Research Paper

A two-gene-based prognostic signature for pancreatic cancer

Shuyi Zhou[1,2], Yuanliang Yan[3,4], Xi Chen[1,3,4], Shuangshuang Zeng[1,3,4], Jie Wei[1,3,4], Xiang Wang[1,3,4], Zhicheng Gong[1,3,4], Zhijie Xu[5]

[1]The Hunan Institute of Pharmacy Practice and Clinical Research, Xiangya Hospital, Central South University, Changsha 410008, China
[2]Department of Hepatobiliary Surgery, Hunan Provincial People's Hospital Xingsha Branch, People's Hospital of Changsha County, Hunan Normal University, Changsha 410008, China
[3]Department of Pharmacy, Xiangya Hospital, Central South University, Changsha 410008, China
[4]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha 410008, China
[5]Department of Pathology, Xiangya Hospital, Central South University, Changsha 410008, China

Correspondence to: Zhijie Xu; **email:** xzj1322007@csu.edu.cn
Keywords: pancreatic cancer, ANLN, HIST1H1C, overall survival, prognosis
Received: December 10, 2019 **Accepted:** June 29, 2020 **Published:** September 23, 2020

ABSTRACT

The purpose of this study was to identify a vital gene signature that has prognostic value for pancreatic cancer based on gene expression datasets from the Cancer Genome Atlas and Gene Expression Omnibus. A total of 34 genes were obtained by the univariate analysis, which were significantly associated with the overall survival of PC patients. After further analysis, *Anillin (ANLN)* and *Histone H1c (HIST1H1C)* were identified and considered to be the most significant prognostic genes among the 34 genes. A prognostic model based on these two genes was constructed, and successfully distinguished pancreatic cancer survival into high-risk and low-risk groups in the training set and testing set. Subsequently, independent predictive factors, including the age, margin condition and risk score, were then employed to construct the nomogram model. The area under curve for the nomogram model was 0.826 at 0.5 years and 0.726 at 1 year, and the C-index of the nomogram model was 0.664 higher than the others variables alone. These findings have indicated that high expression of *ANLN* and *HIST1H1C* predicted poor outcomes for patients with pancreatic cancer. The nomogram model based on the expression of two genes could be valuable for the guidance of clinical treatment.

INTRODUCTION

Pancreatic cancer (PC), the fourth leading cause of cancer-related death in the US, is estimated to cause 227,000 deaths per year worldwide. As the incidence rate of PC in developed countries continues to rise, it will be the second most fatal cancer in 2020 [1]. Although modern cancer chemotherapy and mature surgical techniques cause a modest incremental improvement in patient outcomes [2, 3], with the 5-year overall survival (OS) improving from 5% to 9% and the median OS improving to approximately 11 months compared with the historic benchmark of 5 to 6 months, its prognosis still remains extremely poor [4]. Thus, efforts are needed to develop new ways to treat this fatal disease.

In recent years, precision medicine has offered numerous valuable insights into PC treatment [5]. Individual therapy enables PC patients to have the largest gains with minimum risk. Much more radical treatments, such as combined regimens and extensive radical operation, are preferred for PC patients with high recurrence risks [6]. Therefore, it is extremely necessary to distinguish the high-risk group from all PC patients.

However, there is no sufficient predictive system to predict the outcomes of patients with PC. Traditional risk stratification systems, such as the American Joint Committee on Cancer (AJCC) staging system, have been considered relatively nondiscriminatory for predicting differences in survival among PC patients [7, 8]. With the use of next-generation sequencing and microarray technologies, many studies have found the importance of gene signatures in the initiation, progression and prognosis of human tumors [9–13]. The facilitating investigation of interactions between gene signatures and tumors has made it possible to use signatures to stratify patient risks.

In this study, we aimed to explore the differences in mRNA expression profiles between PC and the adjacent pancreas using The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) datasets. After important prognosis-related genes were identified, we established a two-gene prognostic model that included *Anillin (ANLN)* and *Histone H1c (HIST1H1)* and was applicable for guiding prognostic assessment and treatment decision-making during the early postoperative period.

RESULTS

Common differentially expressed genes between PC and normal tissues

As shown in the analysis process flowchart (Figure 1), after the genomic differential expression analysis of the 3 datasets (GSE28735, GSE62452 and TCGA), there were 222 differentially expressed genes (DEGs) in common (Supplementary Table 1). Using the criteria of P value < 0.01, the number of DEGs in GSE28735, GSE62452 and TCGA are 5250, 7180 and 222, respectively. A Venn diagram was applied to visualize the DEG relationships of the 3 datasets (Figure 2A). We also used a heatmap of the differentially expressed mRNAs to better differentiate normal tissues from PC (Figure 2B and Supplementary Tables 2, 3).

Functional annotation of common DEGs

To further describe the biological functions of the 222 common differentially expressed genes in detail, we performed functional annotation and enrichment analysis using the R package "ClusterProfile". The biological process indicated that genes were enriched for positive regulation of defense response (Figure 3A). For cell component enrichment, the common differentially expressed genes were primarily enriched for apical part of cell (Figure 3B). The molecular function of the genes was enriched mainly for Alcoholism (Figure 3C). The KEGG pathway indicated that the DEGs were mainly enriched for actin binding (Figure 3D).

Screening of the prognostic PC gene signature among DEGs

After a univariate Cox analysis with the bootstrap resampling, we found that 34 of the 222 common DEGs

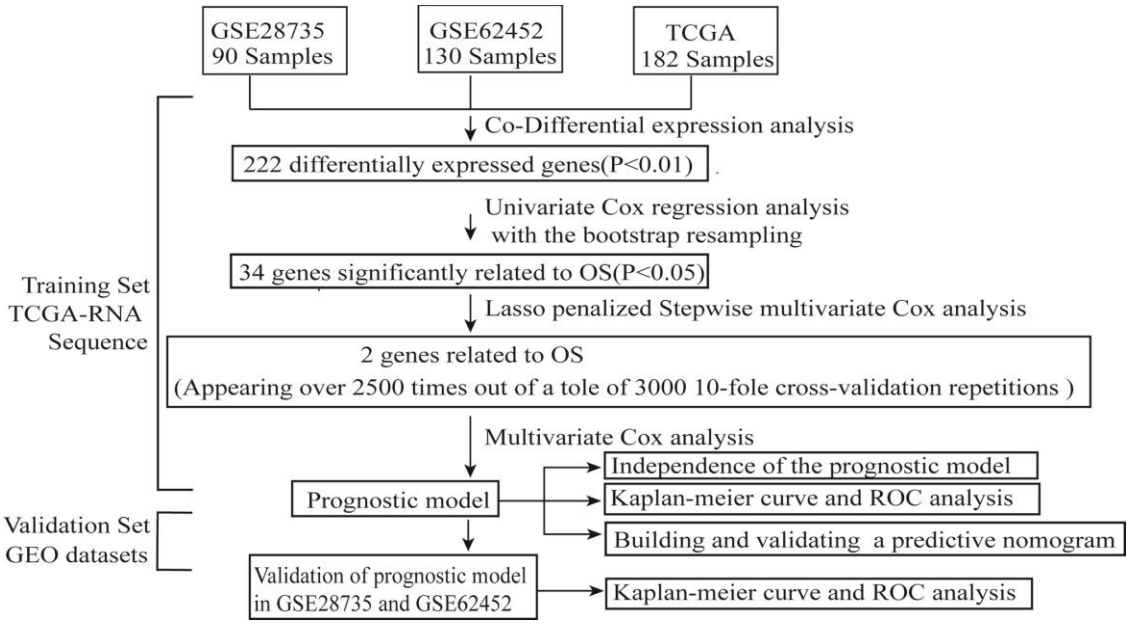


Figure 1. The flowchart of the whole analysis process.

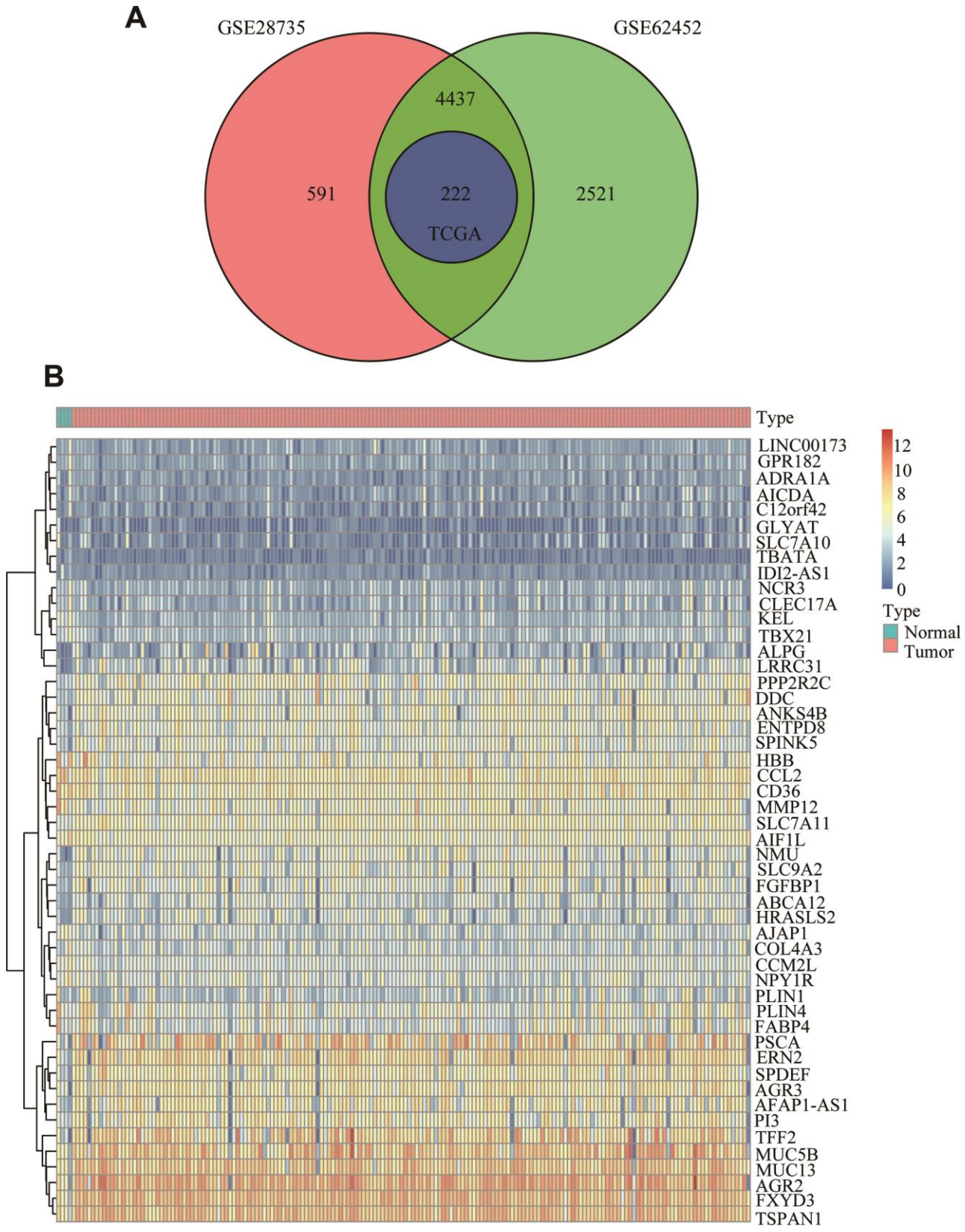


Figure 2. Common differentially expressed genes between PC and normal tissues. (**A**) Venn diagram showing the common DEGs in PC and adjacent normal tissues from the GSE28735, GSE62452 and TCGA datasets. (**B**) Heatmap analysis of the 222 DEGs, which contained the 50 highest expressed genes and the 50 the lowest expressed genes according to the log2FC between normal tissues and cancer tissues from the TCGA datasets.

