OXFORD

Human Molecular Genetics, 2022, Vol. 31, 10, 1705–1719
https://doi.org/10.1093/hmg/ddab343
Advance access publication date 20 November 2021
Bioinformatics Article

Development and validation of prognostic and diagnostic model for pancreatic ductal adenocarcinoma based on scRNA-seq and bulk-seq datasets

Kai Chen[1], Xinxin Liu[1], Weikang Liu[1], Feng Wang[2], Xiaodong Tian[1,*] and Yinmo Yang[1,*]

[1]Department of General Surgery, Peking University First Hospital, Beijing 100034, China
[2]Department of Endoscopy Center, Peking University First Hospital, Beijing 100034, China
*To whom correspondence should be addressed at: Department of General Surgery, Peking University First Hospital, 8th Xishiku Street, Beijing 100034, China. Tel: +86 13911636280; Fax: +86 13911636280; Email: tianxiaodong@pkufh.com

Abstract

The 5-year overall survival (OS) of pancreatic ductal adenocarcinoma (PDAC) is only 10%, partly owing to the lack of reliable diagnostic and prognostic biomarkers. The raw gene-cell matrix for single-cell RNA-seq (scRNA-seq) analysis was downloaded from the GSA database. We drew cell atlas for PDAC and normal pancreatic tissues. The inferCNV analysis was used to distinguish tumor cells from normal ductal cells. We identified differential expression genes (DEGs) by comparing tumor cells and normal ductal cells. The common DEGs were used to conduct prognostic and diagnostic model using univariate and multivariate Cox or logistic regression analysis. Four genes, *MET*, *KLK10*, *PSMB9* and *ITGB6*, were utilized to create risk score formula to predict OS and to establish diagnostic model for PDAC. Finally, we drew an easy-to-use nomogram to predict 2-year and 3-year OSs. In conclusion, we developed and validated the prognostic and diagnostic model for PDAC based on scRNA-seq and bulk-seq datasets.

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is a dismal disease, and the prognosis of patients with PDAC has not been significantly improved recently with 5-year overall survival (OS) of 9–10% (1). The poor prognosis is mainly attributed to low surgical resection rate, chemoradiotherapy resistance and lack of reliable biomarkers for early diagnosis. Most patients have vascular invasion and/or distant metastasis at the time of diagnosis, missing the possibility of radical resection (2,3). In addition, half of patients with PDAC would have tumor recurrence or distant metastasis 2 years after radical resection, with 5-year OS of 25–30% (4). Early diagnosis and radical resection can significantly improve the prognosis of patients, but current serum tumor markers, such as carbohydrate antigen 19-9 (CA19-9) and carcinoembryonic antigen (CEA), have limited specificity and sensitivity to screen patients with early PDAC (5). In addition, accurate prognosis evaluation could provide appropriate clinical decision support for patients. Therefore, it is vital to develop valid prognostic and diagnostic models for PDAC.

Compared with traditional bulk-seq, single-cell RNA-seq (scRNA-seq) could acquire transcriptome data of each cell at unprecedented resolution (6). Recent studies revealed complex intra-tumor heterogeneity in PDAC microenvironment and identified various new cell subpopulations using scRNA-seq. Peng *et al.* (7) conducted scRNA-seq for 24 PDAC and 11 normal pancreatic specimens and found two ductal cell subpopulations in PDAC with different malignancy and cell markers, indicating ductal cell heterogeneity in PDAC. Elyada and colleagues (8) drew cell atlas of human and mouse pancreas and identified three types of cancer-associated fibroblasts (CAFs): myofibroblastic CAFs, inflammatory CAFs and antigen-presenting CAFs, suggesting intricate CAFs' heterogeneity. However, gene expression characteristics of tumor cells in PDAC remain to be further investigated.

The development of bioinformatics and multiomics database makes it easier to explore the expression pattern of malignancies and to construct prognostic and diagnostic models. Various risk score formulas have been proposed to predict the occurrence and prognosis of PDAC based on differential expression genes (DEGs) from bulk-seq datasets, such as The Cancer Genome Atlas (TCGA) or Gene Expression Omnibus (GEO) (9–12). However, bulk-seq only indicates average expression level of the whole tissue, which might lead to the bias for individual tumor cell. In contrast, scRNA-seq could reveal DEGs between tumor cells and normal ductal cells without the interference of stromal and immune cells in pancreatic tissues.

In this study, we aimed to establish and validate prognostic and diagnostic model for PDAC based on scRNA-seq and bulk-seq datasets. We drew the cell atlas of PDAC and normal pancreas using scRNA-seq dataset,

Received: September 24, 2021. **Revised:** November 16, 2021. **Accepted:** November 17, 2021

distinguished tumor cells from normal ductal cells and identified DEGs between them. Next, we combined DEGs from scRNA-seq and DEGs from TCGA and GTEx to get common DEGs and further selected DEGs by univariate Cox proportional hazard regression (Unicox) analysis and LASSO-penalized Cox regression analysis. Then, we performed multivariate Cox proportional hazard regression (Multicox) analysis to construct prognostic model in the train set of TCGA_PAAD. Finally, we conducted internal and external validations for this prognostic model using the validation set of TCGA_PAAD, PACA_AU and GEO datasets, and we developed nomogram to predict 2-year or 3-year OS for PDAC. In addition, we constructed diagnostic model for prognosis related DEGs based on univariate and multivariate logistic regressions.

Results

scRNA-seq delineated cell atlas for PDAC and normal pancreas

To uncover cellular components of tumor microenvironment in PDAC and to discover gene expression profile difference between tumor cells and normal ductal cells, we downloaded single-cell transcriptome sequencing dataset from Genome Sequence Archive (GSA). A total of 24 PDAC (38 201 cells) and 11 normal pancreatic (14 838 cells) specimens were included to construct gene-cell expression matrix (Supplementary Material, Table S1). After cells filtering, normalization, principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction, 26 original clusters were identified (Fig. 2A). According to signature genes of each cell type reported previously (7,8,13), these clusters were classified into nine known cell types, including stellate cells, macrophages, endothelial cells, ductal cells, T cells, fibroblast cells, B cells, acinar cells and endocrine cells (Fig. 2B and C and Supplementary Material, Fig. S1A–C). Interestingly, cluster 25 (unknown) expressed signature genes of macrophages (AIF1 and CD68), stellate cells (RGS5) and fibroblast cells (COL1A1). Next, we identified marker genes of each cell type using FindMarkers algorithm (Fig. 2D and Supplementary Material, Table S2).

We counted the proportion of various cell subpopulations in normal pancreatic tissue (N1–N11) and PDAC (T1–T24) and found that the cellular components varied among distinct specimens, indicating inter-patient heterogeneity (Fig. 3A and B). In particular, PDAC had higher proportion of stellate cells, fibroblast cells and immune cells (B cells, T cells and macrophages) compared with normal pancreatic tissue, which was in line with the abundant extracellular matrix and tumor infiltrating immune cells—hallmark of PDAC (Fig. 3C). Furthermore, we inferred cell-cycle state of cells among different specimens using Seurat package. There were marked difference among PDAC and normal pancreatic specimens for cell-cycle state. The PDAC had higher proportion of G2M phase, suggesting active proliferation ability (Fig. 3D–F).

Copy number alteration analysis distinguished tumor cells from normal ductal cells

Next, we isolated ductal cells to construct gene-cell expression matrix and performed t-SNE analysis. A total of 18 distinct ductal subpopulations were identified (Fig. 3G and Supplementary Material, Fig. S1E and F). We inferred somatic large-scale chromosomal CNVs and calculated CNV scores based on a set of reference normal cells (ductal cells in normal pancreatic specimens, endothelial cells, stellate cells and macrophages) through inferCNV package. The results showed that cluster 4/5/6/8/14/16/17 exhibited significantly higher CNV compared with reference cells and were therefore identified as tumor cells (Supplementary Material, Fig. S1G and H). The original ductal subclusters were classified into Ductals 1–3 and Tumors 1–5 (Fig. 3H). Ductal 1 and Ductal 3 were derived from normal pancreatic specimens and PDAC, respectively, whereas Ductal 2 was derived from both PDAC and normal pancreatic specimens. Inter-patient heterogeneity was detected in tumor cells again, with almost all patients represented in distinct tumor subpopulations. All tumor subpopulations (Tumors 1–5) had higher proportion of G2M phase than Ductal 1, indicating proliferation potential of tumor cells (Supplementary Material, Fig. S1D). Ductal 1 had a high expression of FXYD2, which encodes the sodium-/potassium-transporting ATPase subunit gamma. FXYD2 is expressed in normal pancreatic ductal cells, which is consistent with our finding (7,8). Multiple signature genes for PDAC were detected in most of tumor subpopulations except for Tumor 2, such as CEACAM1, CEACAM5, CEACAM6, TFF1 and TFF2 ($P < 0.001$) (15). All tumor subpopulations expressed higher levels of LAMC2 and MSLN ($P < 0.0001$) (Fig. 3I). In fact, LAMC2 was proposed as a diagnostic biomarker for PDAC (14,15).

Epithelial mesenchymal transition and cancer stem cell properties of tumor subpopulations

In addition, Cluster 4, Cluster 5, Cluster 6, Cluster 8 and Cluster 14 had significantly higher CNV compared with reference cells, thus, they are named Tumors 1–5, respectively (Supplementary Material, Fig. S1G and H). We calculated the CNV scores of Tumors 1–5 again and found that they had higher CNV compared with reference cells, confirming their malignant cell identity (Fig. 4A and B). Based on epithelial mesenchymal transition (EMT) and stem cell markers, we calculated E score, M score and S score. The results indicated that Tumor 2 had the highest M score among ductal subpopulations (Fig. 4C and D and Supplementary Material, Table S3 and Supplementary Material, Fig. S1I). Tumor 2 had a high expression of mesenchymal cell markers, including ACTA2, COL1A1, COL1A2, COL3A1, FN1, MMP2 and MMP7 (Fig. 4E). GO analysis showed that marker genes of Tumor 2 were significantly enriched for extracellular structure organization and constituent (Supplementary Material, Table S4 and Supplementary Material, Fig. S1J and K).

