An integrated mRNA-lncRNA signature for overall survival prediction in cholangiocarcinoma

Citation for published version:
Xu, D, Wei, L, Zeng, L, Mukiibi, R, Xin, H & Zhang, F 2023, 'An integrated mRNA-lncRNA signature for overall survival prediction in cholangiocarcinoma', *Medicine*, vol. 102, no. 39, e35348, pp. 1-11. https://doi.org/10.1097/MD.0000000000035348

Digital Object Identifier (DOI):
10.1097/MD.0000000000035348

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Medicine

Observational Study

Medicine

OPEN

An integrated mRNA–lncRNA signature for overall survival prediction in cholangiocarcinoma

Derong Xu, PhD[a], Lili Wei, BS[b], Liping Zeng, MS[a], Robert Mukiibi, PhD[c], Hongbo Xin, PhD[a], Feng Zhang, PhD[a,*]

Abstract

The combination of mRNA and lncRNA profiles for establishing an integrated mRNA–lncRNA prognostic signature has remained unexplored in cholangiocarcinoma (CCA) patients. We utilized a training dataset of 36 samples from The Cancer Genome Atlas dataset and a validation cohort (GSE107943) of 30 samples from Gene Expression Omnibus. Two mRNAs (*CFHR3* and *PIWIL4*) and 2 lncRNAs (*AC007285.1* and *AC134682.1*) were identified to construct the integrated signature through a univariate Cox regression (*P*-value = 1.35E−02) and a multivariable Cox analysis (*P*-value = 3.07E−02). Kaplan–Meier curve showed that patients with low risk scores had notably prolonged overall survival than those with high risk scores (*P*-value = 4.61E−03). Subsequently, the signature was validated in GSE107943 cohort with an area under the curve of 0.750 at 1-year and 0.729 at 3-year. The signature was not only independent from diverse clinical features (*P*-value = 3.07E−02), but also surpassed other clinical characteristics as prognostic biomarkers with area under the curve of 0.781 at 3-year. Moreover, the weighted gene co-expression network analysis and gene enrichment analyses found that the integrated signature were associated with metabolic-related biological process and lipid metabolism pathway, which has been implicated in the pathogenesis of CCA. Taken together, we developed an integrated mRNA–lncRNA signature that had an independent prognostic value in the risk stratification of patients with CCA.

Abbreviations: AUC = area under the curve, CCA = cholangiocarcinoma, CI = confidence interval, DERNAs = differentially expressed RNAs, GEO = Gene Expression Omnibus, GO = gene ontology, ICCA = intrahepatic cholangiocarcinoma, KEGG = Kyoto Encyclopedia of Genes and Genomes, KM = Kaplan–Meier, OS = overall survival, ROC = receiver operating characteristic, SMD = standard mean difference, TCGA = The Cancer Genome Atlas, TOM = the topological overlap matrix, WGCNA = weighted gene co-expression network analysis.

Keywords: cholangiocarcinoma, proportional hazards models, risk assessment, survival

1. Introduction

Cholangiocarcinoma (CCA) is an aggressive biliary epithelial malignancy arising from within the liver termed as intrahepatic cholangiocarcinoma (ICCA) or more commonly from the extrahepatic bile ducts known as extrahepatic cholangiocarcinoma.[1] According to epidemiological reports in the past decades, CCA is the second most common primary hepatic neoplasm, and its incidence and mortality rate have been rising globally.[2–4] Although surgical resection is a promising curative treatment for CCA patients, only a minority of patients (about 25%) in the early stage are eligible for surgery.[5] Most of the patients are diagnosed at an advanced stage due to the characteristics of asymptomatic disease. These advanced patients usually have worse prognosis with a median overall survival (OS) of 12–15 months.[2,3,6,7] Furthermore, due to molecular heterogeneity and complex etiology of CCA, the commonly used tumor-node-metastasis staging system has shown valuable but insufficient accuracy for prognostic evaluation.[8] Therefore, there is an urgent and critical need to develop novel and more accurate prognostic signatures for CCA patients to distinguish risk stratification and consequently contribute to personalized management and follow-up plans.