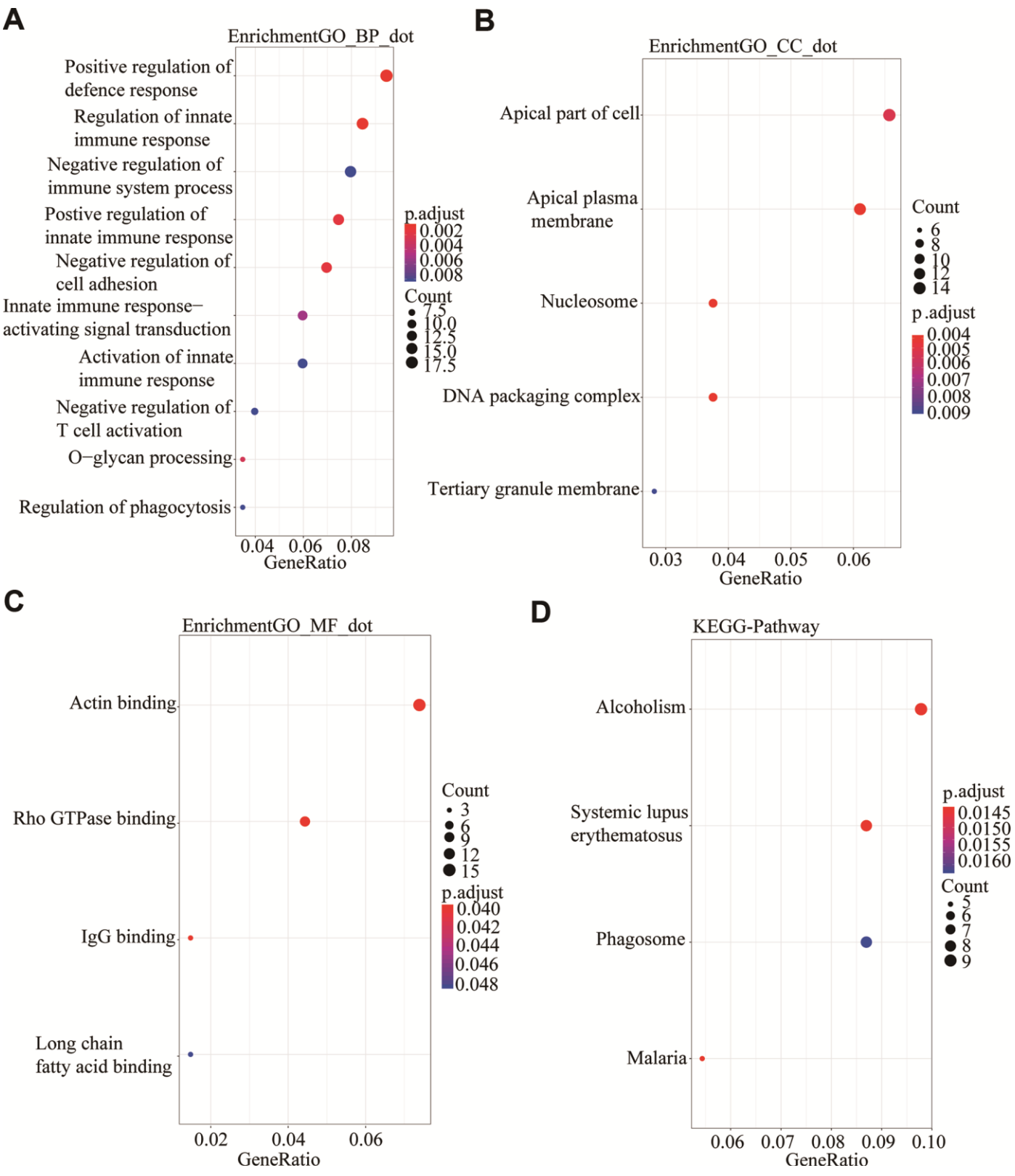


Figure 3. GO and KEGG analysis using the R package "Clusterprofile" for the 222 common DGEs from the three databases. P<0.05 was set as the threshold. (**A**) Biological process. (**B**) Cell component. (**C**) Molecular Function. (**D**) Kyoto Encyclopedia of Genes and Genomes.

were significantly related to PC patient survival (P<0.05). Next, to further reduce the number of genes and overfitting, we subsampled 70% of patients from TCGA dataset (Supplementary Table 4) for analysis at one time and applied a Lasso-penalized regression with a 10-fold cross-validation 3000 times when performing the stepwise multivariate Cox analysis. Two genes, *ANLN* and *HIST1H1C*, appeared over 2500 times among the 3000 10-fold cross-validation repetitions (Figure 4A and Supplementary Table 5). Next, we used a box plot to illustrate that *ANLN* and *HIST1H1C* are constantly and significantly highly expressed in tumor tissues from all the datasets (GSE28735, GSE62452 and TCGA) (Figure 4B–4G and Supplementary Tables 6–8). Furthermore, the data from Cancer Cell Line Encyclopedia (https://portals.broadinstitute.org/ccle) were conducted as the heatmap displaying the elevated *ANLN* and *HIST1H1C* expression levels in several PC cells (Figure 4H and Supplementary Table 9). Next, we wanted to evaluate the protein expression levels of ANLN and HIST1H1C in PC patients. We analyzed the immunohistochemical data from the Human Protein Atlas (http://www.proteinatlas.org/) shown in Figure 4I, and found significantly elevated levels of ANLN and HIST1H1C in the tumor tissues. In addition, the pancancer analysis from GEPIA 2.0 (http://gepia2.cancer-pku.cn/#index) showed that upregulated *ANLN* and *HIST1H1C* transcripts are frequently observed in multiple cancer types, including PC (Supplementary Figure 1). Taken together, *ANLN* and *HIST1H1C* play potential oncogenic roles in most types of human cancers.

The risk stratification and ROC curve indicate the good performance of the two-gene based signature

Using the TCGA dataset, we generated a predictive model based on the expression of two genes, which was characterized by the linear combination of the expression levels of the two genes weighted by their relative coefficient in the multivariate Cox regression. We subsequently calculated the two-gene expression risk score and used X-tile diagrams to produce the optimal cut-off value for the risk score. According to the risk score cut-off point, 42 patients were classified into the high-risk group, and the remaining 118 patients were assigned to the low-risk group. As is shown in Figure 5A and Supplementary Table 10, the Kaplan-meier (K-M) OS curves of the two groups based on the two genes were significantly different (median OS, 1.81 years vs 1.08 years, P=0.00027). To assess the prognostic capacity of the two-gene signature, the area under curve (AUC) of a time-dependent ROC curve was calculated. The AUCs of the two-gene biomarker prognostic model were 0.781, 0.673 and 0.646 for the 0.5-, 1- and 1.5-year survival times, respectively (Figure 5B). To further evaluate the generality of the two-gene biomarker prognostic model, we verified the model with the GEO dataset (GES28735), which contains both mRNA expression and clinical survival data from 45 PC patients. Using the same data management as in the TCGA, we also calculated the two-gene risk score according to the expression levels in GES28735 and the coefficient of the multivariate Cox regression and found an optimal cut-off value for the risk score by means of the X-tile diagrams. A total of 42 PC patients in the GSE28735 dataset were classified into high-risk group (n=11) and low-risk group (n=31). Consistent with the results in the TCGA, and as is shown in Figure 5C and Supplementary Table 11, the K-M OS curves indicated that the OS of PC patients included in the GSE28735 data in the high-risk group was significantly lower than that in the low-risk group (median OS 0.58 years vs 2.08 years, P=0.0016). Moreover, the time-dependent ROC analyses for the survival prediction of the prognostic model obtained AUCs of 0.624 at 0.5 years, 0.692 at 1 year and 0.664 at 1.5 years (Figure 5D). In addition, we further verified the risk model with the GEO dataset (GSE62452), which contains both mRNA expression and clinical survival data from 69 PC patients. A total of 66 PC patients in the GSE62452 dataset were classified into high-risk group (n=33) and low-risk group (n=33). Consistent with the results in the TCGA and GSE28735, and as is shown in Figure 5E and Supplementary Table 12, the K-M OS curves indicated that the OS of PC patients in the high-risk group from GSE62452 was significantly lower than that in the low-risk group (median OS 1.26 years vs 2.09 years, P=0.002). Moreover, the time-dependent ROC analyses for the survival prediction of the prognostic model obtained AUCs of 0.729 at 2 year and 0.824 at 3 years (Figure 5F). Above all, we concluded that the two-gene signatures were able to predict the prognosis in PC patients.

Building and validating a predictive nomogram

Univariate and multivariate Cox regression analyses were employed to detect the independent predictive ability of the two-gene-based prognostic model in the abovementioned TCGA PC cohort with detailed clinical information. Using the univariate Cox regression analysis, we found that the prognostic model and age had prognostic values, while the others variables did not significantly correlate with OS (Figure 6). Considering that the margin condition almost reached statistical significance (P=0.054) and might affect the prognosis of PC patients according to the clinical experience, we incorporated age, margin condition and prognostic model into the multivariate Cox regression analysis. As a result, both the age and prognostic model were independent prognostic factors, and the margin condition nearly reached statistical significance (Figure 6).