Table 1. Clinicopathological characteristics of patient cohorts for diagnostic and prognostic models

	TCGA_PAAD (*n* = 177)	**PACA_AU (*n* = 91)**	**GSE57495 (*n* = 63)**	**GSE71729 (*n* = 357)**	**GSE62452 (*n* = 130)**	**GSE15471 (*n* = 78)**	**GSE16515 (*n* = 52)**
Platform	HTSeq (RNA-seq)	HiSeq (RNA-seq)	GPL15048 (gene chip)	GPL20769 (gene chip)	GPL6244 (gene chip)	GPL570 (gene chip)	GPL570 (gene chip)
Gender							
Male	97	47	NA	NA	NA	NA	34
Female	80	43	NA	NA	NA	NA	18
Unknown	0	1	NA	NA	NA	NA	0
Age (years)							
Median (range)	65 (36–89)	67 (36–86)	NA	NA	NA	NA	68.5 (49–84)
Histological type							
PDAC	164	72	63	145	69	39	36
IPMN	NA	7	0	0	0	0	0
Neuroendocrine carcinoma	6	0	0	0	0	0	0
Others	7	12	0	212	61	39	16
Location							
Head	129	NA	NA	NA	NA	NA	NA
Body	15	NA	NA	NA	NA	NA	NA
Tail	14	NA	NA	NA	NA	NA	NA
Others	19	NA	NA	NA	NA	NA	NA
T stage							
T1	7	NA	NA	NA	NA	NA	NA
T2	24	NA	NA	NA	NA	NA	NA
T3	141	NA	NA	NA	NA	NA	NA
T4	3	NA	NA	NA	NA	NA	NA
Others	2	NA	NA	NA	NA	NA	NA
N stage							
N0	49	NA	NA	NA	NA	NA	NA
N1	119	NA	NA	NA	NA	NA	NA
Others	9	NA	NA	NA	NA	NA	NA
AJCC stage							
I	21	NA	13	NA	7	NA	NA
IIA	28	NA	17	NA	18	NA	NA
IIB	118	NA	33	NA	66	NA	NA
III	3	NA	0	NA	26	NA	NA
IV	4	NA	0	NA	13	NA	NA
Others	3	NA	0	NA	0	NA	NA
Margin status							
R0	83	NA	NA	NA	NA	NA	NA
R1	41	NA	NA	NA	NA	NA	NA
Others	53	NA	NA	NA	NA	NA	NA

Construction and internal validation of prognostic model based on TCGA database

To better understand gene expression pattern of tumor cells in PDAC, we compared Ductal 1 with Tumors 1–5, and found 604 DEGs using FindMarkers algorithm (Fig. 5A). On the other hand, we found 2615 DEGs between PDAC (TCGA) and normal pancreatic tissue (matched GTEx). Then, 222 common DEGs were used to construct gene expression matrix accompanied by clinical follow-up data using TCGA_PAAD dataset (Fig. 5B and Table 1). GO and GSEA analyses suggested that DEGs were significantly related to immune response, antigen processing and presentation and EMT (Supplementary Material, Fig. S2A–F). We identified 68 genes related to OS by univariate regression analysis (Supplementary Material, Table S5). To avoid overfitting during the prognostic model construction, we performed LASSO-penalized Cox regression analysis and finally selected 10 genes from previous OS-related genes (Fig. 5C and D).

Next, 177 subjects in TCGA_PAAD were randomly divided into train set and validation set in 2:1. The Multicox analysis was employed to construct final prognostic model in train set. The risk score for each subject was calculated as follows: risk score $(t) = h0\ (t)\ *\ \exp(MET\ *\ 0.3839 + ITGB6\ *\ 0.1881 + PSMB9\ *0.3586 + KLK10\ *0.0838)$. Subjects were divided into low-risk group and high-risk groups according to the median cutoff value. The Kaplan–Meier curve (KM) showed that subjects in the high-risk group had significantly shorter OS than those in the low-risk group ($P = 4.71e - 07$) (Fig. 5E). Time-dependent receiver operating characteristic curve (ROC) was used to evaluate the accuracy of predicting 1-year, 1.5-year and 2-year OSs, and the area under curve (AUC)

values were 0.777, 0.74 and 0.738, respectively (Fig. 5F). The higher-risk score indicated the worse prognosis (Fig. 5G and H). Then, we conducted internal validation for prognostic model. Consistent with previous results in train set, subjects with high-risk score had worse prognosis in the validation set ($P = 2.832e-03$, 1-/1.5-/2-year AUC: 0.673/0.731/0.806) and all sets ($P = 9.274e-09$, 1-/1.5-/2-year AUC: 0.749/0.738/0.741) (Fig. 5I–L and Supplementary Material, Fig. S2G–J).

External validation of prognostic model

To further evaluate the reliability of prognostic model, we downloaded the gene expression matrix and clinical follow-up data of PDAC from ICGA and GEO databases. Subjects in the high-risk group had significantly shorter OS than those in the low-risk group based on four prognosis-related signatures in all external validation sets for PACA_AU ($P = 3.688e-02$), GSE57495 ($P = 1.919e-03$) and GSE71729 ($P = 2.514e-03$) (Fig. 6A–C). Their 1-/1.5-/2-year AUC values were 0.714/0.629/0.532, 0.742/0.833/0.856 and 0.665/0.69/NA.

Nomogram for predicting the survival for PDAC patients

We downloaded complete clinicopathological characteristics of subjects in TCGA_PAAD and performed Unicox and Multicox analyses. The Unicox results showed that *N* stage [hazard ratio (HR): 1.95, 95% confidence interval (CI): 1.03–3.67, $P = 0.0398$], margin status (HR: 1.72, 95% CI: 1.02–2.92, $P = 0.0431$) and risk score (HR: 1.97, 95% CI: 1.42–2.75, $P < 0.0001$) were significantly correlated to the OS of subjects (Fig. 6D). Moreover, *N* stage (HR: 1.78, 95% CI: 1.01–3.15, $P = 0.0486$), margin status (HR: 1.73, 95% CI: 1.06–2.83, $P = 0.0281$) and risk score (HR: 1.86, 95% CI: 1.37–2.53, $P < 0.0001$) were also independent prognostic factors for PDAC (Fig. 6E).

Furthermore, we established an easy-to-use and clinically adaptable prognostic nomogram. The subject with higher total points was associated with worse 2-year and 3-year OSs (Fig. 6F). The calibration curve showed good correlation between nomogram-predicted OS and actual OS, indicating the accuracy of this nomogram (Fig. 6G).

The performance of prognostic model in different age and N stage subgroups

To further evaluate the accuracy and reliability of prognostic model in different ages and *N* stages, we classified subjects in TCGA_PAAD into subgroups (age <= 65 and age > 65; *N*0 and *N*1). For age subgroups, subjects with high-risk score had significantly shorter OS compared with those with low-risk score (age <= 65: $P = 4.226e-06$, 1-/1.5-/2-year AUC: 0.76/0.785/0.823; age > 65: $P = 1.869e-03$, 1-/1.5-/2-year AUC: 0.738/0.683/0.636) (Supplementary Material, Fig. S2K–N). For *N* stage subgroups, we got similar results (*N*0: $P = 3.582e-04$, 1-/1.5-/2-year AUC: 0.729/0.806/0.87; *N*1: $P = 2.56e-03$, 1-/1.5-/2-year AUC: 0.738/0.677/0.649) (Supplementary Material, Fig. S2O–R).

Construction of diagnostic model based on GEO database

In order to evaluate the diagnostic value of four prognosis-related signatures in PDAC, we compared the gene expressions of PDAC and normal pancreatic tissue in GES62452. The primary PDAC had significantly higher expression level of *MET*, *KLK10*, *PSMB9* and *ITGB6* than the normal pancreatic tissue (Fig. 7A). Univariate analysis showed that the high expression of *MET*, *KLK10*, *PSMB9* and *ITGB6* was significantly correlated with PDAC (Fig. 7B). However, only *PSMB9* (OR: 4.26, 95% CI: 1.56–13.12, *P*-value = 7.03e − 03), *ITGB6* (OR: 1.92, 95% CI: 1.37–2.81, *P*-value = 3.28e − 04) and *KLK10* (OR: 7.99, 95% CI: 2.24–33.33, *P*-value = 2.323e − 03) were independent diagnostic factors for PDAC (Fig. 7C). The equation of diagnostic model was $\text{logitP}\ (Y = 1) = -24.05 + (PSMB9 * 1.4493 + ITGB6 * 0.4263 + KLK10 * 1.9274)$.

Similarly, we established diagnostic nomogram to visualize the results of multivariate logistic regression. The subjects with higher total points had higher incidence of PDAC (Fig. 7D). The calibration curve showed great agreement between nomogram-predicted and actual PDAC probability, and C-index was 0.873 (Fig. 7E).

External validation of diagnostic model

We retrieved gene expression matrix from GSE71729, GSE15471 and GSE16515 as an external validation set for the diagnostic model. Consistent with the previous result, primary PDAC had a significantly higher expression of *MET*, *KLK10*, *PSMB9* and *ITGB6* compared with the normal pancreatic tissue in GSE15471 and GSE16515 (Supplementary Material, Fig. S3B and C). In addition, both primary and metastatic PDAC had significantly higher expression of *MET*, *KLK10* and *ITGB6* in GSE71729 (Supplementary Material, Fig. S3A). The diagnostic model performed well in external validation set (GSE71729: C-index = 0.898; GSE15471: C-index = 0.948; GSE16515: C-index = 0.938) (Supplementary Material, Fig. S3D–F).