Numerous literatures have documented the significant involvement of dysregulated mRNAs and lncRNAs in the initiation and progression of CCA.[9–11] For instance, YTHDF2 has been demonstrated to promote ICCA progression by increasing CDKN1B mRNA degradation.[12] Similarly, MUC13 has been

DX and LW contributed equally to this work.

This study was supported by the Natural Science Foundation of JiangXi Province (20212BAB216077), the Natural Science Foundation of Chongqing City (CSTB2023NSCQ-MSX0780) and the National Natural Science Foundation of China (31601034). The funding body played no role in the design of the study and data collection, analysis, and interpretation as well as drafting the manuscript.

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are publicly available.

Posted history: This manuscript was previously posted to Research Square: doi: https://www.researchsquare.com/article/rs-99682/v1.

[a] The National Engineering Research Center for Bioengineering Drugs and the Technologies, The Institute of Translational Medicine, Nanchang University, Nanchang, China, [b] Department of Oral and Maxillofacial Surgery, The First Affiliated Hospital of Nanchang University, Nanchang, China, [c] The Roslin Institute, University of Edinburgh, Midlothian, United Kingdom.

[*] Correspondence: Feng Zhang, The National Engineering Research Center for Bioengineering Drugs and the Technologies, The Institute of Translational Medicine, Nanchang University, Nanchang 330031, China (e-mail: fengzhang0709@hotmail.com).

How to cite this article: Xu D, Wei L, Zeng L, Mukiibi R, Xin H, Zhang F. An integrated mRNA–lncRNA signature for overall survival prediction in cholangiocarcinoma. Medicine 2023;102:39(e35348).

Received: 31 August 2022 / Received in final form: 10 August 2023 / Accepted: 1 September 2023

http://dx.doi.org/10.1097/MD.0000000000035348

implicated in accelerating ICCA progression via EGFR/PI3K/AKT pathways.[13] JUND/linc00976 have been found to promote progression and metastasis of CCA by regulating the miR-3202/GPX4 axis.[14] Additionally, the upregulation of HCG18 has been observed to expedite growth and metastasis of CCA through mediating miR-424-5p/SOX9 axis.[15] Furthermore, mRNAs and lncRNAs were reported to exhibit distinct tissue-specific expression patterns and correlate strongly with development and progression of CCA, which offer valuable insights into CCA prognosis. Ruys et al reported 77 prognostic biomarkers using an immunohistochemical analysis.[16] A three-miRNA signature for prognosis[17] and a 7-mRNA biomarker for recurrence-free survival prediction[18] have been identified from global transcriptome profile of CCA patients. While combined mRNA and lncRNA signatures have demonstrated substantial prognostic value across various cancers,[19,20] a comprehensive investigation into the potential integration of mRNAs and lncRNAs expression across whole transcriptome for predicting overall survival in CCA remains unexplored.

To discover potentially novel biomarkers with higher prognostic prediction of CCA, we initially identified the differentially expressed mRNAs and lncRNAs between cholangiocarcinoma and normal tissues by analyzing high-throughput data downloaded from The Cancer Genome Atlas (TCGA) database. We then developed an independent mRNA–lncRNA signature using a univariate Cox regression analysis and a stepwise multivariable Cox analysis for the identified mRNAs and lncRNAs. Moreover, we also evaluated expression levels of the detected biomarkers across various datasets using meta-analysis, and assessed prognostic performance of the signature in an external dataset (GSE107943) as an independent biomarker. Finally, the module eigengene related to prognostic RNAs were determined by weighted gene co-expression network analysis (WGCNA), and biological functions related to the signature were investigated through gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis. Taken together, the mRNA–lncRNA signature identified in the current study would contribute to improve prognosis accuracy, and thus facilitate individualized management in CCA patients.