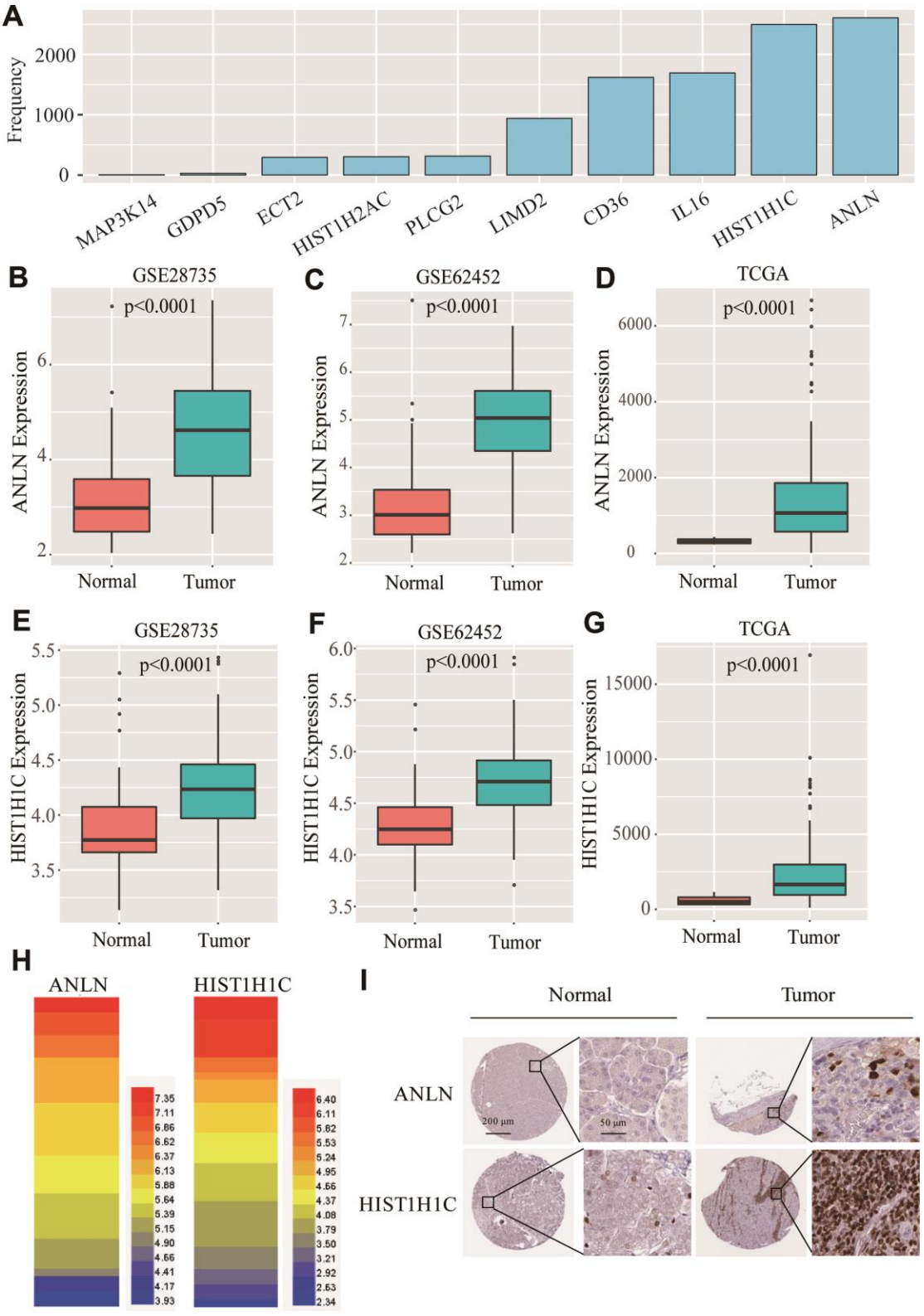


Figure 4. Screening of the two-gene-based signatures in PC patients. (**A**) After 3000 analyses, *ANLN* and *HIST1H1C* appeared more than 2500 times as the independent prognostic genes among 35 survival related genes. (**B**–**D**) The boxplot shows that ANLN were constantly high-expressed in GSE28735, GSE62452 and TCGA. (**E**–**G**) The boxplot shows that HIST1H1C were constantly high-expressed in GSE28735, GSE62452 and TCGA. (**H**) The heatmap of *ANLN* and *HIST1H1C* mRNA expression in PC cells from Cancer Cell Line Encyclopedia. (I) the Human Protein Atlas project shows representative immunohistochemical images of ANLN and HIST1H1C in PC tissues compared with surrounding normal tissues.

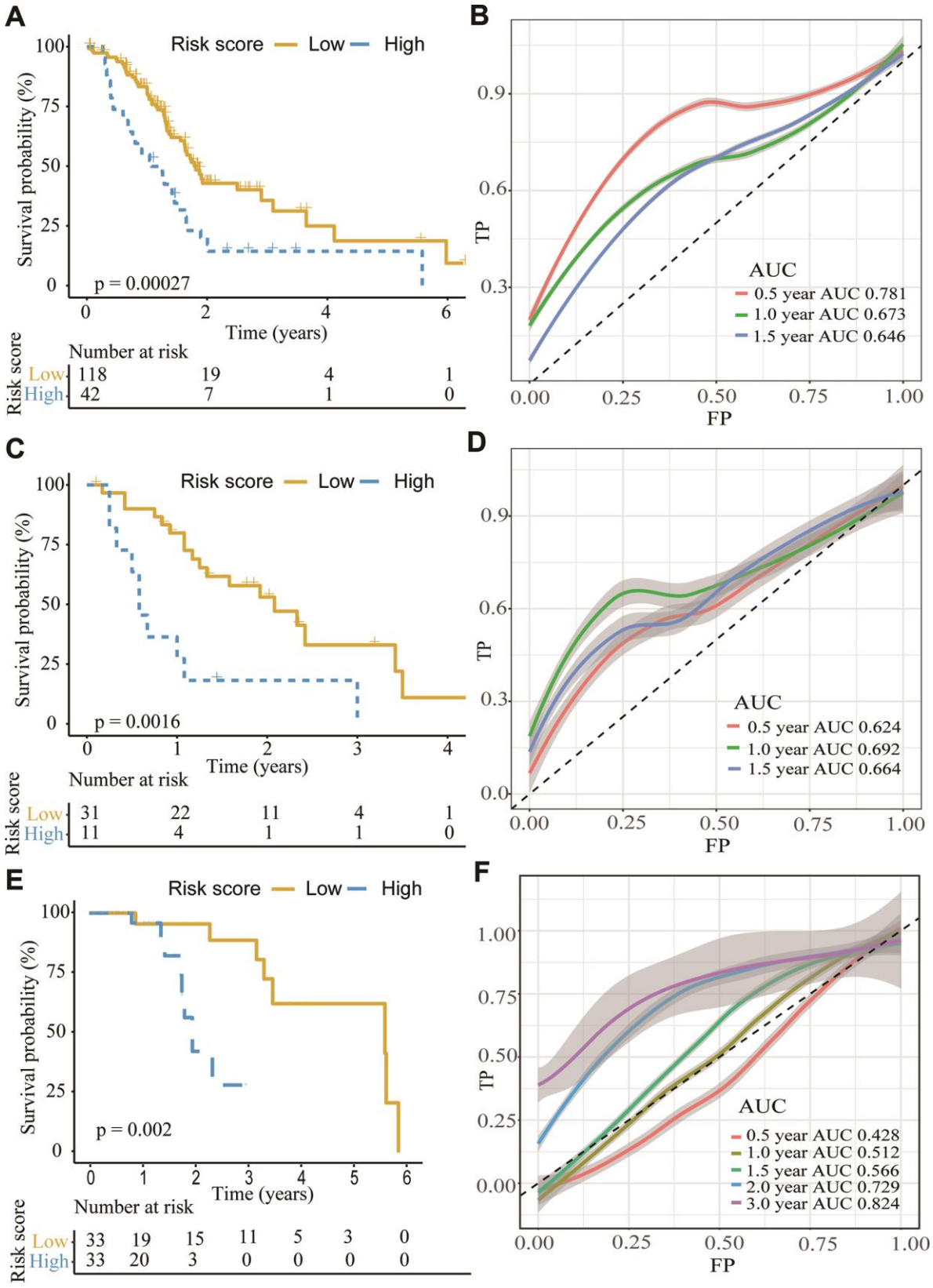


Figure 5. The K-M plot showed a lower overall survival in the high risk group compared to the low risk group divided by the optimal cut-off point. (**A**, **B**) K-M and time-dependent ROC curves for the prognostic model based on *ANLN* and *HIST1H1C* expression in the TCGA PC cohort. (**C**, **D**) K-M and time-dependent ROC curves for the prognostic model based on *ANLN* and *HIST1H1C* expression in the GSE28735. (**E**, **F**) K-M and time-dependent ROC curves for the prognostic model based on *ANLN* and *HIST1H1C* expression in the GSE62452.

To make the prognostic model clinical applicable, a nomogram was applied to predict the probability of the 0.5-, 1-, and 1.5-year OS in the TCGA cohort. The predictive factors in the nomogram include age, margin condition and prognostic model. The C-index of the nomogram model was 0.664, which was higher than the others variables alone (Figure 7A and Supplementary Table 13). Moreover, calibration plots were also used to visualize the performance of the nomograms, with more overlap with the gray-line representing better performance (Figure 7B–7D). The AUCs of the different models were also calculated; the nomogram's AUC was 0.826 at 0.5 years and 0.726 at 1 year, which was the largest compared to the other models (Figure 8). These findings demonstrated that the nomogram built with the combined model is the best nomogram to predict survival for patients with PC, when compared with nomograms built with a single prognostic factor, and demonstrated significance for facilitating patient counseling, decision-making and follow-up arranging.

DISCUSSION

Pancreatic cancer is one of the deadliest malignant tumors worldwide, with a gradually increasing morbidity each year. Despite the availability of improving surgical techniques, progress in PC treatment still remains slow, as the 5-year overall survival rate has only improved to 9% in the period 2006 to 2012 in the US [4, 14, 15]. The reasons for this phenomenon could be various, though one specific reason might be critical, namely, the lack of a risk assessment system, which makes it difficult to make therapy decisions for an individual patient with PC.