Model validation using patient cohort from our department

To further validate the diagnostic and prognostic models, we examined the expression levels of *MET*, *KLK10*, *PSMB9* and *ITGB6 in vitro*. Compared with normal ductal cell (HPNE), PDAC cells expressed higher levels of *MET*, *KLK10*, *PSMB9* and *ITGB6*, which further verified tumor cells are abnormally enriched for genes mentioned before (Fig. 8A–D). Subsequently, we retrieved PDAC specimens from our department and found that PDAC had a significantly higher expression of *MET*, *KLK10*, *PSMB9* and *ITGB6* than matched adjacent normal pancreatic tissue (Fig. 8E–H). Moreover, we found that the established diagnostic model could be validated with our data, and C-index was 0.777 (Fig. 8I).

Next, we employed KM curve to evaluate the correlation between OS and relapse-free survival (RFS) and

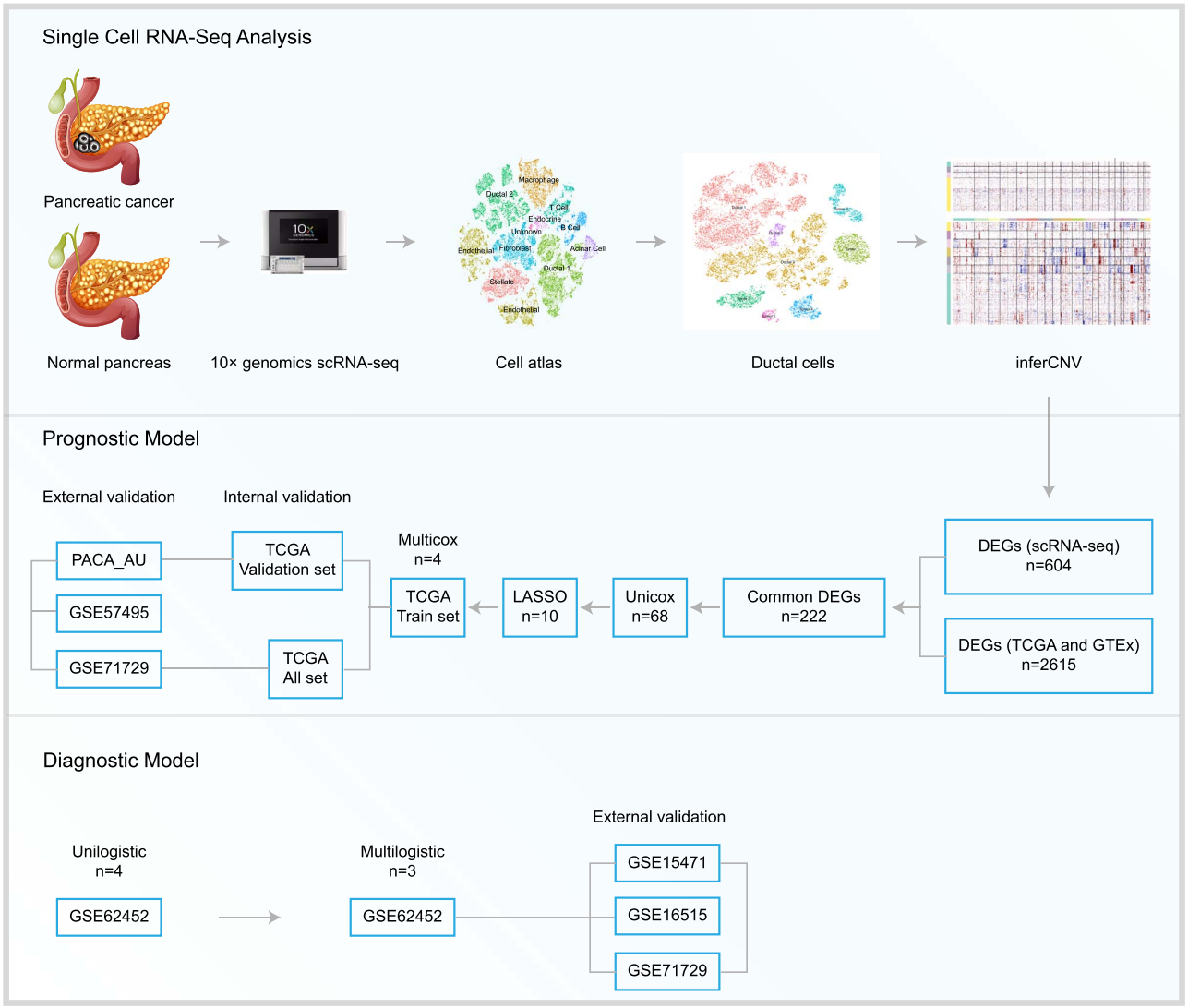

Figure 1. Graphical scheme describing the study design. We first delineated cell atlas of PDAC and normal pancreas using scRNA-seq datasets, then distinguished tumor cells from normal ductal cells by inferCNV analysis. The common DEGs of scRNA-seq and TCGA versus GTEx analyses were used to construct prognostic and diagnostic models. We also conducted intern and external validations for them.

four prognosis-related signatures. Subjects with higher expression of *MET*, *KLK10* and *ITGB6* had significantly worse OS and RFS ($P < 0.05$). Subjects with higher expression of *PSMB9* had shorter OS and RFS, but there was no statistical significance (Supplementary Material, Fig. S4A and B). In addition, we also compared the expression of them in the protein level between PDAC and normal pancreatic tissue by IHC using the Human Protein Atlas (HPA) database. The results suggested that PDAC had significantly enriched expression of *MET*, *PSMB9* and *ITGB6* (Supplementary Material, Fig. S4C).

Discussion

PDAC is characterized by intra-tumor and inter-patient heterogeneities, which brings huge challenge for effective treatment of PDAC. Recently, a growing body of researches based on scRNA-seq confirmed inter-patient heterogeneity, indicating that tumor cells from different patients had distinct gene expression profiles and malignant behavior (16–18). Consistent with previous findings, we identified multiple tumor cell subpopulations (Tumors 1–5) belonging to different patients, respectively, by scRNA-seq analysis, and these subpopulations had different CNV profiles and EMT scores. Moreover, Tumor 2 had the highest M score and higher expression of mesenchymal cell markers, suggesting EMT of tumor cells. Tumor 2 subpopulation was derived from patient T9 (moderately poorly differentiated, 36 years old, CA19-9: 11.2, vascular invasion and perineural invasion). Previous studies demonstrated EMT was significantly related to tumorigenesis, chemoresistance and metastasis (19–21). Thus, we speculated that the EMT might play a pivotal role in tumorigenesis for patient T9 and cause highly aggressive behavior.

Besides clinicopathological characteristics, risk score based on gene expression pattern could be the prognostic and diagnostic signatures for malignancies. With

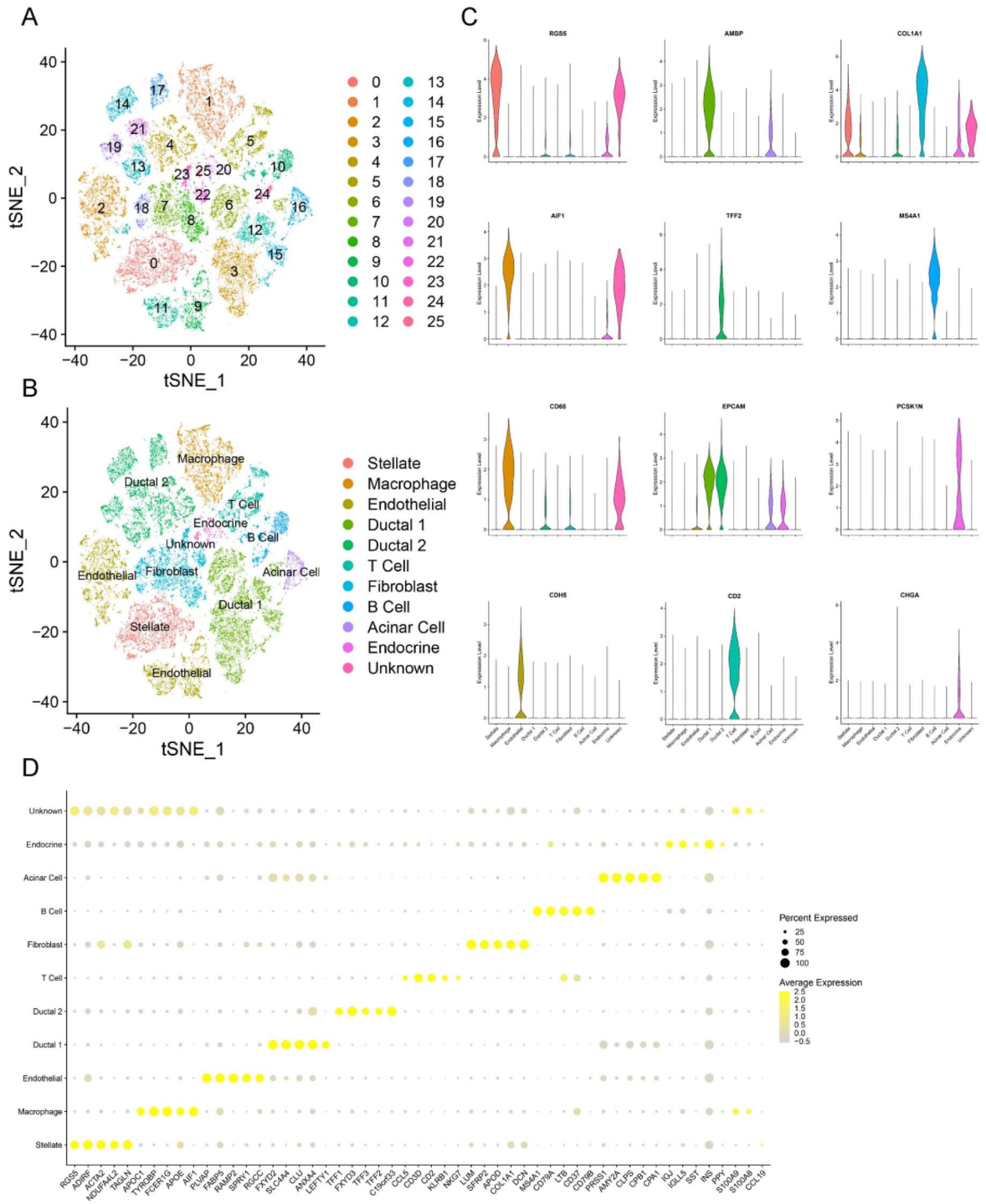

Figure 2. scRNA-seq delineates cell atlas of pancreas. (**A**, **B**) The t-SNE plot showing the original cluster (A) and named cell subpopulations (B). (**C**) Violin plots showing the expression level of known cell-type-specific markers to demonstrate the identity of each cluster. (**D**) Bubble plot showing the Top5 marker genes across all clusters. Size of dots represents the proportion of cells expressing a particular marker, and intensity of color indicates the average expression level.