2. Materials and methods

2.1. CCA datasets and patient information

The training data termed as TCGA-CHOL contained 45 samples (36 CCA tumor tissues and 9 adjacent normal tissues) that were obtained from the TCGA portal on July 3, 2019.[21] The external validation cohort (Accession: GSE107943)[22] with 30 CCA tumor tissues and 27 adjacent normal tissues was retrieved from Gene Expression Omnibus (GEO) database.[23] All samples from the 2 cohorts were sequenced across the whole transcriptome by RNA-seq high-throughput sequencing platform. The read counts and corresponding clinical information were publicly available and used in the current study. The samples in both training and validation datasets possessed mRNA and lncRNA expression data as well as complete survival information including survival status, survival time, and classic clinicopathological features. The data collection and processing followed the publication guidelines provided by TCGA (http://cancergenome.nih.gov/publications/publicationguidelines) and GEO database, thus ethical approval was not required for this study.

2.2. Screening of differentially expressed RNAs

The expression profiles of 20,271 mRNAs and 14,852 lncRNAs in total were obtained from TCGA database. We then re-annotated all RNAs in training and validation cohorts based on Gencode V30 (https://www.gencodegenes.org/). After removing the RNAs with mean expression value lower than one and a median read counts equal to zero across all samples, we obtained 17,010 mRNAs and 6390 lncRNAs for differential expression RNAs screening. The edgeR[24] and DESeq2[25] R packages were independently utilized for detecting the differentially expressed RNAs (DERNAs) between tumor tissues and normal or adjacent tissues of CCA patients. We adjusted the *P*-value by false discovery rate as proposed by the Benjamini-Hochberg procedure to limit the occurrence rate of false positives.[26] The RNAs were considered as differently expressed at a threshold of $|\log_2(\text{fold change})| \geq 1.5$ and false discovery rate < 0.05. The overlapped DERNAs identified by edgeR and DESeq2 packages were then considered for further downstream analyses.

2.3. The mRNA–lncRNA signature construction and validation

To uncover candidate prognostic RNAs related to OS of CCA patients, we performed univariate Cox regression analysis for the overlapped DERNAs through the survival R package. The candidate DERNAs with significant *P*-values (*P*-value < 0.01) were then subjected to a stepwise multivariable Cox analysis to select the optimal combination for predicting survival outcome. Subsequently, we combined the expression levels of those selected RNAs with the multivariable Cox regression coefficients as weights to construct a risk score model for each patient. The formula was listed as follows:

$$\text{risk} \quad \text{score} = \sum_{i=1}^{n} coefficient\,(RNA_i) * expression\,(RNA_i)\,.$$

where n is the total number of molecules used to calculate the risk scores including mRNAs and lncRNAs, coefficient (RNA_i) corresponds to multivariable Cox regression coefficient of the *i*th RNA, and expression (RNA_i) represents expression level of the *i*th RNA. To confirm the expression level of selected markers among other dependent datasets, we retrieved the whole GEO database and collected dataset meeting 2 criteria: firstly, it possessed mRNA or lncRNA expression data of normal or adjacent tissues; secondly, their sample size was more than 3. The 6 datasets (GEO accessions: GSE26566,[27] GSE57555,[28] GSE31370,[29] GSE76297,[30] and GSE32879[31]) with a total of 413 samples that included 228 tumor tissues and 185 normal or adjacent tissues. However, there was no appropriate cohort for confirmation of lncRNAs expression status. Comprehensive meta-analyses for the selected cohorts were performed by the meta R package.[32] The inconsistency (I^2) test and the Cochran Q test were utilized to assess heterogeneity. When I^2 was >50% or *P*-value was lower than .01, the random effect model was applied, otherwise, the fixed effect model was implemented to weight the standard mean difference (SMD). Finally, the overall SMD and a 95% confidence interval (CI) were employed to measure the general expression differences of selected biomarkers across various CCA groups.

The optimal cutoff selection of risk scores to classify CCA patients into the high or low risk score groups in the training dataset was determined by "cutp-function" from the survMisc package in R. The hazard function would choose cut point with the maximal sensitivity and specificity for survival rate.[33] Kaplan–Meier (KM) method was applied to evaluate survival differences between the high-risk and low-risk groups, and statistical significance was obtained by the log-rank test. Time-dependent receiver operating characteristic (ROC) analysis was conducted by the timeROC package[34] to compare prognostic performance for predicting OS through calculating the area under the curve (AUC) value. Besides, we validated the risk score model in the external independent dataset and the combined dataset (comprised of TCGA-CHOL training cohort and the GSE107943 validation dataset) through the KM method and ROC analysis, respectively.

