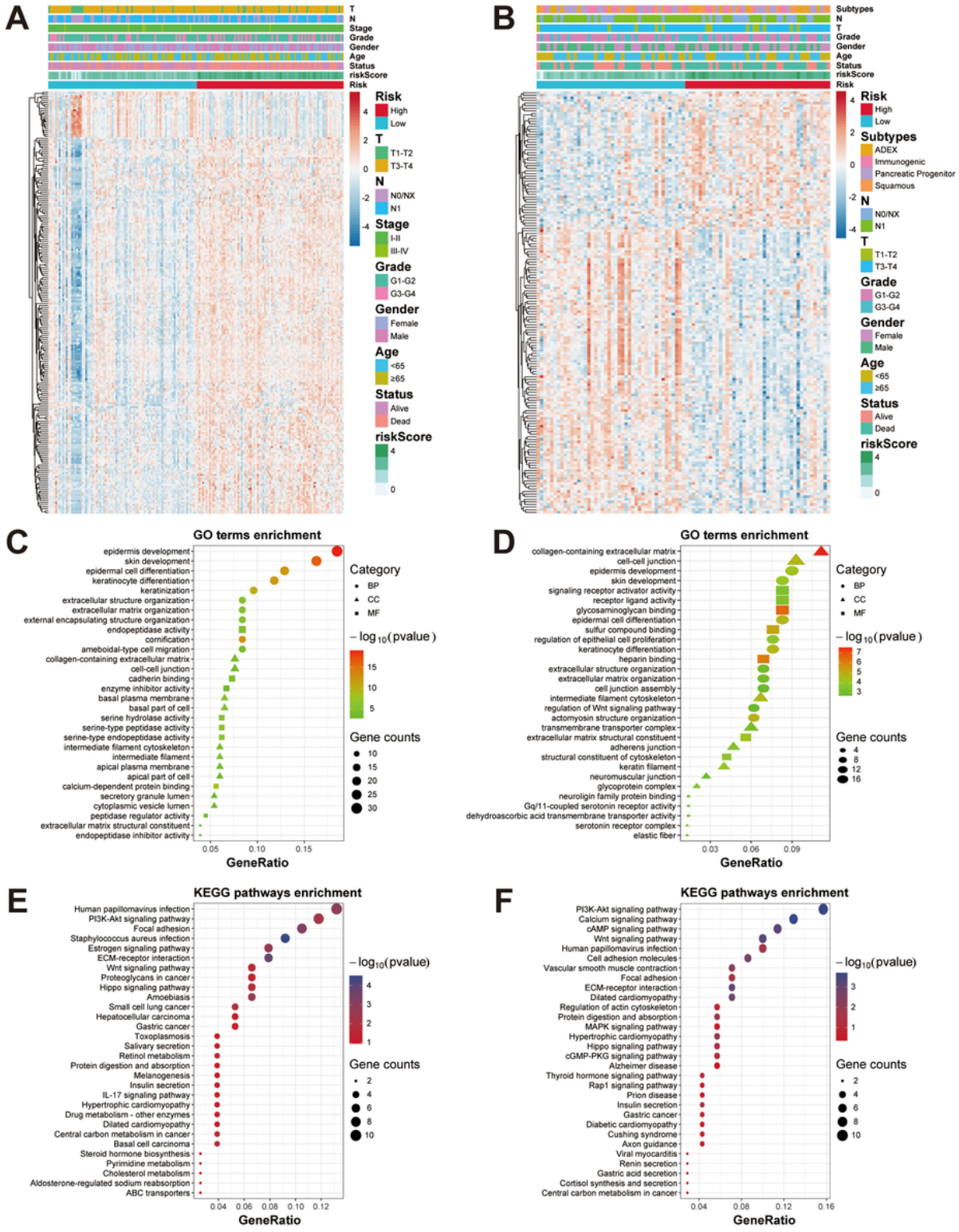

Figure 4

Differential gene expression analysis, GO and KEGG enrichment analysis between two risk groups.

(A and B) Heatmap of the DEGs between the high-risk and low-risk groups in TCGA (A) and ICGC (B) cohorts. **(C–F)** Representative terms of GO enrichment analysis and representative pathways of KEGG

enrichment analysis between the high-risk and low-risk groups in TCGA (C and E) and ICGC (D and F) cohorts.