the development of sequencing technology, distinct molecular subtypes related to diagnosis and prognosis of PDAC were reported (22–26). In this study, we performed CNV analysis to identify tumor cells from all ductal cells and compared the gene expression profile between them. The composition of PDAC was so complicated with 70% stromal constitutes that the identification of many oncogenic drivers and prognosis-related signatures were largely ignored owing to the limitation of traditional bulk-seq. However, the scRNA-seq approach could directly compare gene expression profile between tumor cells and normal cells with single-cell resolution (6,27–29). Therefore, DEGs from scRNA-seq are more reliable to further identify

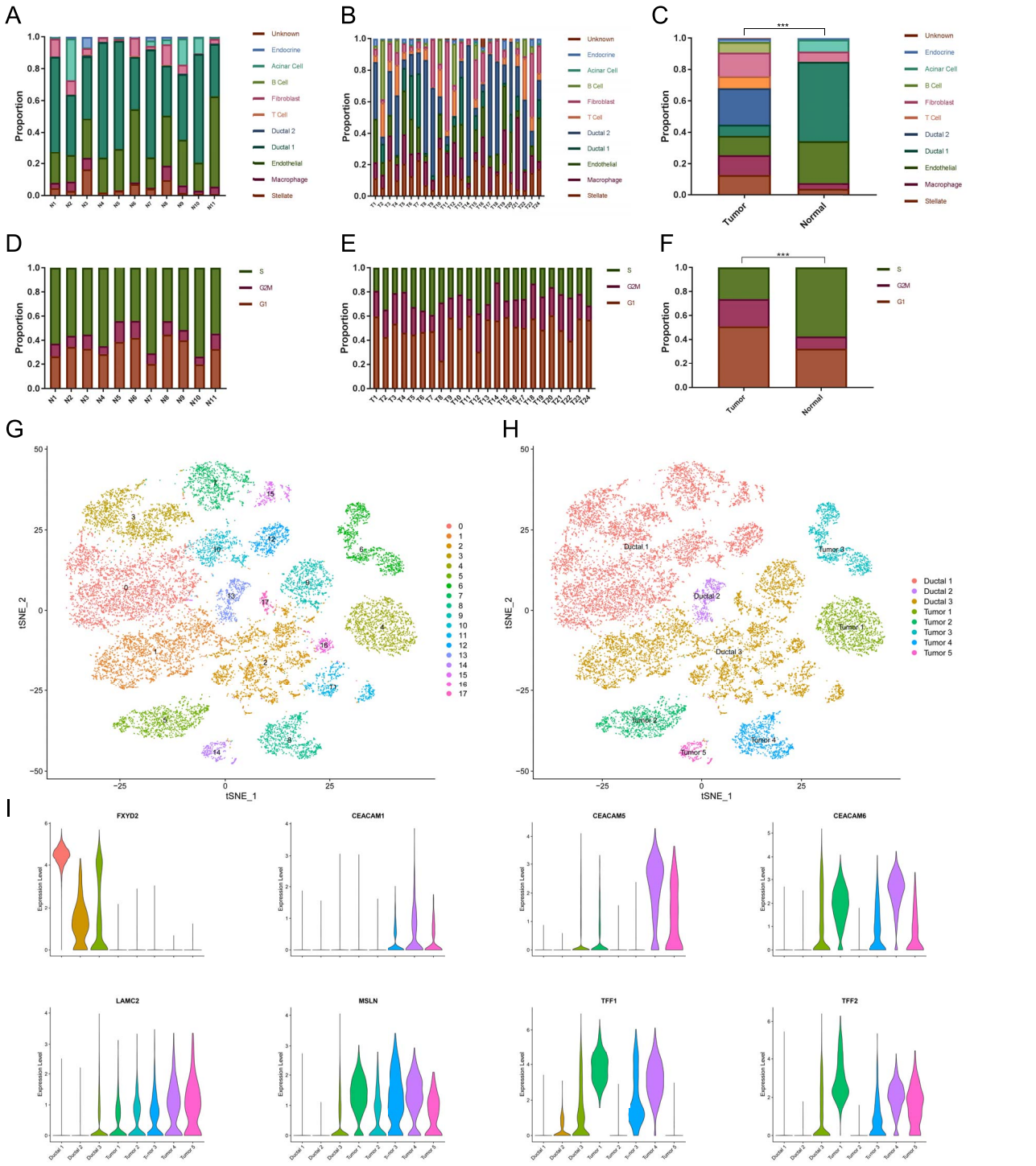

Figure 3. The different cellular constituents and cell-cycle status between PDAC and normal pancreatic specimens. (**A**–**C**) Proportion of various cell subpopulations among normal pancreatic specimens (A), PDAC specimens (B) and PDAC versus normal pancreatic specimens (C). (**D**–**F**) Proportion of G1/S/G2M phase among normal pancreatic specimens (D), PDAC specimens (E) and PDAC versus normal pancreatic specimens (F). (**G**, **H**) Subclustering of the ductal cell subpopulations for original clusters (G) and named ductal cell subpopulations (H). (**I**) Violin plots showing the expression level of selected ductal cell type markers among ductal cell subpopulations.

prognosis-related signatures when compared with bulk-seq.

Reliable prognostic and diagnostic models must be validated by various datasets. Multiple prognostic models have been developed for PDAC based on TCGA or International Cancer Genome Consortium (ICGC) database. Nevertheless, some of them did not conduct the necessary external validation of developed models, reducing the reliability of models (9–11,30). Some of them incorporated all patients with different pathologic types of pancreatic cancer, including PDAC, intraductal papillary mucinous neoplasm (IPMN), neuroendocrine carcinoma

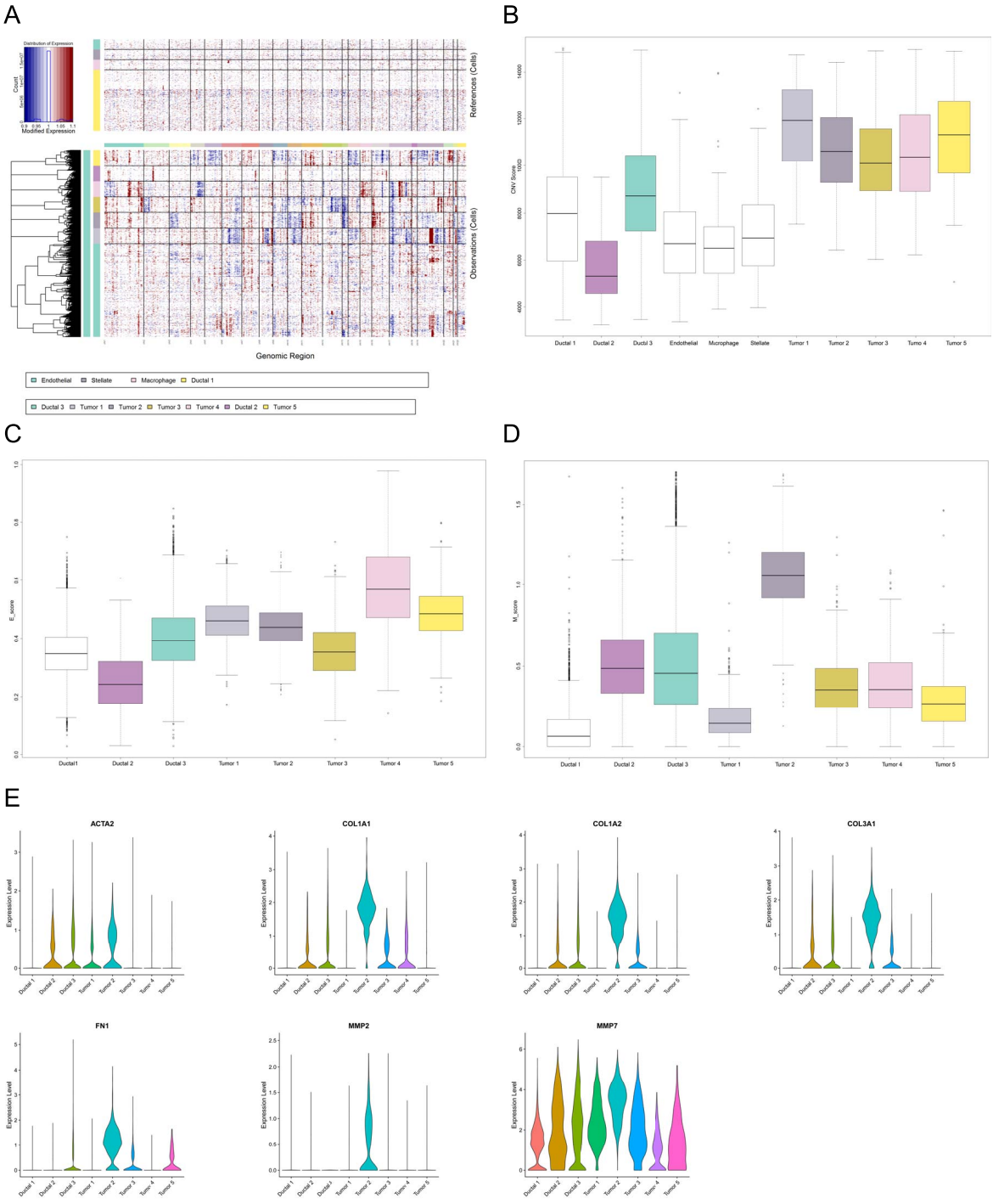

Figure 4. The CNV profile analysis distinguishes tumor cells. (**A**) Heatmap showing large-scale CNV profile of each ducal cell and reference cell subpopulation; the red and blue colors represent high and low CNV level, respectively. (**B**) Boxplot showing the CNV score of each subpopulation; white boxes represent reference cells. (**C**, **D**) Boxplot showing the E score and M score of each ductal subpopulation. (**E**) Violin plot showing the expression level of mesenchymal cell markers among ductal subpopulations.