2.4. Weighted gene co-expression network analysis and functional enrichment analysis

To investigate the functional roles behind the integrated mRNA–lncRNA signature, we conducted Pearson correlation coefficient between the prognostic RNAs and all mRNAs in TCGA-CHOL dataset. *P*-value < .05 and correlation coefficient > 0.6 were chosen as the threshold to consider mRNAs co-expressed with the identified mRNAs and lncRNAs in the signature. Subsequently, WGCNA was performed on co-expression mRNAs and prognostic mRNAs to explore co-expression modules associated with the risk model using WGCNA package.[35] Soft power parameter with 12 was used to construct the topological overlap matrix, and a dynamic hybrid cut method with a minimum module size of 30 genes was implemented to detect co-expression clusters. The relationship between ME and risk scores was evaluated and further plotted as a heatmap. The genes from co-expression modules significantly related to the mRNA–lncRNA integrated signature were then subjected to GO and KEGG functional enrichment analysis through the clusterProfiler R package.[36] The cutoff of *q*-value < 0.05 was used to identify both significantly enriched GO terms and KEGG pathways.

3. Results

3.1. Clinical characteristics of CCA datasets

Descriptive statistics for clinical characteristics of all the CCA patients involved in the current study were summarized in Table 1. The training set from TCGA contained 36 CCA patients with a mean follow-up time of 806 days, ranging from 10 to 1976 days. The mean age of individuals from TCGA was 64. There were 18 (50%) patients alive at the time of the last follow-up. The 30 CCA patients from GSE107943 selected as validation set had a mean follow-up time of 334 days (ranging from 14 to 1140), the average age of 66 at initial pathologic diagnosis, and more than half of patients (17) dead during follow-up times.

3.2. Differentially expressed mRNAs and lncRNAs in CCA

Using the expression profiles from the TCGA-CHOL dataset, we compared mRNAs and lncRNAs expression level between 36 CCA tumor and 9 adjacent normal samples. A total of 4787 DEmRNAs (Fig. 1A) and 1950 DElncRNAs (Fig. 1B) were identified by DESeq2, whereas 4907 DEmRNAs (Fig. 1C) and 2216 DElncRNAs (Fig. 1D) were detected by edgeR. The 4628 DEmRNAs (Fig. 1E) and 1810 DElncRNAs (Fig. 1F) identified by both packages were utilized for further downstream analyses. The heatmap plots showed that CCA samples were clearly distinguished from normal tissues based on the top 200 DEmRNAs and DElncRNAs (Fig. 1G and H).

3.3. Development of integrated mRNA–lncRNA signature in the training cohort

To detect prognostic biomarkers in CCA patients, we carried out a univariate Cox proportional hazards regression analysis for each of the 4628 DEmRNAs and 1810 DElncRNAs in the discovery dataset. A total of 6 mRNAs (*ACRV1*, *TMEM121B*, *PIWIL4*, *GOLGA8M*, *CFHR3*, and *FUT4*) and 3 lncRNAs (*AC134682.1*, *AC007285.1*, and *AC138430.1*) with *P*-value < 0.01 were determined to be significantly associated with overall survival. We subsequentially performed a meta-analysis for the 6 mRNAs using random effect model due to the *P*-value of heterogeneity test <.05 as shown in Figure 2. The *TMEM121B* and *GOLGA8M* were eliminated due to the inconsistently differential expression level across these 5 CCA cohorts with pooled SMD of −0.15 (95% CI: −0.65 to 0.36) and −1.48 (95% CI: −3.84 to 0.87), respectively. The 4 mRNAs and 3 lncRNAs were further subjected to a stepwise multivariate Cox regression analysis. The optimally integrated mRNA–lncRNA signature was determined with the low value of Akaike information criterion and significance tests for each RNA,[37] which included 2 mRNAs (*CFHR3* and *PIWIL4*) and 2 lncRNAs (*AC007285.1* and *AC134682.1*). The chromosomal position, hazard ratio, *P*-value and coefficient of these 4 prognostic RNAs in CCA are provided in Table 2. Among these 4 RNAs, only *CFHR3* had a positive coefficient, which suggested that higher expression of this mRNA was related to shorter survival of the CCA patients, whereas the other 3 RNAs were potentially protective factors since their negative coefficients, implying that higher expression level of these genes was associated with greater survival time. The pooled SMD of *CFHR3* and *PIWIL4* were −2.80 (95% CI: −4.13 to −1.47) and 1.46 (95% CI: 0.51 to 2.41), respectively, which provided additional confidence of prognostic value since these 2 mRNAs were also differently expressed cross various independent CCA cohorts.