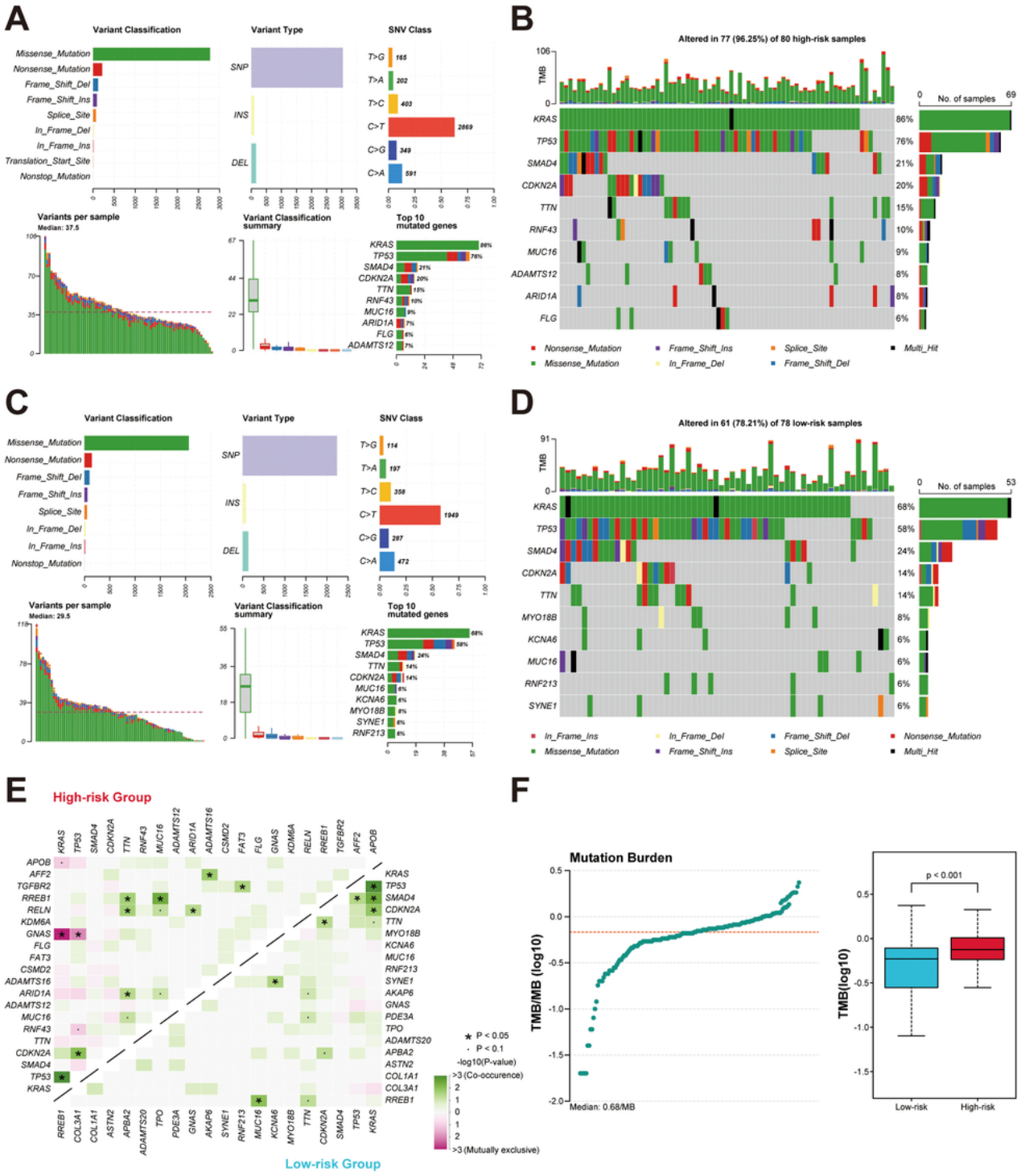

Figure 5

Somatic mutation profiles between two risk groups in TCGA cohort.

(A–D) MAF-summary plots and waterfall charts of somatic mutations in the high-risk group (A and B) and low-risk group (C and D). The top 10 mutated genes were shown.

(E) Correlation heatmaps of co-occurrence and mutually exclusive mutations in the high-risk and low-risk groups. **(F)** Distribution of TMB (left) and comparison between two risk groups (right).