As is known to all, treatment for pancreatic cancer has experienced some changes from curative resection alone to surgical treatment with adjuvant chemotherapy, which has provided patients with many kinds of therapeutic regimens. In addition, owing to the enormous and high-quality clinical trials focused on adjuvant chemotherapy regimens after surgery, combination chemotherapy has been found to lead to longer survival in select PC patients [16]. However, the greater concomitant toxicity of combination chemotherapy has aroused a new debate: should combination chemotherapy be the first-line chemotherapy for all PC patients, especially when the patients are of a low recurrence risk? Moreover, the course of adjuvant chemotherapy and the frequency of reexamination after discharge also remain controversial

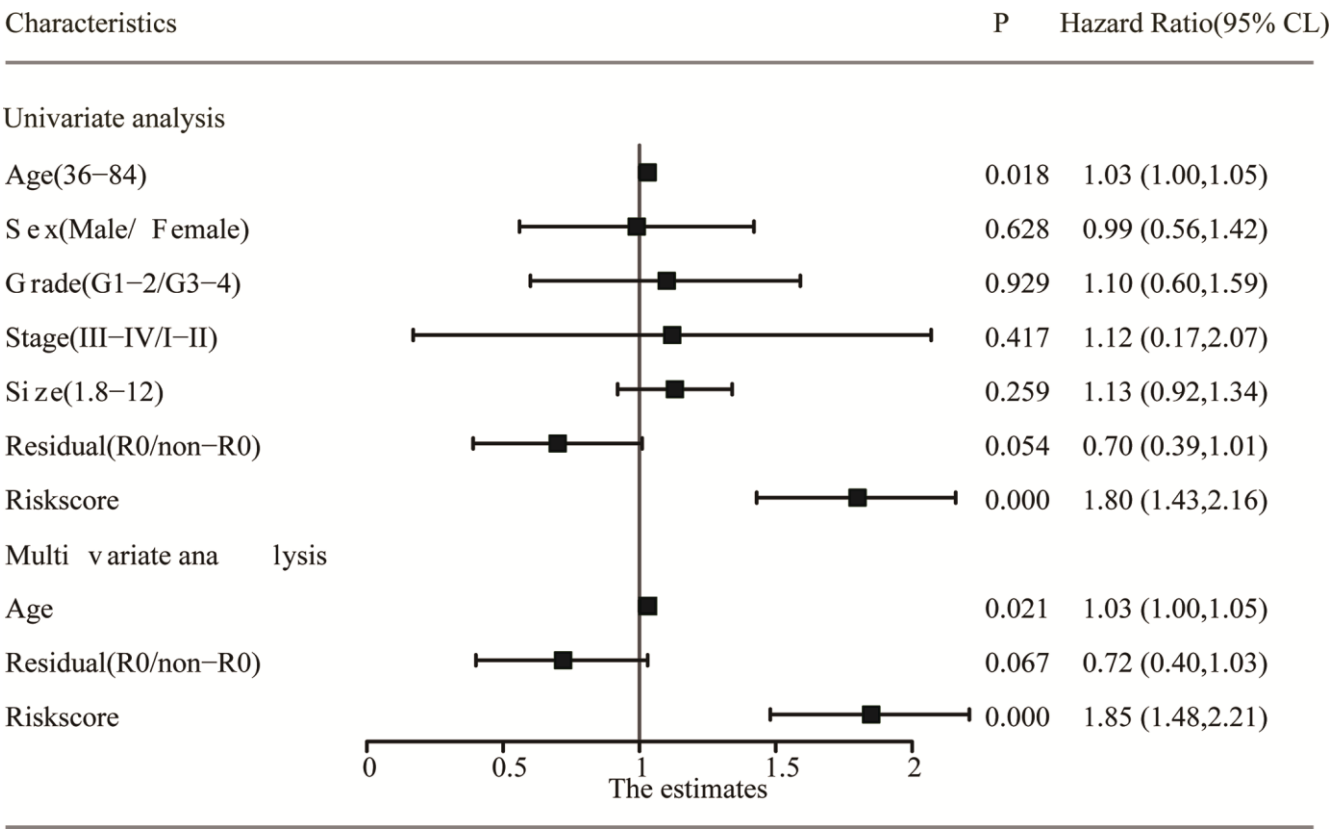


Figure 6. Univariate and multivariate analysis of the risk score and clinicopathological characteristics with OS. The residual (R0/non-R0) and risks core indicated the margin condition and prognostic model, respectively.

[17, 18]. It would be much better if there were an accurate and clinical applicable criterion to assess the possibility of recurrence in PC patients. It is important to note that resectability status alone is not a reliable prognostic factor in PC; even in potential curative resection patients, median survival outcomes were similar to nonradical resection patients [19]. Except for anatomical considerations, CA 19-9 levels are the most commonly used antigen in the clinic to assess the resectability and prognosis of PC patients [20]. However, CA 19-9 is also elevated in other benign conditions and multiple cancer types with limited sensitivity [21]. Several studies have recently characterized whole-genome changes occurring in PC by analyzing the mutational landscape of these deadly diseases [21]. Owing to these findings, the discovery of gene signatures to assess PC patient prognosis is of practicable and great value.

Recently, using the overlapping analysis, Yan et al. [22] identified four survival-related genes (*LYRM1*, *KNTC1*, *IGF2BP2* and *CDC6*) in four public PC datasets. And the predictive nomogram based these four survival-related genes shows robust performance in predicting PC prognosis. However, the disturbances arising from outliers do not fit well in their framework. Therefore, more reliable predictive model needs to be built and optimized. In our study, we selected genes that were

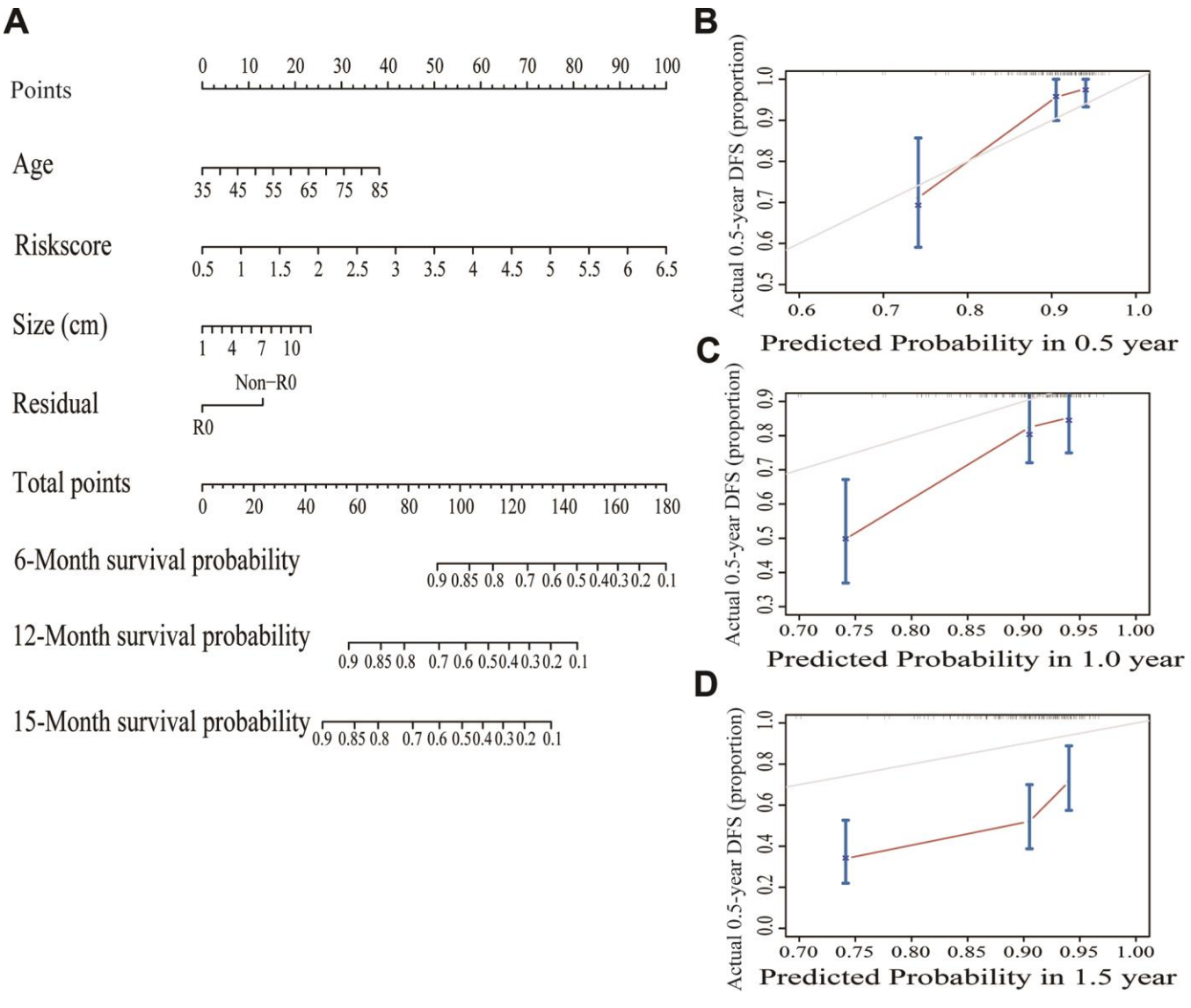


Figure 7. Building and validating a predictive nomogram. (**A**) A nomogram to predict survival probability at 6, 12 and 15 months after surgery. (**B**) Calibration curve for the nomogram when predicting 6 months of overall survival. (**C**) Calibration curve for the nomogram when predicting 12 months of overall survival. (**D**) Calibration curve for the nomogram when predicting 15 months of overall survival.

significantly highly/lowly expressed in all three datasets. Among the 35 survival-related genes, aside from the two genes that have not been reported to have a relationship with tumor progression or prognosis in any cancer types, 16 (45.7%) genes have been reported to be related to tumor progression and prognosis. Moreover, 11 (31.4%) genes have been reported to be related with the occurrence and outcome of PC, which demonstrated the correctness and repeatability of our data mining methods and results. To further minimize the scope of the survival-related genes, we applied multivariate Cox regression complex with a Lasso regression. To reduce the overfitting and avoid disturbances arising from outliers, we also used a 10-fold cross validation and performed the whole analytical process 3000 times. We selected the most frequent genes (*ANLN* and *HIST1H1C*) to construct the predicting model, which is different from the model conducted by Yan's group [22]. Then, we proved that the prognostic signature performed well for the discrimination of the high and low risk groups using both the TCGA PC datasets and GEO datasets. At the same time, the AUCs of the two-gene biomarker prognostic model were plotted, which demonstrated the acceptable predictive value of the two genes. Consistent with previous studies, the AJCC staging system allowed the prediction of clinical outcomes for PC patients. In spite of the anatomical extent of the cancer, which was assessed using a staging system, biological heterogeneity might play a critical role in pancreatic carcinoma. Compared with a single factor, our nomogram is more powerful for prediction and may became a reference item in the clinic in the future.