To build integrated mRNA–lncRNA signature for survival prediction in CCA patients, we calculated the risk scores for each individual using expression level of the 2 mRNAs and the 2 lncRNAs weighted by their regression coefficients from above multivariate

Table 1

The clinicopathological features of CCA patients in training and independent validation set.

Variables		Training set (n = 36)	Independent validation set (n = 30)	Combined set (n = 66)
Follow-up (days)	Mean (range)	806 (10–1976)	334 (14–1140)	625 (10–1976)
Tumor stage	I/II	28 (77.78%)	21 (70.00%)	49 (74.24%)
	III/IV	8 (22.22%)	9 (30.00%)	17 (25.76%)
Age, years	<60	11 (30.56%)	8 (26.67%)	19 (28.79)
	≥60	25 (69.44%)	22 (73.33%)	47 (71.21%)
Gender	Female	20 (55.56%)	6 (20.00%)	26 (39.39%)
	Male	16 (44.44%)	24 (80.00%)	40 (60.61%)
Histological type	ICCA	30 (83.33%)	30 (100.00%)	62 (93.94%)
	ECCA	6 (16.67%)	0 (0.00%)	4 (6.06%)
Residual tumor	R0	28 (77.78%)	/	/
	R1	5 (13.89%)	/	/
	RX	3 (8.33%)	/	/
Histologic grade	G1/G2	16 (44.44%)	/	/
	G3/G4	20 (55.56%)	/	/
Survival status	Alive	18 (50.00%)	13 (43.33%)	31 (46.97%)
	Dead	18 (50.00%)	17 (56.67%)	35 (53.03%)

CCA = cholangiocarcinoma.