or others (12,31,32). Patients with different pathologic subtypes who had totally different prognosis should not be mixed together. In our study, we only incorporated patients with PDAC into model construction and validation. The whole TCGA_PAAD dataset was divided into train set and validation set randomly to avoid potential bias. We constructed prognostic model in train set, then performed both internal validation and external validation. Furthermore, we developed an easy-to-use prognostic nomogram combining clinicopathological characteristics and risk score to predict 2-year and 3-year OSs. In addition, we also constructed diagnostic model for

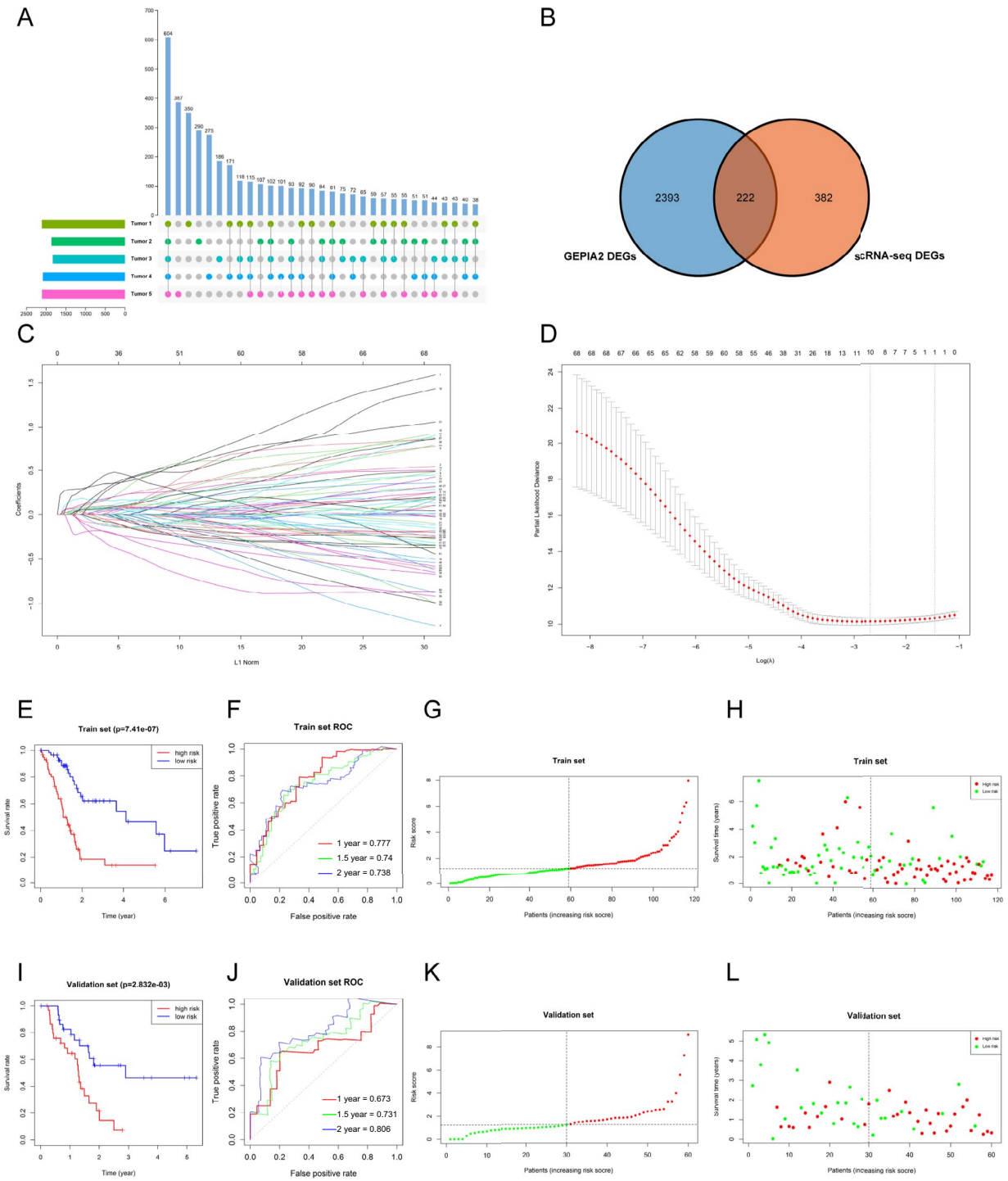

Figure 5. Construction and validation of prognostic model in TCGA_PAAD dataset. (**A**) DEGs between tumor cell and normal ductal cell subpopulations are shown in Upsetplot. (**B**) The overlapping area showing the common DEGs of scRNA-seq and GEPIA2 analyses in Vennplot. (**C**, **D**) Variable selection using LASSO regression, the correlation between coefficients and the number of variable (C), and the first dashed line showing the cutoff value we selected, indicating minimal deviance (D). (**E**–**H**) Construction of prognostic model in train set in TCGA_PAAD, KM curve showing different OSs between high and low-risk group (E), ROC curve was used to evaluate the accuracy of prognostic model for 1-/1.5-/2-year OS (F), risk score distribution of subjects in train set (G) and survival status scatter plot (H). (**I**–**L**) Internal validation of prognostic model in validation set in TCGA_PAAD.

prognosis-related genes by univariate and multivariate logistic regression analysis, and conducted external validation using GSE71729, GSE15471 and GSE16515 datasets and our data.

Overall, we developed the relatively reliable prognostic and diagnostic models for PDAC. However, there are several limitations in this study: (1) These models were constructed based on bulk-seq datasets instead of scRNA-seq dataset because of limited sample size, although we utilized scRNA-seq analysis to identify DEGs for model construction. With the popularity of scRNA-seq technology and decrease of sequencing cost,

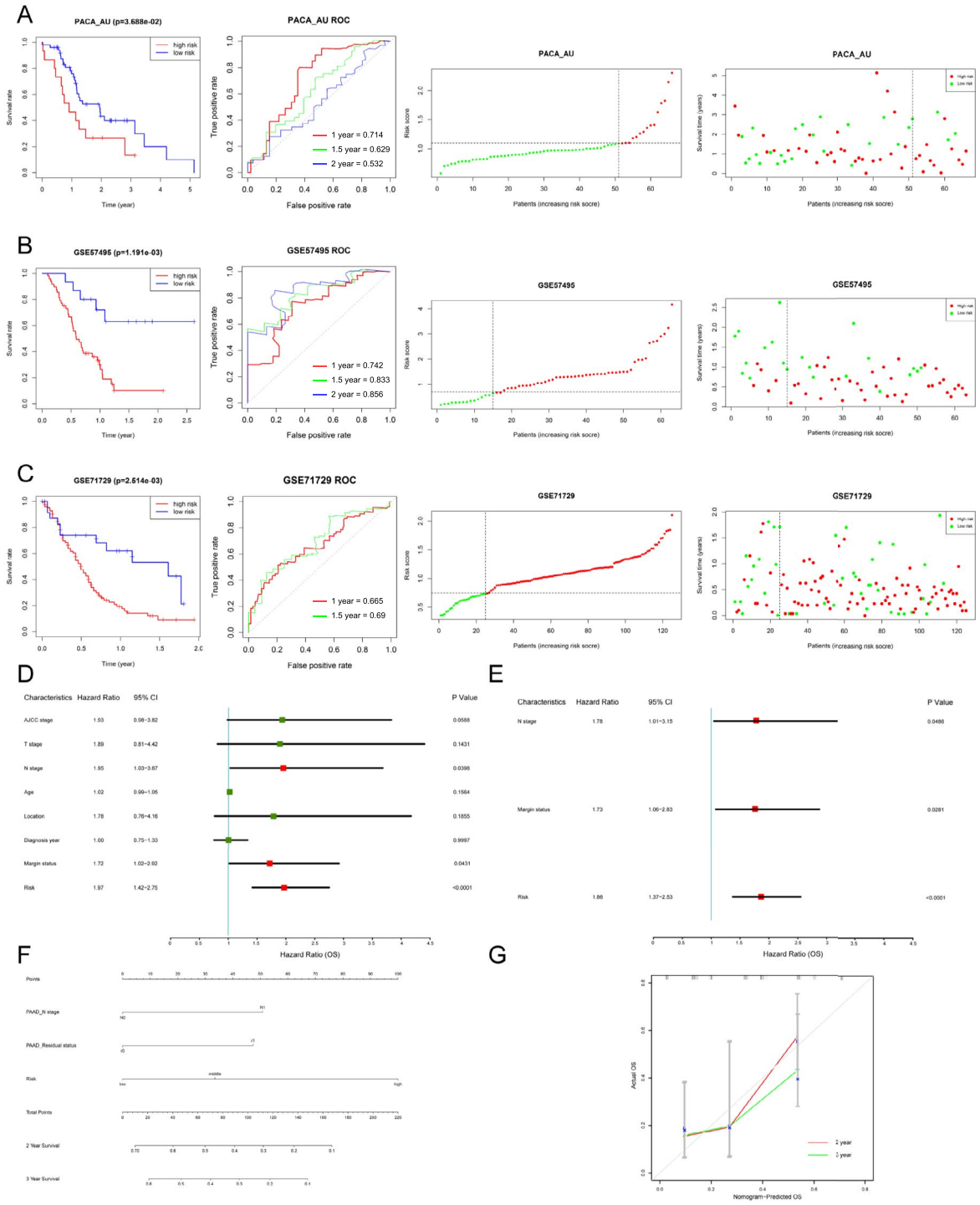

Figure 6. Construction of nomogram for predicting OS in PDAC. (**A**–**C**) External validation of prognostic model in PACA_AU, GSE57495 and GSE71729. (**D**, **E**) Unicox and Multicox analyses were performed to find the risk factors of OS in PDAC; red boxes represent $P < 0.05$ in the forestplot. (**F**) The prognosis-nomogram was drawn to predict 2-year and 3-year OSs for PDAC. (**G**) Calibration curve showing the agreement between actual and nomogram-predicted OS; the gray diagonal line is reference line.