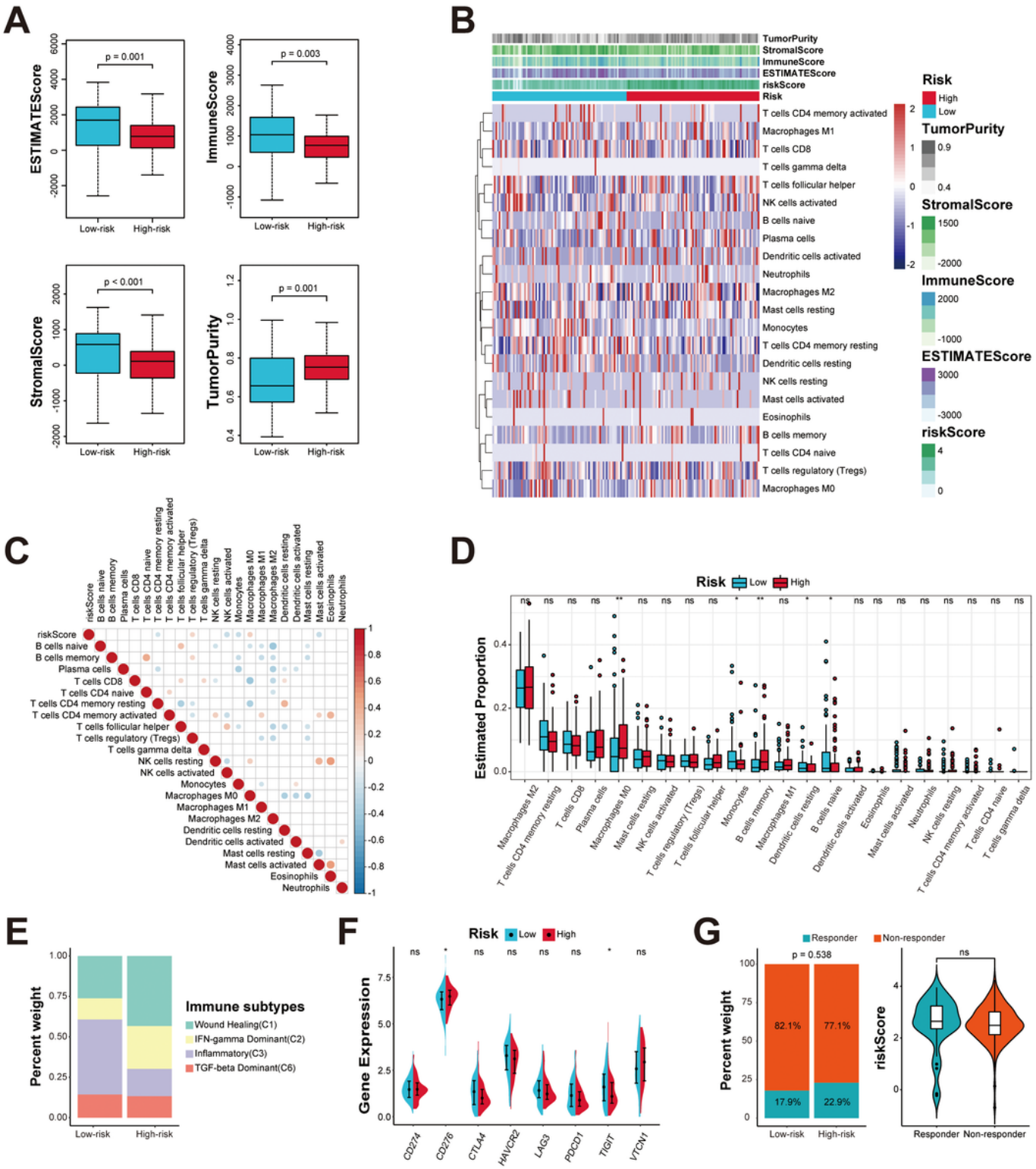

Figure 6

Estimation of immune cell infiltration and prediction of ICB responses in TCGA cohort.

(A) Comparison of estimate score, immune score, stromal score and tumor purity between two risk groups. **(B)** Heatmap displaying the abundances of 22 types immune cells. **(C)** Correlation heatmap of 22 types immune cells and the risk score. **(D)** Comparison of CIBERSORT scores of 22 types immune cells between two risk groups. **(E)** Proportions of four immune subtypes in two risk groups. **(F)** The expression levels of eight immune checkpoints between two risk groups. **(G)** Comparison of ICB response rates between two risk groups and the risk score between responders and non-responders. ns, not significant; *, $P < 0.05$; **, $P < 0.01$.