ANLN, an actin-binding protein, is essential for assembly of the cleavage furrow during the late stages of mitosis and acts as the central organizer at the cytokinetic machinery hub [23]. Owing to its critical role, studies have observed overexpressed *ANLN* in several types of cancers [24–26], and it has also been proven to be correlated with poor prognoses in breast cancer, lung cancer and hepatocellular carcinoma [27–29]. However, the role of *ANLN* in PC remains unclear. To the best of our knowledge, our study was the first to validate that high-expressed *ANLN* is associated with low overall survival by means of bioinformatic analysis. The pathogenesis of *ANLN* in tumor progression might act as a cell cycle regulator, enabling the promotion of tumor growth by decreasing apoptosis and DNA damage, though the detailed mechanisms require further investigation [30]. *HIST1H1C*, another PC molecular marker, is a basic nuclear protein responsible for interaction with the linker DNA between nucleosomes and functions to compact chromatin into higher order structures. Until now, however, only one study has reported *HIST1H1C* as a hub gene among the DEGs in nonfunctional pituitary adenomas, and one study illustrated that *HIST1H1C* is involved with tumor growth in pancreatic cancer [31]. There are few studies focused on *HIST1H1C*, which makes the significance of this gene in PC uncertain and worthy of future investigations.

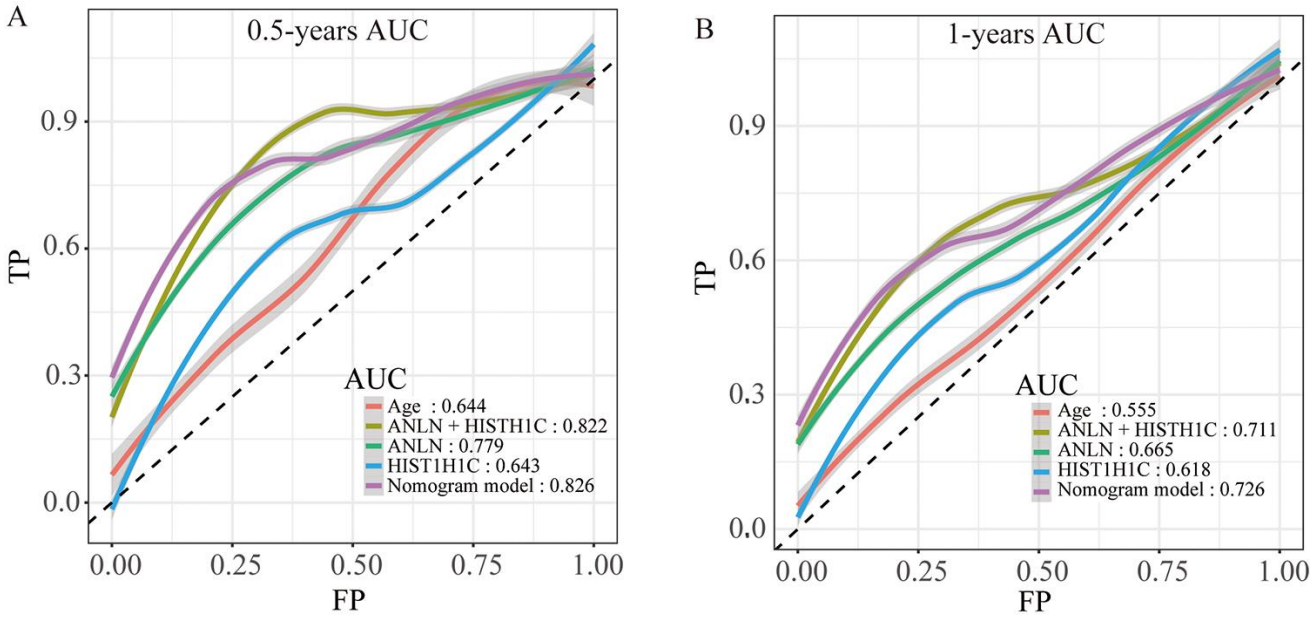


Figure 8. The time-dependent ROC curves of the nomogram and single variables in the model predicting the overall survival after surgery at 0.5 years (**A**) and 1 year (**B**).

Nevertheless, there are some limitations of our study. First, considering the poor prognosis of pancreatic carcinomas, there were not enough patients with an OS over 3 years, meaning that this approach could be inaccurate if we are going to predict the more long-term outcomes from patients with PC. Second, due to the lack of particular clinical data in GSE28735 and GSE62452, we were unable to perform external validation of our nomogram in those GES databases, which means the nomogram should be further validated using multicenter clinical trials and prospective studies.

In conclusion, two genes have been identified as having a prognostic significance in pancreatic carcinoma using a relatively rigid regression model method. For the first time, we report that *ANLN* and *HIST1H1C* are related to the clinical outcome of PC patients. We also constructed a nomogram comprising the prognostic models to assist clinicians treating patients with PC in a personalized way.

MATERIALS AND METHODS

Patients and samples

The mRNA expression and corresponding clinical data of PC patients were obtained from the TCGA dataset that contained 178 PC tissues and 4 adjacent noncancerous pancreatic tissues, which is in accordance with the report from Lu's group [32]. This TCGA dataset was downloaded from Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) at January 2019. The processed mRNA expression data of patients with pancreatic ductal adenocarcinoma were downloaded from two GEO datasets (GSE28735 and GSE62452) [33, 34] that contained the microarray gene-expression profiles of 69 pancreatic tumors and 61 adjacent nontumor tissues as well as the microarray gene-expression profiles of 45 matching pairs of pancreatic tumors and adjacent nontumor tissues from 45 separate patients. We screened potential GEO datasets according to the following inclusion criteria: 1) PC samples and nonmalignant adjacent tissues, 2) studies with more than 45 PC samples or adjacent noncancerous tissues, 2) expression profiling by array, and 3) all samples from Homo sapiens. As afore mentioned [10], the datasets with samples from other organisms or cell lines, those that performed the expression profiling by high-throughput sequencing, those with non-coding RNA profiling by high-throughput sequencing, those with genome variation profiling by genome tiling array, and those with methylation profiling by array or single nucleotide polymorphism genotyping by array were excluded.

RNA-seq data quantification and analysis

Initially, the raw PC mRNA expression profile counts were downloaded from the TCGA database (https://portal.gdc.cancer.gov/), GSE28735 and GSE62452 (https://www.ncbi.nlm.nih.gov/gds). Second, we calculated the DEGs from the TCGA data by means of the limma package and from the GEO data by means of the "edgeR" package. The DEGs from the datasets with a $P<0.01$ were selected and a Venn diagram was plotted. Only the DEGs in all three datasets were considered for subsequent analysis.

Functional enrichment analysis

The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the common DEGs in the three datasets were performed using the R package "clusterProfiler" for Annotation and Visualization.

Constructing the gene-related prognostic model

We initially employed a univariate COX regression analysis of the TCGA datasets to investigate the correlation between patient OS and the expression levels of each gene, which was considered to be significant when the $P<0.05$. We then used the R package "glmnet" to perform the Lasso-penalized Cox regression along with a 10-fold cross-validation to further screen significant genes with prognosis [35, 36]. During the Lasso-penalized Cox regression selection, we randomly subsampled 70% of TCGA patients with 3000 replacements and selected the genes with repeat occurrence frequencies greater than 2500 times. We adopted the largest lambda value such that the error was within one standard error of the minimum, which was called "1-se" lambda [37]. Subsequently, a multivariate Cox regression analysis with a stepwise method was performed to assess the contribution of a gene as an independent prognosis factor for patient survival and to further select the best model. The robustness of all the above-mentioned models was validated by bootstrap procedures [38]. Finally, the risk score (RS) of the mRNA signature was calculated according to the following formula: $\text{RiskSore} = \sum_{i=1}^{n}(coef_i \times Expr_i)$, where Expr was the mRNA expression and coef is the mRNA Lasso coefficient. The optimal cut-off value of RS was determined by X-tile plots version 3.6.1 (Yale University School of Medicine, New Haven, CT, USA) [39]. Based on the cut-off value of RS, patients were classified into the low-risk group and high-risk group, and the K-M curves for the different groups were plotted. Time-dependent receiver operating characteristic (ROC) curve analysis was performed to

evaluate the predictive value of the prediction model. Furthermore, the prognostic gene signature was also validated using the GEO dataset (GSE28735).

Independence of the prognostic gene signature from other clinical characteristics

To prove the independent predictive power of the prognostic gene signature when other clinical variables (such as age, sex, histologic grade, AJCC stage, tumor size and tumor residual) were present in patients with PC, univariate and multivariate Cox regression analyses were performed with the clinical characteristics and gene prognostic model set as independent variables and the OS set as the dependent variable. The robustness of all the above-mentioned models was also validated by bootstrap procedures [38]. All reported P values were two-sided. The hazard ratio (HR) and 95% confidence intervals (CI) were calculated.

Building and validating a predictive nomogram

To improve the generalization ability and applicability of the model, we combined clinical variables with prognostic power with the gene signature model, and we applied the nomogram for the prediction of PC patient prognosis. In this study, the combined model based on all independent prognostic factors selected by the multivariable Cox regression analysis was used to construct a nomogram to assess the probability of the 0.5-, 1-, and 1.5-year OS for patients with PC. Subsequently, the C-index was calculated using R with the "survival" package. Then, to prove that the combined model has an advantage over other single variables, time-dependent ROC curves were obtained using the R package "pROC" [40].

Statistical analyses

The levels of *ANLN* and *HIST1H1C* that were differentially expressed between cancerous and adjacent noncancerous pancreatic tissues were estimated using Student's t-test with SPSS 12.0 software. In addition, the significantly annotated KEGG and GO were calculated in clusterProfiler, and false discovery rate was analyzed using Benjamini correction. Significance of survival analysis was performed by K-M curve with log-rank test. A P value less than 0.05 is considered statistically significant.

AUTHOR CONTRIBUTIONS

Acquisition of Data: SY Zhou and YL Yan. Analysis and Interpretation of Data: J Wei, X Chen, and X Wang. Conception and Design: YL Yan and ZJ Xu. Data Curation: SS Zeng. Development of Methodology: YL Yan, ZC Gong, and ZJ Xu. Writing the Manuscript: SY Zhou and ZJ Xu.

CONFLICTS OF INTEREST

No potential conflicts of interest were disclosed.

FUNDING

This work was supported by the Natural Science Foundation of Hunan Province (No. 2020JJ5934, 2019JJ50932), the National Natural Science Foundation of China (No. 81803035, 81703036), and the China Postdoctoral Science Foundation (No. 2017M610510).