large-scale scRNA for patients with PDAC are required to identify novel tumor subtypes and construct precise prognostic and diagnostic model. (2) Both bulk-seq and scRNA-seq could be completed only when resection specimens are retrieved after operation. These kinds of prognostic models are eligible for improving the strategies of adjunctive therapy; patients with high risk might need to receive other targeted therapy or immunotherapy other than regular chemotherapy. It is better to conduct a prognostic model according to signatures that could be tested before operation indeed. Liquid biopsy, such as circulating tumor cells (CTCs) and exosomal miRNAs could be the most promising alternatives (33–35). (3) The functions of prognosis-related genes in PDAC, including *MET*, *KLK10*, *PSMB9* and *ITGB6*, remain to be elucidated. (4) LASSO

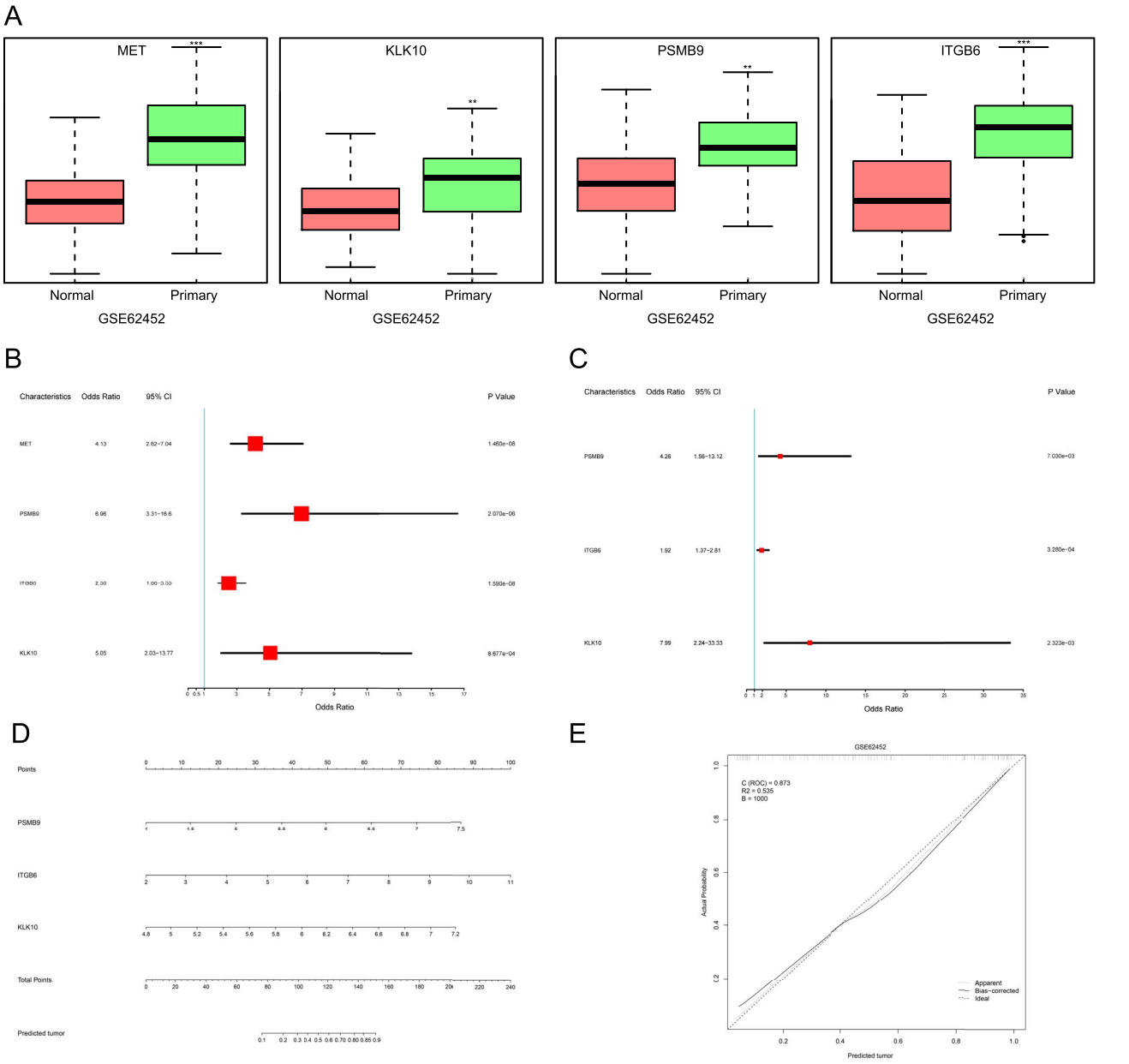

Figure 7. Construction of diagnostic model. (**A**) Boxplot showing the expression level of four prognosis-related genes among normal pancreas and PDAC in GSE62452. (**B**, **C**) Univariate and multivariate logistic regression analyses were used to select risk factors of the occurrence of PDAC; red boxes represent $P < 0.05$ in the forestplot. (**D**) The diagnosis-nomogram was drawn to predict the occurrence PDAC. (**E**) Calibration curve showing the agreement between actual and nomogram-predicted PDAC; the gray diagonal line is reference line.

conducted after Unicox analysis will cause model overfitting.

In conclusion, we constructed and validated the prognostic and diagnostic models for PDAC based on scRNA-seq and bulk-seq datasets. In addition, we established an easy-to-use nomogram combining risk score, *N* stage and margin status, which will help identify high-risk PDAC patients.

Materials and Methods

Patient cohorts and study design

The raw gene-cell matrix for scRNA-seq analysis was downloaded from GSA database (https://bigd.big.ac.cn/gsa) (access number: CRA00116) (36). A total of 24 PDAC and 11 normal pancreatic specimens were included in the scRNA-seq analysis. None of the patients received radiotherapy or chemotherapy before operation (clinicopathological characteristics are shown in Supplementary Material, Table S1). RNA-seq and clinical data of TCGA_PAAD and PACA_AU were downloaded from TCGA and ICGC, respectively, updated to April 20, 2021. We only focused on subjects diagnosed with PDAC and excluded subjects with neuroendocrine or acinar cell carcinoma. The gene microarray data were downloaded from GEO database under access numbers GSE57495, GSE71729, GSE62452, GSE15471 and GSE16515 using GEOquery package (37–41). The clinicopathological