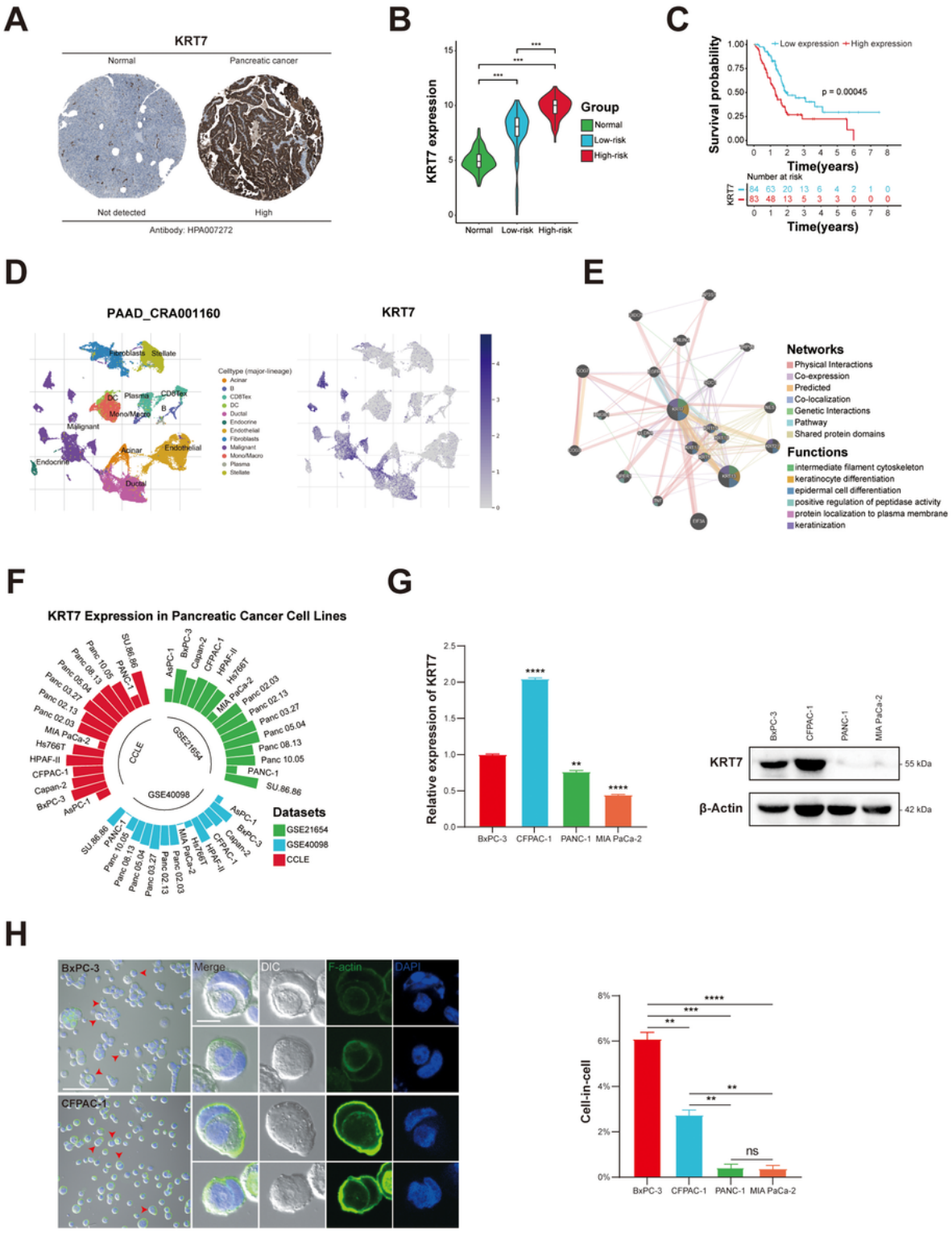

Figure 7

KRT7 was associated with prognosis of patients and CIC formation.

(A) Representative immunochemistry images of KRT7 in normal pancreas tissue and PC tissue from HPA database. **(B)** Comparison of *KRT7* expression between normal samples (GTEx dataset) and tumor samples (TCGA dataset). TCGA patients were stratified into the low-risk and high-risk groups based on

the risk score of individuals. **(C)** The Kaplan-Meier survival analysis based on *KRT7* expression in TCGA dataset. **(D)** The UMAP plots of diverse cell types in PDAC tissues (left) and expression levels of *KRT7* among different cell types (right) based on single-cell transcriptome analysis. **(E)** Related networks and functions of *KRT7*. **(F)** *KRT7* expression levels among 15 PC cell lines in three independent datasets. **(G)** The expression levels of *KRT7* mRNA (left) and protein (right) in 4 PC cell lines. **(H)** Representative immunofluorescent images of typical CIC structures for BxPC-3 and CFPAC-1 cell lines, and the frequencies of CIC formation in 4 PC cell lines. Red arrows indicate CIC structures. Scale bar: 100 μm (left) and 10 μm (right). Data represent means ± SD from three independent experiments. ns, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