REFERENCES

1. Fu Y, Liu S, Zeng S, Shen H. The critical roles of activated stellate cells-mediated paracrine signaling, metabolism and onco-immunology in pancreatic ductal adenocarcinoma. Mol Cancer. 2018; 17:62. https://doi.org/10.1186/s12943-018-0815-z PMID:29458370

2. Mollinedo F, Gajate C. Novel therapeutic approaches for pancreatic cancer by combined targeting of RAF→MEK→ERK signaling and autophagy survival response. Ann Transl Med. 2019 (Suppl 3); 7:S153. https://doi.org/10.21037/atm.2019.06.40 PMID:31576360

3. Chen X, Xu Z, Zeng S, Wang X, Liu W, Qian L, Wei J, Yang X, Shen Q, Gong Z, Yan Y. The molecular aspect of antitumor effects of protease inhibitor nafamostat mesylate and its role in potential clinical applications. Front Oncol. 2019; 9:852. https://doi.org/10.3389/fonc.2019.00852 PMID:31552177

4. Tesfaye AA, Azmi AS, Philip PA. miRNA and gene expression in pancreatic ductal adenocarcinoma. Am J Pathol. 2019; 189:58–70. https://doi.org/10.1016/j.ajpath.2018.10.005 PMID:30558723

5. Li YJ, Wu JY, Wang JM, Hu XB, Cai JX, Xiang DX. Gemcitabine loaded autologous exosomes for effective and safe chemotherapy of pancreatic cancer. Acta Biomater. 2020; 101:519–30. https://doi.org/10.1016/j.actbio.2019.10.022 PMID:31629893

6. Chiorean EG, Coveler AL. Pancreatic cancer: optimizing treatment options, new, and emerging targeted therapies. Drug Des Devel Ther. 2015; 9:3529–45. https://doi.org/10.2147/DDDT.S60328 PMID:26185420

7. Helm J, Centeno BA, Coppola D, Melis M, Lloyd M, Park JY, Chen DT, Malafa MP. Histologic characteristics

enhance predictive value of American joint committee on cancer staging in resectable pancreas cancer. Cancer. 2009; 115:4080–89. https://doi.org/10.1002/cncr.24503 PMID:19626671

8. Wang L, Xiong L, Wu Z, Miao X, Liu Z, Li D, Zou Q, Yang Z. Expression of UGP2 and CFL1 expression levels in benign and Malignant pancreatic lesions and their clinicopathological significance. World J Surg Oncol. 2018; 16:11. https://doi.org/10.1186/s12957-018-1316-7 PMID:29347944

9. Song J, Xu Q, Zhang H, Yin X, Zhu C, Zhao K, Zhu J. Five key lncRNAs considered as prognostic targets for predicting pancreatic ductal adenocarcinoma. J Cell Biochem. 2018; 119:4559–69. https://doi.org/10.1002/jcb.26598 PMID:29239017

10. Yan Y, Xu Z, Qian L, Zeng S, Zhou Y, Chen X, Wei J, Gong Z. Identification of CAV1 and DCN as potential predictive biomarkers for lung adenocarcinoma. Am J Physiol Lung Cell Mol Physiol. 2019; 316:L630–43. https://doi.org/10.1152/ajplung.00364.2018 PMID:30604627

11. Liu B, Liu J, Liu K, Huang H, Li Y, Hu X, Wang K, Cao H, Cheng Q. A prognostic signature of five pseudogenes for predicting lower-grade gliomas. Biomed Pharmacother. 2019; 117:109116. https://doi.org/10.1016/j.biopha.2019.109116 PMID:31247469

12. Jiang Y, Mei W, Gu Y, Lin X, He L, Zeng H, Wei F, Wan X, Yang H, Major P, Tang D. Construction of a set of novel and robust gene expression signatures predicting prostate cancer recurrence. Mol Oncol. 2018; 12:1559–78. https://doi.org/10.1002/1878-0261.12359 PMID:30024105

13. Cheng Q, Huang C, Cao H, Lin J, Gong X, Li J, Chen Y, Tian Z, Fang Z, Huang J. A novel prognostic signature of transcription factors for the prediction in patients with GBM. Front Genet. 2019; 10:906. https://doi.org/10.3389/fgene.2019.00906 PMID:31632439

14. Song M, Sun M, Xia L, Chen W, Yang C. miR-19b-3p promotes human pancreatic cancer capan-2 cells proliferation by targeting phosphatase and tension homolog. Ann Transl Med. 2019; 7:236. https://doi.org/10.21037/atm.2019.04.61 PMID:31317006

15. Subedi P, Nembrini S, An Q, Zhu Y, Peng H, Yeh F, Cole SA, Rhoades DA, Lee ET, Zhao J. Telomere length and cancer mortality in American Indians: the strong heart study. Geroscience. 2019; 41:351–61. https://doi.org/10.1007/s11357-019-00080-4 PMID:31230193

16. Ducreux M, Seufferlein T, Van Laethem JL, Laurent-Puig P, Smolenschi C, Malka D, Boige V, Hollebecque A, Conroy T. Systemic treatment of pancreatic cancer revisited. Semin Oncol. 2019; 46:28–38. https://doi.org/10.1053/j.seminoncol.2018.12.003 PMID:30638624

17. Balaban EP, Mangu PB, Khorana AA, Shah MA, Mukherjee S, Crane CH, Javle MM, Eads JR, Allen P, Ko AH, Engebretson A, Herman JM, Strickler JH, et al. Locally Advanced, Unresectable Pancreatic Cancer: American Society of Clinical Oncology Clinical Practice Guideline. J Clin Oncol. 2016; 34:2654–68. https://doi.org/10.1200/JCO.2016.67.5561 PMID:27247216

18. Tempero MA, Malafa MP, Al-Hawary M, Asbun H, Bain A, Behrman SW, Benson AB 3rd, Binder E, Cardin DB, Cha C, Chiorean EG, Chung V, Czito B, et al. Pancreatic adenocarcinoma, version 2.2017, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw. 2017; 15:1028–61. https://doi.org/10.6004/jnccn.2017.0131 PMID:28784865

19. Reni M, Zanon S, Balzano G, Nobile S, Pircher CC, Chiaravalli M, Passoni P, Arcidiacono PG, Nicoletti R, Crippa S, Slim N, Doglioni C, Falconi M, Gianni L. Selecting patients for resection after primary chemotherapy for non-metastatic pancreatic adenocarcinoma. Ann Oncol. 2017; 28:2786–92. https://doi.org/10.1093/annonc/mdx495 PMID:28945895

20. Chang JC, Kundranda M. Novel diagnostic and predictive biomarkers in pancreatic adenocarcinoma. Int J Mol Sci. 2017; 18:667. https://doi.org/10.3390/ijms18030667 PMID:28335509

21. Sun L, Burnett J, Guo C, Xie Y, Pan J, Yang Z, Ran Y, Sun D. CPA4 is a promising diagnostic serum biomarker for pancreatic cancer. Am J Cancer Res. 2015; 6:91–96. PMID:27073726

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

23. Wang G, Shen W, Cui L, Chen W, Hu X, Fu J. Overexpression of anillin (ANLN) is correlated with colorectal cancer progression and poor prognosis. Cancer Biomark. 2016; 16:459–65. https://doi.org/10.3233/CBM-160585 PMID:27062703

24. Zhang M, Wang F, Xiang Z, Huang T, Zhou WB. LncRNA XIST promotes chemoresistance of breast cancer cells to doxorubicin by sponging miR-200c-3p to upregulate ANLN. Clin Exp Pharmacol Physiol. 2020; 47:1464–72. https://doi.org/10.1111/1440-1681.13307 PMID:32198770

25. Wang Z, Chen J, Zhong MZ, Huang J, Hu YP, Feng DY, Zhou ZJ, Luo X, Liu ZQ, Jiang WZ, Zhou WB. Overexpression of ANLN contributed to poor prognosis of anthracycline-based chemotherapy in breast cancer patients. Cancer Chemother Pharmacol. 2017; 79:535–43. https://doi.org/10.1007/s00280-017-3248-2 PMID:28243684

26. Zhou W, Wang Z, Shen N, Pi W, Jiang W, Huang J, Hu Y, Li X, Sun L. Knockdown of ANLN by lentivirus inhibits cell growth and migration in human breast cancer. Mol Cell Biochem. 2015; 398:11–19. https://doi.org/10.1007/s11010-014-2200-6 PMID:25223638

27. Lian YF, Huang YL, Wang JL, Deng MH, Xia TL, Zeng MS, Chen MS, Wang HB, Huang YH. Anillin is required for tumor growth and regulated by miR-15a/miR-16-1 in HBV-related hepatocellular carcinoma. Aging (Albany NY). 2018; 10:1884–901. https://doi.org/10.18632/aging.101510 PMID:30103211

28. Long X, Zhou W, Wang Y, Liu S. Prognostic significance of ANLN in lung adenocarcinoma. Oncol Lett. 2018; 16:1835–40. https://doi.org/10.3892/ol.2018.8858 PMID:30008873

29. Magnusson K, Gremel G, Rydén L, Pontén V, Uhlén M, Dimberg A, Jirström K, Pontén F. ANLN is a prognostic biomarker independent of ki-67 and essential for cell cycle progression in primary breast cancer. BMC Cancer. 2016; 16:904. https://doi.org/10.1186/s12885-016-2923-8 PMID:27863473

30. Zeng S, Yu X, Ma C, Song R, Zhang Z, Zi X, Chen X, Wang Y, Yu Y, Zhao J, Wei R, Sun Y, Xu C. Transcriptome sequencing identifies ANLN as a promising prognostic biomarker in bladder urothelial carcinoma. Sci Rep. 2017; 7:3151. https://doi.org/10.1038/s41598-017-02990-9 PMID:28600503

31. Song D, Chaerkady R, Tan AC, García-García E, Nalli A, Suárez-Gauthier A, López-Ríos F, Zhang XF, Solomon A, Tong J, Read M, Fritz C, Jimeno A, et al. Antitumor activity and molecular effects of the novel heat shock protein 90 inhibitor, IPI-504, in pancreatic cancer. Mol Cancer Ther. 2008; 7:3275–84. https://doi.org/10.1158/1535-7163.MCT-08-0508 PMID:18852131

32. Lu H, Niu F, Liu F, Gao J, Sun Y, Zhao X. Elevated glypican-1 expression is associated with an unfavorable prognosis in pancreatic ductal adenocarcinoma. Cancer Med. 2017; 6:1181–91. https://doi.org/10.1002/cam4.1064 PMID:28440066

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

34. Yang S, He P, Wang J, Schetter A, Tang W, Funamizu N, Yanaga K, Uwagawa T, Satoskar AR, Gaedcke J, Bernhardt M, Ghadimi BM, Gaida MM, et al. A novel MIF signaling pathway drives the Malignant character of pancreatic cancer by targeting NR3C2. Cancer Res. 2016; 76:3838–50. https://doi.org/10.1158/0008-5472.CAN-15-2841 PMID:27197190