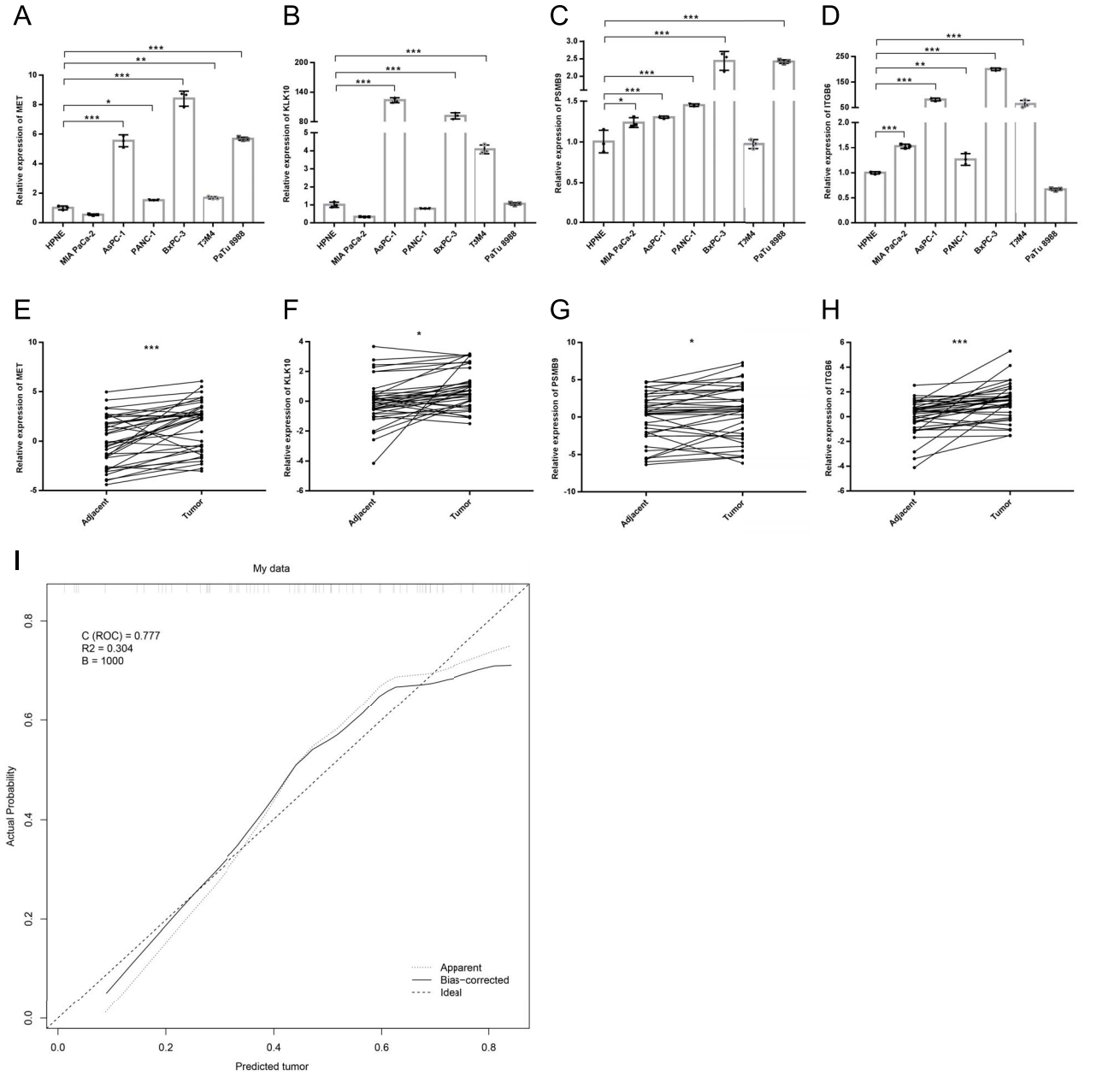

Figure 8. Validation of diagnostic model using our data. (**A**–**D**) RT-qPCR was performed to show the expression levels of *MET*, *KLK10*, *PSMB9* and *ITGB* among pancreatic cell lines. (**E**–**H**) The relative expression levels of *MET*, *KLK10*, *PSMB9*, and *ITGB6* between tumor and tumor-adjacent tissues were shown. (**I**) Calibration curve showing the performance of diagnostic model in our dataset.

characteristics of patients for prognostic and diagnostic model development are shown in Table 1. A total of 37 postoperative PDAC and matched adjacent normal pancreatic specimens were retrieved from the Department of General Surgery of Peking University First Hospital for validation test. This study was approved by Ethics Committee of Peking University First Hospital (Approval No. 2019-147) and was conducted in accordance with ethical guidelines (Declaration of Helsinki). Written informed consent was obtained from all participants.

This study employed a three-phase design: in the initial scRNA-seq analysis phase, we identified the cell atlas for PDAC and normal pancreas and identified DEGs between tumor cells and normal ductal cells. In the second phase, we developed and validated a prognostic and diagnostic model using TCGA_PAAD, PACA_AU and gene microarray data from the GEO datasets. In the third phase, we evaluated the reliability of the model by real-time quantitative PCR (RT-qPCR) assays using our data and IHC data from HPA. The flow chart of this study has been depicted in Fig. 1.

Cell culture

Human ductal cell line (hTERT-HPNE) and pancreatic cancer cell lines, MIA PaCa-2, AsPC-1, BxPC-3, PANC-1 and T3M4, were bought from ATCC. Pancreatic cancer lines PaTu 8988 was provided from PharmLab (China). All cell lines were authentic by short tandem repeats profile.

The hTERT-HPNE, MIA PaCa-2 and PANC-1 (DMEM, Gibco, USA) and AsPC-1, BxPC-3 and T3M4 (RPMI 1640, Gibco) were cultured in cell culture dishes (NEST Biotechnology, China) in humidified incubator at 37°C with 5% CO2.

scRNA-seq analysis

All specimens were merged as an original seurat object using Seurat (v3.2.3) R toolkit (42). This object was filtered to remove unqualified cells (<200 genes/cell, >10% mitochondrial genes, transcripts/cell <1000 or >20 000) and genes (<10 cells/gene) and was normalized (LogNormalize). The percentage of mitochondria genes and total counts were used to scale data. Next, 2000 highly variable genes were selected for PCA. The 'harmony' method was used to integrate the dataset from different specimens. Significant principle components were identified by JackStraw analysis. Cell atlas was visualized using t-SNE analysis.

Cluster marker genes were found through one-vs-rest binary classification metrics. The cell type of each cluster was identified by aligning marker genes to known signature genes reported in previous studies and CellMarker database (http://biocc.hrbmu.edu.cn/CellMarker/) (43). The known signature genes were AMBP, CFTR, FXYD2, KRT18 and KRT8 (ductal cells); CD68 and AIF1 (macrophages); MS4A1, CD79A, CD79B and VPREB3 (B cells); CDH5, RAMP2, PLVAP and VWF (endothelial cells); RGS5, NDUFA4L2, ADIRF and TAGLN (stellate cells); CD3D, CD3E and CD2 (T cells); LUM, COL1A1 and DCN (fibroblast cells); PRSS1 and REG1A (acinar cells); PCSK1N, INS, PPY and SST (endocrine cells).

Cellular components and cell-cycle analysis

We exported the meta.data from the seurat object and counted the proportion of cell subpopulations in PDAC and normal pancreatic specimens. The cell-cycle score of each cell was calculated using CellCycleScoring algorithm in the Seurat package, then each cell was classified into three statuses, including G1, S and G2M. We counted the proportion of cell-cycle statuses in PDAC and normal pancreatic specimens and compared it between tumor cells and normal ductal cells.

The identification and annotation of DEGs

FindMarker algorithm was utilized to identify DEGs between groups in scRNA-seq analysis (logfc.threshold = 0.5, q-value < 0.05). In addition, DEGs between PDAC and normal pancreatic tissues were identified based on TCGA and GTEx database using GEPIA2 online tools (44). The common DEGs were shown by the Upset plot and Venn plot. We performed DEGs' annotation by GO, KEGG ('clusterProfiler' R package, v3.18.0) and GSEA (GSEA tool, v4.0.1) analyses by running default parameters.

CNV inferring

Somatic large-scale chromosomal CNV score was calculated using 'inferCNV' R package. A raw counts matrix of scRNA-seq, annotation file and gene/chromosome position file were prepared according to data requirements (https://github.com/broadinstitute/inferCNV). We selected normal ductal cells, endothelial cells, stellate cells and macrophages as reference normal cells. The CNV score was calculated as quadratic sum of CNV region.

Construction and validation of prognostic model

To construct the prognostic model, we imported TCGA_PAAD datasets into R tool and classified it into train set and validation set randomly in 2:1. The common DEGs identified by both scRNA-seq and GEPIA2 analyses were used for Unicox analysis for OS. Significant OS-related genes were selected ($P < 0.001$) to further perform variable selection using LASSO-penalized Cox regression analysis. Then, LASSO-selected genes were subjected to Multicox analysis using 'survival' R package (v3.2.7). Risk score = $h0^{*}\ e^{\wedge}\sum_{i=0}^{n}\exp()$. Patients were classified into high-risk group and low-risk group based on the median risk score. KM curve was used to compare the OS between two groups. Time-dependent ROC was used to evaluate the accuracy of prognostic model by 'survivalROC' R package (v1.0.3). Then, TCGA_PAAD validation set and PACA_AU, GSE57495, GSE71729 were employed for the internal and external validations of the prognostic model. Finally, Unicox and Multicox analyses were performed to test the correlation between risk score, clinicopathological characteristics and OS. A nomogram was developed using 'rms' R package (v6.2.0) to predict 2-year and 3-year OSs in TCGA_PAAD, and the calibration curve was used to evaluate the accuracy of nomogram-predicted OS.

Construction and validation of diagnostic model

In order to test the diagnostic value of prognosis-related genes, univariate and multivariate logistic regression were performed to construct the diagnostic model using GSE62452 dataset. Similarly, a nomogram was developed to visualize the results of multivariate logistic regression, and calibration curve was used to evaluate the accuracy of nomogram-predicted PDAC. In addition, GSE71729, GSE15471, GSE16515 and patient cohort from our department were used to validate the reliability of this diagnostic model. We drew box plot to compare the gene expression difference between PDAC and normal pancreatic tissue.

RNA extraction and RT-qPCR

Total RNA from human PDAC and adjacent normal tissue were extracted by standard TRIzol/chloroform extraction method (Invitrogen, USA). First-strand of cDNAs were synthesized from the 2 μg total RNA with ReverTra Ace qPCR RT kit (TOYOBO, Japan) according to the manufacturer's instructions. RT-qPCR was performed by SYBR Green Realtime PCR Master Mix (TOYOBO) using AB7500 machine. The following primers were used: MET (forward primer: CCCGAAGTGTAAGCCCAACT, reverse primer: AGGATACTGCACTTGTCGGC); PSMB9

(forward primer: CCATCGAGTCATCTTGGGCA, reverse primer: ACCATACCAGGTTTTGGCCC); KLK10 (forward primer: TCGAGTAGGGGATGACCACC, reverse primer: ATGGACAACAGAGCGAGTGG); ITGB6 (forward primer: TGCGTCTCTGAAGATGGAGTG, reverse primer: GGGTCACCACAGGTAGGACA).

Statistical analysis

All statistical analyses were performed in R tool (v.4.0.3). RT-qPCR assays were performed in three replicates and repeated three times independently. The KM method and the corresponding log-rank test were performed to identify the prognostic value of marker genes. Statistical significance was defined as $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$.

Supplementary Material

Supplementary material is available at *HMG* online.

Acknowledgements

We would like to thank Zebin Mao for experiment suggestions and discussion, Zhengkui Zhang, Jisong Li for collecting pancreatic specimens and Baofa Sun and Dali Han for original scRNA-seq data; we are also grateful to the authors of R packages we used.

Funding

This study was supported by The Natural Science Foundation of China (NOs. 81871954 and 81672353) and Beijing Municipal Natural Science Foundation (NO. 7212111).

Conflict of Interest statement: None declared.

Ethics statement

This study was approved by Ethics Committee of Peking University First Hospital (Approval No. 2019-147) and was conducted in accordance with ethical guidelines (Declaration of Helsinki). Written informed consent was obtained from all participants.