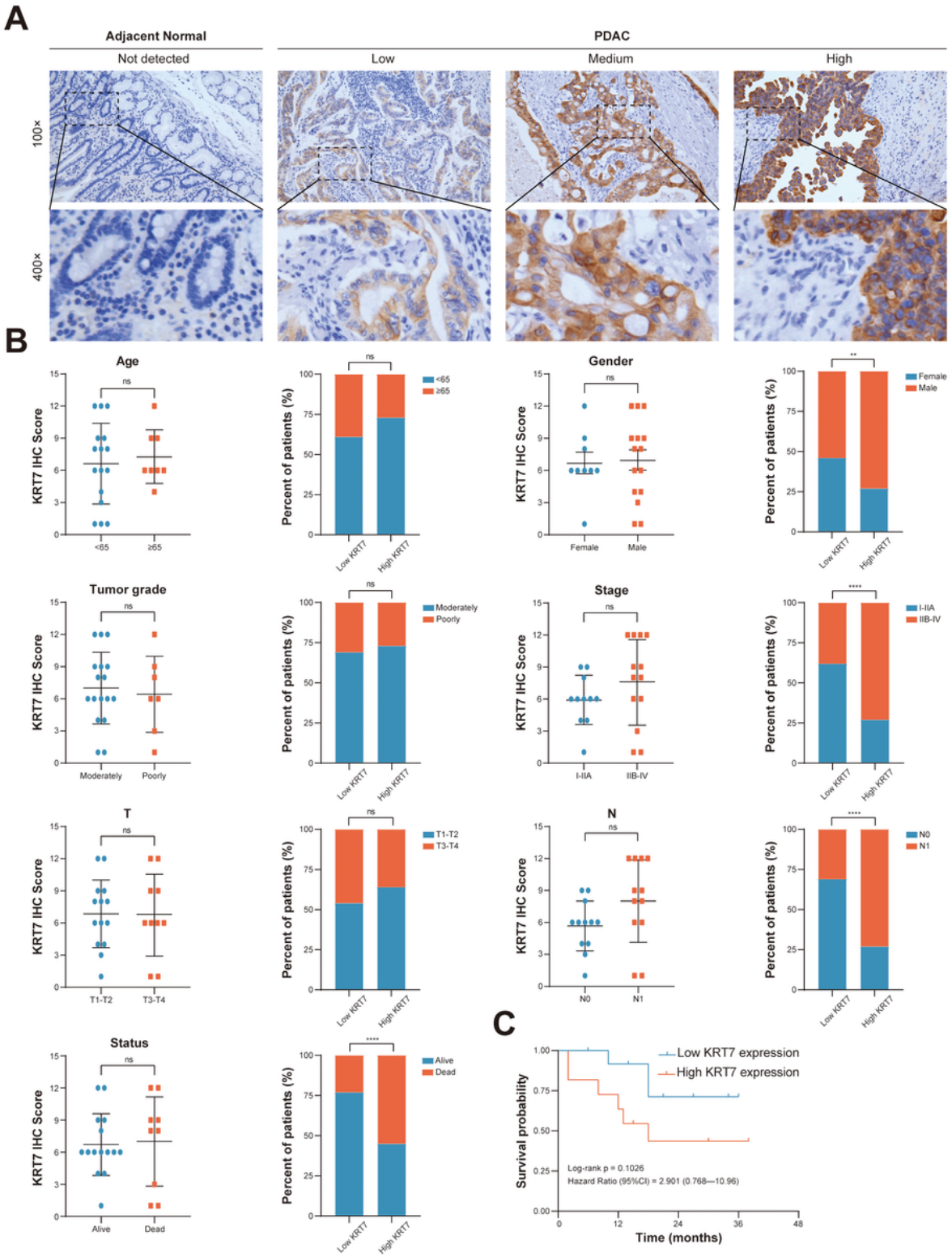

Figure 8

Validation of KRT7 in an independent PDAC cohort.

(A) Representative images of IHC staining of KRT7 in PDAC samples. **(B)** The comparison of KRT7 IHC scores between different clinicopathological subgroups and proportions of clinicopathological characteristics in low and high KRT7 expression groups. Low expression, IHC scores 1–6; high expression,

IHC scores 8–12. **(C)** The Kaplan-Meier survival analysis of 24 PDAC patients by KRT7 expression levels. ns, not significant; **, $P < 0.01$ and ****, $P < 0.0001$ of chi-square test.

Supplementary Files

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

- SupplementaryTable1.docx
- SupplementaryTable2.docx
- SupplementaryFigure1.tif
- SupplementaryFigure2.tif
- SupplementaryFigure3.tif
- SupplementaryFigure4.tif
- SupplementaryFigure5.tif
- SupplementaryFigure6.tif