35. Huang JL, Fu YP, Jing CY, Yi Y, Sun J, Gan W, Lu ZF, Zhou J, Fan J, Qiu SJ. A novel and validated prognostic nomogram based on liver fibrosis and tumor burden for patients with hepatocellular carcinoma after curative resection. J Surg Oncol. 2018; 117:625–33. https://doi.org/10.1002/jso.24895 PMID:29165812

36. Wu J, Zhou L, Huang L, Gu J, Li S, Liu B, Feng J, Zhou Y. Nomogram integrating gene expression signatures with clinicopathological features to predict survival in operable NSCLC: a pooled analysis of 2164 patients. J Exp Clin Cancer Res. 2017; 36:4. https://doi.org/10.1186/s13046-016-0477-x PMID:28057025

37. Wong KK, Rostomily R, Wong ST. Prognostic gene discovery in glioblastoma patients using deep learning. Cancers (Basel). 2019; 11:53. https://doi.org/10.3390/cancers11010053 PMID:30626092

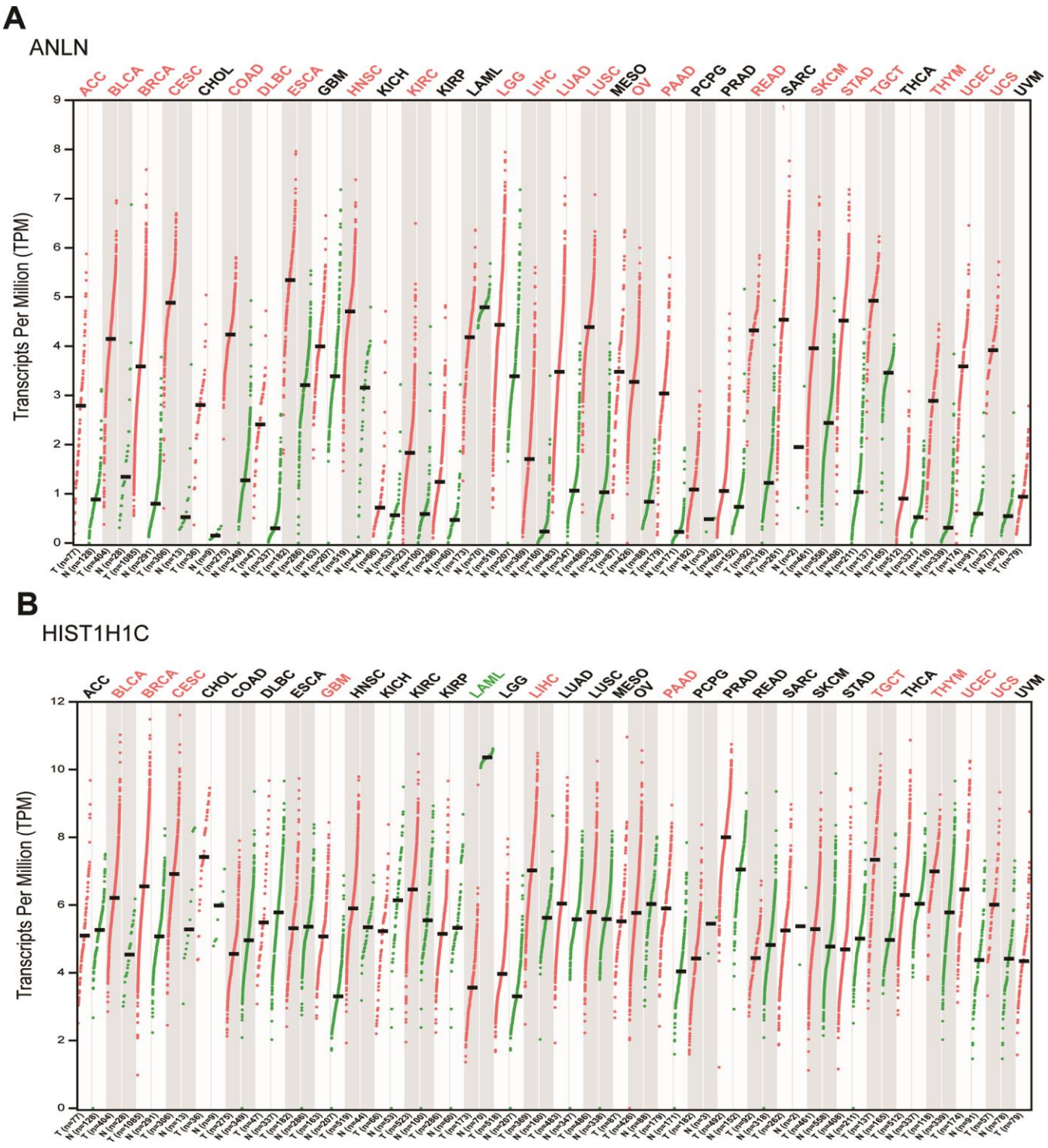
38. Mavridis K, Gueugnon F, Petit-Courty A, Courty Y, Barascu A, Guyetant S, Scorilas A. The oncomiR miR-197 is a novel prognostic indicator for non-small cell lung cancer patients. Br J Cancer. 2015; 112:1527–35. https://doi.org/10.1038/bjc.2015.119 PMID:25867273

39. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. Clin Cancer Res. 2004; 10:7252–59. https://doi.org/10.1158/1078-0432.CCR-04-0713 PMID:15534099

40. Huitzil-Melendez FD, Capanu M, O'Reilly EM, Duffy A, Gansukh B, Saltz LL, Abou-Alfa GK. Advanced hepatocellular carcinoma: which staging systems best predict prognosis? J Clin Oncol. 2010; 28:2889–95. https://doi.org/10.1200/JCO.2009.25.9895 PMID:20458042

SUPPLEMENTARY MATERIALS

Supplementary Figure



Supplementary Figure 1. The pancancer analysis of *ANLN* and *HIST1H1C* expression across multiple cancer types in GEPIA 2.0 database. Red tumor abbreviations indicated significant upregulation of *ANLN* and *HIST1H1C* in cancers, while green tumor abbreviations indicated significant downregulation of *ANLN* and *HIST1H1C.*

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Tables 1, 2, 3, 4, 6, 7, 8, 10.

Supplementary Table 5. 3000 times lasso cox regression genes.

	Gene	Times
1	ECT2	316
2	ANLN	2640
3	HIST1H2BD	0
4	NEK2	0
5	HIST1H2AC	326
6	CD36	1692
7	HIST1H2BJ	0
8	ZMAT1	0
9	CHEK1	0
10	KIF4A	0
11	CDC6	0
12	BCL11A	0
13	HIST1H2BC	0
14	CLK4	0
15	HIST1H1C	2543
16	HIST1H2BK	0
17	PLCG2	312
18	MAP3K14	1
19	GSDMC	0
20	SCML4	0
21	ZBTB32	0
22	HRASLS2	0
23	HIST1H4I	0
24	GDPD5	30
25	GYPC	0
26	RNF166	0
27	CCM2L	0
28	LIMD2	948
29	OXER1	0
30	IL16	1673
31	PSTPIP1	0
32	AGER	0
33	VENTX	0
34	GH1	0
35	BTNL9	0

Supplementary Table 9. The expression of ANLN and HIST1H1C in PC cells from Cancer Cell Line Encyclopedia (CCLE).

Gene	ANLN	HIST1H1C
SNU410_PANCREAS	7.595567	5.063576
PANC0403_PANCREAS	7.45979	3.238598
PANC1_PANCREAS	7.032564	4.064839
PK45H_PANCREAS	6.952955	5.389367
SW1990_PANCREAS	6.893646	4.444045
PATU8902_PANCREAS	6.839072	3.806673
SNU213_PANCREAS	6.79091	3.951536
KP2_PANCREAS	6.65843	5.718855
HS766T_PANCREAS	6.294668	6.366739
PANC0203_PANCREAS	6.291476	4.468557
PANC1005_PANCREAS	6.249681	5.16116
PATU8988T_PANCREAS	6.243406	4.120279
T3M4_PANCREAS	6.232112	6.643391
KP3_PANCREAS	6.17958	4.369459
PK1_PANCREAS	6.102597	3.550379
PANC0327_PANCREAS	6.077964	6.297954
YAPC_PANCREAS	6.05971	3.701167
HUPT3_PANCREAS	6.056466	6.20422
KP4_PANCREAS	5.910993	6.217787
PANC0504_PANCREAS	5.89602	4.113109
SUIT2_PANCREAS	5.895513	3.811658
TCCPAN2_PANCREAS	5.875302	6.281926
PANC0213_PANCREAS	5.815443	4.933527
CAPAN2_PANCREAS	5.763245	3.244718
MIAPACA2_PANCREAS	5.712785	5.638969
PANC0813_PANCREAS	5.665741	4.26619
SNU324_PANCREAS	5.623243	4.435857
PK59_PANCREAS	5.584925	4.309652
CAPAN1_PANCREAS	5.458695	6.691079
BXPC3_PANCREAS	5.431424	5.106647
CFPAC1_PANCREAS	5.410521	6.612041
HPAFII_PANCREAS	5.405624	4.803013
L33_PANCREAS	5.322749	4.259711
HPAC_PANCREAS	5.19894	3.818123
HUPT4_PANCREAS	5.196609	3.970606
DANG_PANCREAS	5.165471	3.057551
PATU8988S_PANCREAS	5.028114	4.90915
PSN1_PANCREAS	4.380796	3.110652
ASPC1_PANCREAS	4.184791	3.775559
SU8686_PANCREAS	3.933033	2.336863
QGP1_PANCREAS	3.925441	4.861757

Supplementary Table 11. The two-gene expression risk stratification in GSE28735.