References

1. Siegel, R.L., Miller, K.D. and Jemal, A. (2020) Cancer statistics, 2020. *CA Cancer J. Clin.*, **70**, 7–30.
2. Ryan, D.P., Hong, T.S. and Bardeesy, N. (2014) Pancreatic adenocarcinoma. *N. Engl. J. Med.*, **371**, 2140–2141.
3. Vincent, A., Herman, J., Schulick, R., Hruban, R.H. and Goggins, M. (2011) Pancreatic cancer. *Lancet*, **378**, 607–620.
4. Garrido-Laguna, I. and Hidalgo, M. (2015) Pancreatic cancer: from state-of-the-art treatments to promising novel therapies. *Nat. Rev. Clin. Oncol.*, **12**, 319–334.
5. Yang, Y. (2020) Current status and future prospect of surgical treatment for pancreatic cancer. *Hepatobiliary. Surg. Nutr.*, **9**, 89–91.
6. Tanay, A. and Regev, A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature*, **541**, 331–338.
7. Peng, J., Sun, B., Chen, C., Zhou, J., Chen, Y., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G. *et al.* (2019) Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.*, **29**, 725–738.
8. Elyada, E., Bolisetty, M., Laise, P., Flynn, W.F., Courtois, E.T., Burkhart, R.A., Teinor, J.A., Belleau, P., Biffi, G., Lucito, M.S. *et al.* (2019) Cross-species single-cell analysis of pancreatic ductal adenocarcinoma reveals antigen-presenting cancer-associated fibroblasts. *Cancer. Discov.*, **9**, 1102–1123.
9. Gu, M., Sun, J., Zhang, S., Chen, J., Wang, G., Ju, S. and Wang, X. (2021) A novel methylation signature predicts inferior outcome of patients with PDAC. *Aging (Albany NY)*, **13**, 2851–2863.
10. Yue, P., Zhu, C., Gao, Y., Li, Y., Wang, Q., Zhang, K., Gao, S., Shi, Y., Wu, Y., Wang, B. *et al.* (2020) Development of an autophagy-related signature in pancreatic adenocarcinoma. *Biomed. Pharmacother.*, **126**, 110080.
11. Wen, X., Shao, Z., Chen, S., Wang, W., Wang, Y., Jiang, J., Ma, Q. and Zhang, L. (2020) Construction of an RNA-binding protein-related prognostic model for pancreatic adenocarcinoma based on TCGA and GTEx databases. *Front. Genet.*, **11**, 610350.
12. Li, M., Li, H., Chen, Q., Wu, W., Chen, X., Ran, L., Si, G. and Tan, X. (2020) A novel and robust long noncoding RNA panel to predict the prognosis of pancreatic cancer. *DNA Cell Biol.*, **39**, 1282–1289.
13. Chen, K., Wang, Q., Li, M., Guo, H., Liu, W., Wang, F., Tian, X. and Yang, M. (2021) Single-cell RNA-seq reveals dynamic change in tumor microenvironment during pancreatic ductal adenocarcinoma malignant progression. *EBioMedicine*, **66**, 103315.
14. Garg, M., Braunstein, G. and Koeffler, H.P. (2014) LAMC2 as a therapeutic target for cancers. *Expert. Opin. Ther. Tar.*, **18**, 979–982.
15. Kosanam, H., Prassas, I., Chrystoja, C.C., Soleas, I., Chan, A., Dimitromanolakis, A., Blasutig, I.M., Rückert, F., Gruetzmann, R., Pilarsky, C. *et al.* (2013) Laminin, gamma 2 (LAMC2): a promising new putative pancreatic cancer biomarker identified by proteomic analysis of pancreatic adenocarcinoma tissues. *Mol. Cell. Proteomics*, **12**, 2820–2832.
16. Sun, Y., Wu, L., Zhong, Y., Zhou, K., Hou, Y., Wang, Z., Zhang, Z., Xie, J., Wang, C., Chen, D. *et al.* (2021) Single-cell landscape of the ecosystem in early-relapse hepatocellular carcinoma. *Cell*, **184**, 404, e416–421.
17. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.
18. Wang, R., Dang, M., Harada, K., Han, G., Wang, F., Pizzi, M.P., Zhao, M., Tatlonghari, G., Zhang, S., Hao, D. *et al.* (2021) Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma. *Nat. Med.*, **27**, 141–151.
19. Pastushenko, I. and Blanpain, C. (2019) EMT transition states during tumor progression and metastasis. *Trends Cell Biol.*, **29**, 212–226.
20. Ren, J. and Liang, Q. (2019) HMGB1 promotes the proliferation and invasion of oral squamous cell carcinoma via activating epithelial-mesenchymal transformation. *Biocell*, **43**, 199–206.
21. Yang, C. and Tian, Y. (2019) SPAG9 promotes prostate cancer growth and metastasis. *Biocell*, **43**, 207–214.
22. Bailey, P., Chang, D.K., Nones, K., Johns, A.L., Patch, A.M., Gingras, M.C., Miller, D.K., Christ, A.N., Bruxner, T.J., Quinn, M.C. *et al.* (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, **531**, 47–52.
23. Raffenne, J. and Cros, J. (2018) Molecular characterisation defines several subtypes of pancreatic ductal adenocarcinoma. *B. Cancer.*, **105**, 55–62.
24. Aguirre, A.J., Nowak, J.A., Camarda, N.D., Moffitt, R.A., Ghazani, A.A., Hazar-Rethinam, M., Raghavan, S., Kim, J., Brais, L.K., Ragon, D. *et al.* (2018) Real-time genomic characterization of advanced

pancreatic cancer to enable precision medicine. *Cancer. Discov.*, **8**, 1096–1111.
25. Puleo, F., Nicolle, R., Blum, Y., Marisa, L., Demetter, P., Quertinmont, E., Svrcek, M., Elarouci, N., Iovanna, J., Franchimont, D. *et al.* (2018) Stratification of pancreatic ductal adenocarcinomas based on tumor and microenvironment features. *Gastroenterology*, **155**, 1999, e1993–e2013.
26. Halbrook, C.J. and Lyssiotis, C.A. (2017) Employing metabolism to improve the diagnosis and treatment of pancreatic cancer. *Cancer Cell*, **31**, 5–19.
27. Hosein, A.N., Brekken, R.A. and Maitra, A. (2020) Pancreatic cancer stroma: an update on therapeutic targeting strategies. *Nat. Rev. Gastroenterol. Hepatol.*, **17**, 487–505.
28. Navin, N.E. (2015) The first five years of single-cell cancer genomics and beyond. *Genome Res.*, **25**, 1499–1507.
29. Shema, E., Bernstein, B.E. and Buenrostro, J.D. (2019) Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat. Genet.*, **51**, 19–25.
30. Hou, J., Wang, Z., Li, H., Zhang, H. and Luo, L. (2020) Gene signature and identification of clinical trait-related m(6) A regulators in pancreatic cancer. *Front. Genet.*, **11**, 522.
31. Luo, L., Li, Y., Huang, C., Lin, Y., Su, Y., Cen, H., Chen, Y., Peng, S., Ren, T. and Xie, R. (2021) A new 7-gene survival score assay for pancreatic cancer patient prognosis prediction. *Am. J. Cancer Res.*, **11**, 495–512.
32. Jiang, P., Yang, F., Zou, C., Bao, T., Wu, M., Yang, D. and Bu, S. (2021) The construction and analysis of a ferroptosis-related gene prognostic signature for pancreatic cancer. *Aging (Albany NY)*, **13**, 10396–10414.
33. Ye, Q., Ling, S., Zheng, S. and Xu, X. (2019) Liquid biopsy in hepatocellular carcinoma: circulating tumor cells and circulating tumor DNA. *Mol. Cancer*, **18**, 114.
34. Kalluri, R. (2016) The biology and function of exosomes in cancer. *J. Clin. Invest.*, **126**, 1208–1215.
35. Kalluri, R. and LeBleu, V.S. (2020) The biology, function, and biomedical applications of exosomes. *Science*, **367**, eaau6977.
36. Wang, Y., Song, F., Zhu, J., Zhang, S., Yang, Y., Chen, T., Tang, B., Dong, L., Ding, N., Zhang, Q. *et al.* (2017) GSA: genome sequence archive<sup/>. *Genomics. Proteomics. Bioinformatics.*, **15**, 14–18.
37. Chen, D.T., Davis-Yadley, A.H., Huang, P.Y., Husain, K., Centeno, B.A., Permuth-Wey, J., Pimiento, J.M. and Malafa, M. (2015) Prognostic fifteen-gene signature for early stage pancreatic ductal adenocarcinoma. *PLoS One*, **10**, e0133562.
38. Moffitt, R.A., Marayati, R. and Flate, E.L. (2015) Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.*, **47**, 1168–1178.
39. Yang, S., He, P., Wang, J., Schetter, A., Tang, W., Funamizu, N., Yanaga, K., Uwagawa, T., Satoskar, A.R., Gaedcke, J. *et al.* (2016) A novel MIF signaling pathway drives the malignant character of pancreatic cancer by targeting NR3C2. *Cancer Res.*, **76**, 3838–3850.
40. Idichi, T., Seki, N., Kurahara, H., Yonemori, K., Osako, Y., Arai, T., Okato, A., Kita, Y., Arigami, T., Mataki, Y. *et al.* (2017) Regulation of actin-binding protein ANLN by antitumor miR-217 inhibits cancer cell aggressiveness in pancreatic ductal adenocarcinoma. *Oncotarget*, **8**, 53180–53193.
41. Li, L., Zhang, J.W., Jenkins, G., Xie, F., Carlson, E.E., Fridley, B.L., Bamlet, W.R., Petersen, G.M., McWilliams, R.R. and Wang, L. (2016) Genetic variations associated with gemcitabine treatment outcome in pancreatic cancer. *Pharmacogenet. Genomics*, **26**, 527–537.
42. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
43. Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.
44. Tang, Z., Li, C., Kang, B., Gao, G., Li, C. and Zhang, Z. (2017) GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic. Acids Res.*, **45**, W98–W102.