	ID	HIST1H1C	ANLN	Surtime	Surstat	Riskscore	Groupby
1	GSM711904	4.59351	3.478818	4.25	1	0.401283	Low risk
2	GSM711906	5.383048	6.50617	0.58	1	1.851105	High risk
3	GSM711908	4.642691	6.541408	0.25	1	3.62942	High risk
4	GSM711910	3.707255	2.865696	3.5	1	0.555705	Low risk
5	GSM711914	4.629382	5.674897	3	1	1.945416	High risk
6	GSM711916	4.828843	4.864498	0.17	1	0.902076	Low risk
7	GSM711922	4.808846	4.835817	1.58	1	0.89889	Low risk
8	GSM711924	4.515929	5.470503	1.08	1	1.849385	High risk
9	GSM711926	4.718713	4.651367	1.33	1	0.849599	Low risk
10	GSM711928	4.044202	4.589405	3.42	1	1.464419	Low risk
11	GSM711930	4.380305	5.038249	0.25	1	1.516759	High risk
12	GSM711932	4.133895	6.320266	1	1	4.815825	High risk
13	GSM711934	4.482482	4.605157	2.08	1	1.009769	Low risk
14	GSM711936	4.709863	2.9729	3.17	0	0.250142	Low risk
15	GSM711938	4.472675	2.621358	0.08	0	0.237873	Low risk
16	GSM711940	5.041278	5.580661	1.08	1	1.266404	Low risk
17	GSM711942	4.77368	5.394913	1.92	1	1.396585	Low risk
18	GSM711944	4.523715	5.181269	0.92	1	1.485881	Low risk
19	GSM711946	3.95196	3.259936	2.42	1	0.598995	Low risk
20	GSM711948	4.34825	3.772391	2.33	1	0.616691	Low risk
21	GSM711950	4.441237	3.769534	2.33	0	0.567338	Low risk
22	GSM711952	5.284748	5.607847	1.17	1	1.044138	Low risk
23	GSM711954	4.415777	5.321105	0.58	1	1.809254	High risk
24	GSM711956	5.069863	4.410287	2.33	0	0.523712	Low risk
25	GSM711958	5.338916	5.728907	2	0	1.088248	Low risk
26	GSM711960	4.75965	4.458094	2	0	0.71143	Low risk
27	GSM711962	4.692492	6.405433	0.67	1	3.145077	High risk
28	GSM711964	4.87009	5.140327	1.83	0	1.065114	Low risk
29	GSM711966	5.500636	5.748288	1.75	0	0.958246	Low risk
30	GSM711968	4.71547	5.310038	1.75	0	1.380871	Low risk
31	GSM711970	5.056518	4.657617	0.75	1	0.635193	Low risk
32	GSM711972	5.393925	6.276303	1.42	0	1.549226	High risk
33	GSM711974	4.565349	5.451879	0.5	1	1.747143	High risk
34	GSM711976	4.734871	4.955239	1.33	0	1.046703	Low risk
35	GSM711978	4.637025	4.249571	0.42	1	0.679697	Low risk
36	GSM711980	4.19272	4.349029	0.92	0	1.07827	Low risk
37	GSM711982	4.914539	6.970197	0.33	1	3.918385	High risk
38	GSM711984	4.047127	2.808281	0.83	1	0.395805	Low risk
39	GSM711986	5.846707	5.333412	0.83	0	0.522338	Low risk
40	GSM711988	4.780226	4.680945	1.25	1	0.82275	Low risk
41	GSM711990	4.518276	3.429367	0.42	1	0.413314	Low risk
42	GSM711992	4.661236	4.14586	1.08	1	0.616738	Low risk

Supplementary Table 12. The two-gene expression risk stratification in GSE62452.

	Sample	Fultime	Fustate	Type	Express_ANLN	Express_HIST1HIC	Riskscore	Groupby
1	GSM1527105	4.258333	1	Tumor	3.28581	4.15587	0.69245	Low risk
2	GSM1527107	0.575	1	Tumor	6.37726	4.98483	3.641451	High risk
3	GSM1527109	0.225	1	Tumor	6.27981	4.12032	1.334185	High risk
4	GSM1527111	3.466667	1	Tumor	2.72845	3.3195	0.235374	Low risk
5	GSM1527115	2.991667	1	Tumor	5.47523	4.14786	1.141673	High risk
6	GSM1527117	0.2	1	Tumor	4.65815	4.3633	1.20571	High risk
7	GSM1527123	1.625	1	Tumor	4.63507	4.33638	1.163162	High risk
8	GSM1527125	1.05	1	Tumor	5.27076	4.1233	1.058728	High risk
9	GSM1527127	1.333333	1	Tumor	4.42231	4.23132	0.982557	High risk
10	GSM1527129	3.408333	1	Tumor	4.39364	3.68734	0.526347	Low risk
11	GSM1527131	0.233333	1	Tumor	4.76962	3.9566	0.779813	Low risk
12	GSM1527133	0.966667	1	Tumor	6.14909	3.71616	0.818008	Low risk
13	GSM1527135	2.058333	1	Tumor	4.45058	4.10903	0.860839	Low risk
14	GSM1527137	3.333333	0	Tumor	2.82307	4.17296	0.633992	Low risk
15	GSM1527139	0.1	1	Tumor	2.48363	3.94183	0.450665	Low risk
16	GSM1527141	1.1	1	Tumor	5.38821	4.62335	1.919456	High risk
17	GSM1527143	1.933333	1	Tumor	5.20154	4.40832	1.439838	High risk
18	GSM1527145	0.9	1	Tumor	5.02687	4.06697	0.93839	Low risk
19	GSM1527147	2.416667	1	Tumor	3.0661	3.52321	0.32084	Low risk
20	GSM1527149	2.308333	1	Tumor	3.55992	3.9235	0.566902	Low risk
21	GSM1527151	2.3	0	Tumor	3.56466	3.98591	0.609194	Low risk
22	GSM1527154	1.15	1	Normal	2.30865	3.24949	0.197173	Low risk
23	GSM1527155	0.566667	1	Tumor	5.07569	4.01034	0.890008	Low risk
24	GSM1527157	2.35	0	Tumor	4.21571	4.71055	1.613475	High risk
25	GSM1527159	0.816667	1	Tumor	5.63999	4.9959	3.106545	High risk
26	GSM1527161	1.966667	0	Tumor	4.26289	4.30718	1.03195	High risk
27	GSM1527163	0.641667	1	Tumor	6.21958	4.25898	1.539909	High risk
28	GSM1527165	1.816667	0	Tumor	4.95109	4.51506	1.533367	High risk
29	GSM1527167	1.766667	0	Tumor	5.56318	5.1566	3.662255	High risk
30	GSM1527169	1.758333	0	Tumor	5.08181	4.21906	1.129573	High risk
31	GSM1527171	0.741667	1	Tumor	4.47359	4.67485	1.645144	High risk
32	GSM1527173	1.441667	0	Tumor	6.17811	5.04493	3.722144	High risk
33	GSM1527175	0.533333	1	Tumor	5.21309	4.07266	0.986266	High risk
34	GSM1527177	1.366667	0	Tumor	4.75762	4.28066	1.123428	High risk
35	GSM1527179	0.383333	1	Tumor	4.04649	4.23222	0.901257	Low risk
36	GSM1527181	0.883333	0	Tumor	4.1425	3.71913	0.514734	Low risk
37	GSM1527183	0.35	1	Tumor	6.79223	4.4891	2.284431	High risk
38	GSM1527185	0.858333	1	Tumor	2.80611	3.64402	0.346414	Low risk
39	GSM1527187	0.808333	0	Tumor	5.15503	5.44705	4.631561	High risk
40	GSM1527189	1.241667	1	Tumor	4.50642	4.38521	1.193225	High risk
41	GSM1527191	0.375	1	Tumor	3.26562	4.09829	0.645597	Low risk
42	GSM1527193	1.075	1	Tumor	3.95943	4.3022	0.956221	Low risk
43	GSM1527196	0.791667	1	Tumor	4.54332	3.84922	0.654933	Low risk
44	GSM1527198	0.525	1	Tumor	5.90911	4.32916	1.551464	High risk
45	GSM1527199	0.075	1	Tumor	4.35018	4.08732	0.820502	Low risk
46	GSM1527200	0.491667	1	Tumor	3.4362	3.42219	0.311788	Low risk
47	GSM1527202	0.816667	1	Tumor	4.83738	4.44265	1.375509	High risk
48	GSM1527204	0.441667	1	Tumor	6.22746	4.1379	1.344611	High risk

49	GSM1527205	1.791667	1	Tumor	3.64831	4.32354	0.911297	Low risk
50	GSM1527207	1.183333	1	Tumor	6.01381	4.34821	1.624452	High risk
51	GSM1527209	2.666667	1	Tumor	5.50742	4.31662	1.393137	High risk
52	GSM1527210	1.908333	1	Tumor	5.93411	5.57191	6.39679	High risk
53	GSM1527212	3.825	1	Tumor	3.47327	3.88402	0.531243	Low risk
54	GSM1527213	1.825	1	Tumor	5.24938	3.28717	0.407756	Low risk
55	GSM1527215	3.5	0	Tumor	3.08215	3.39736	0.279167	Low risk
56	GSM1527216	3.191667	0	Tumor	3.19303	3.81927	0.462455	Low risk
57	GSM1527218	1.141667	1	Tumor	7.66299	4.93645	4.647977	High risk
58	GSM1527219	0.908333	1	Tumor	5.13991	4.24769	1.18276	High risk
59	GSM1527220	1.775	1	Tumor	6.3008	4.59253	2.291621	High risk
60	GSM1527223	0.775	1	Tumor	3.63132	3.69955	0.447	Low risk
61	GSM1527225	1.658333	1	Tumor	3.98875	4.95504	2.020132	High risk
62	GSM1527227	5.9	0	Tumor	3.59565	3.98568	0.61344	Low risk
63	GSM1527228	5.666667	0	Tumor	3.56692	4.14986	0.734199	Low risk
64	GSM1527230	5.641667	0	Tumor	3.65724	4.28682	0.875908	Low risk
65	GSM1527232	4.141667	1	Tumor	3.11183	4.27017	0.75713	Low risk
66	GSM1527234	0.266667	1	Tumor	3.60256	4.10509	0.703624	Low risk

Supplementary Table 13. The C-index of nomogram model.

Univariate analysis					
Age(36-84)	1.00	1.05	0.018	1.03	1.03(1.00,1.05)
Sex(Male/Female)	0.56	1.42	0.628	0.99	0.99(0.56,1.42)
Grade(G1-2/G3-4)	0.60	1.59	0.929	1.10	1.10(0.60,1.59)
Stage(III-IV/I-II)	0.17	2.07	0.417	1.12	1.12(0.17,2.07)
Size(1.8-12)	0.92	1.34	0.259	1.13	1.13(0.92,1.34)
Residual(R0/non-R0)	0.39	1.01	0.054	0.70	0.70(0.39,1.01)
Prognostic model()	1.43	2.16	0.000	1.80	1.80(1.43,2.16)
Multivariate analysis		0.000			
Age	1.00	1.05	0.021	1.03	1.03(1.00,1.05)
Residual	0.40	1.03	0.067	0.72	0.72(0.40,1.03)
Riskscore	1.48	2.21	0.000	1.85	1.85(1.48,2.21)

Items	**C-index**
Age	0.55
Residual	0.577
riskscore	0.647
nomogram model	0.664
ANLN	0.633
HIST1H1C	0.553