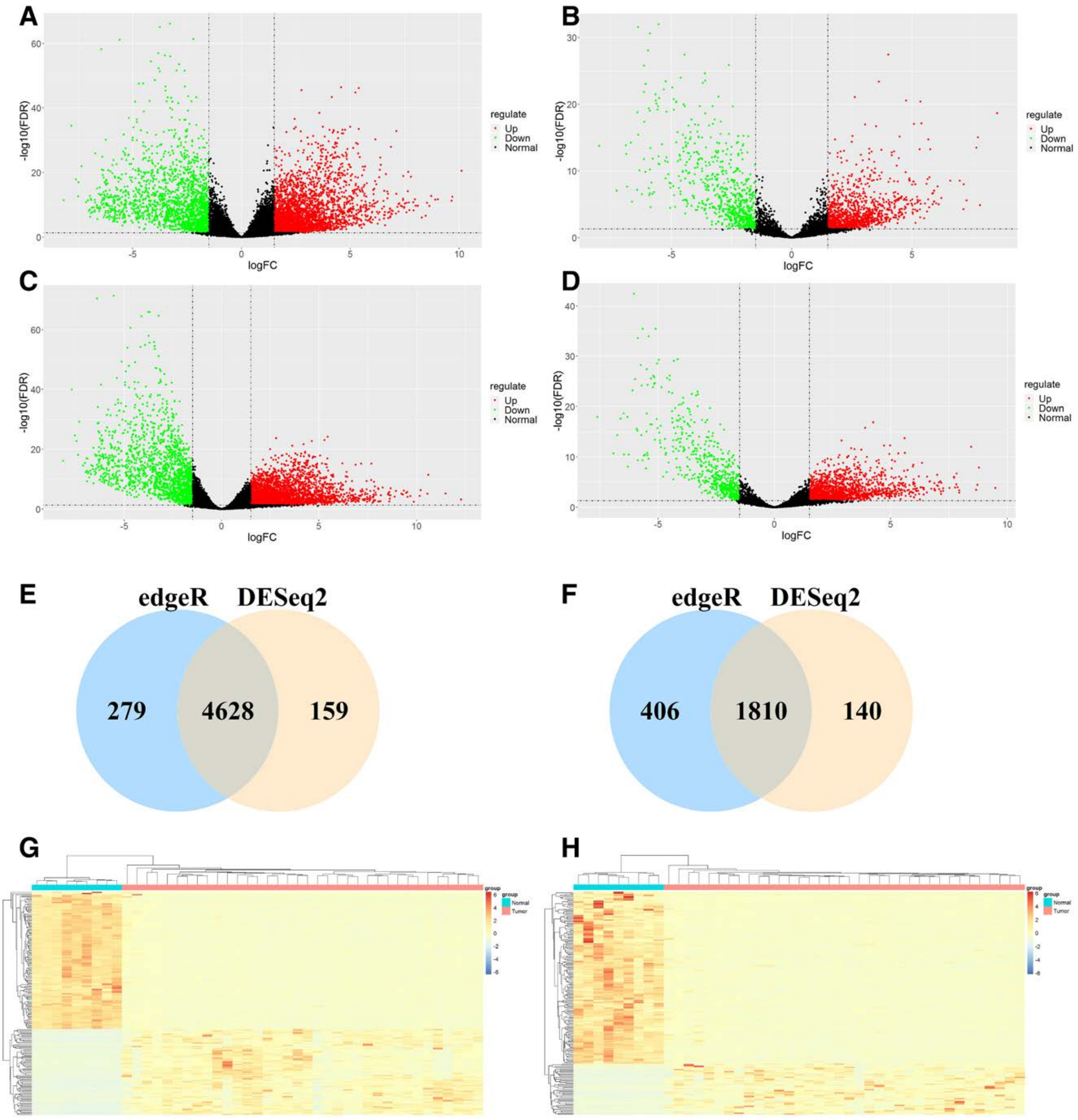

Figure 1. Identification of differentially expressed mRNAs (DEmRNAs) and lncRNAs (DElncRNAs). DEmRNAs (A) and DElncRNAs (B) were identified using the DESeq2 package; DEmRNAs (C) and DElncRNAs (D) were identified using the edgeR package; Venn diagram comparing DEmRNAs (E) and DElncRNAs (F) between edgeR and DESeq2 package; The unsupervised hierarchical clustering heatmap with top 200 differentially expressed mRNAs (G) and top 200 differentially expressed lncRNAs (H) through DESeq2.

Cox analysis as follows: risk score = (3.18 × expression value of *CFHR3*) + (−1.62 × expression value of *PIWIL4*) + (−2.97 × expression value of *AC007285.1*) + (−1.95 × expression value of *AC134682.1*). The patients were then categorized into high-risk group (23 patients) and low-risk group (13 patients) based on the optimal cutoff point (−0.14) determined by "cutp" function from survMisc package (Fig. 3A). The survival status and the expression pattern of the 4 prognostic RNAs for each CCA patient in the discovery cohort are presented in Figure 3A as well. The KM curve with a log-rank test suggested that patients in the low-risk group have significantly longer survival time compared to the patients in a high-risk group (Fig. 3B). Additionally, the univariate Cox regression model (Table 3) showed a 6.46-fold increase (*P*-value = 1.35E−02) of hazard ratio in the high-risk group compared to the low-risk group for OS. Time-dependent ROC curve for the risk score model in the training cohort (shown in Fig. 3C) revealed an AUC of 0.872 and 0.790 for 1-year and 3-year OS prediction, respectively, which implied that the integrated mRNA–lncRNA signature possessed a high specificity and sensitivity.

3.4. Validation for the prognostic prediction value of integrated mRNA–lncRNA signature in the independent validation cohort

To evaluate robustness of the integrated mRNA–lncRNA signature for prognosis in CCA patients, we validated its prognostic ability in an independent cohort (GSE107943)

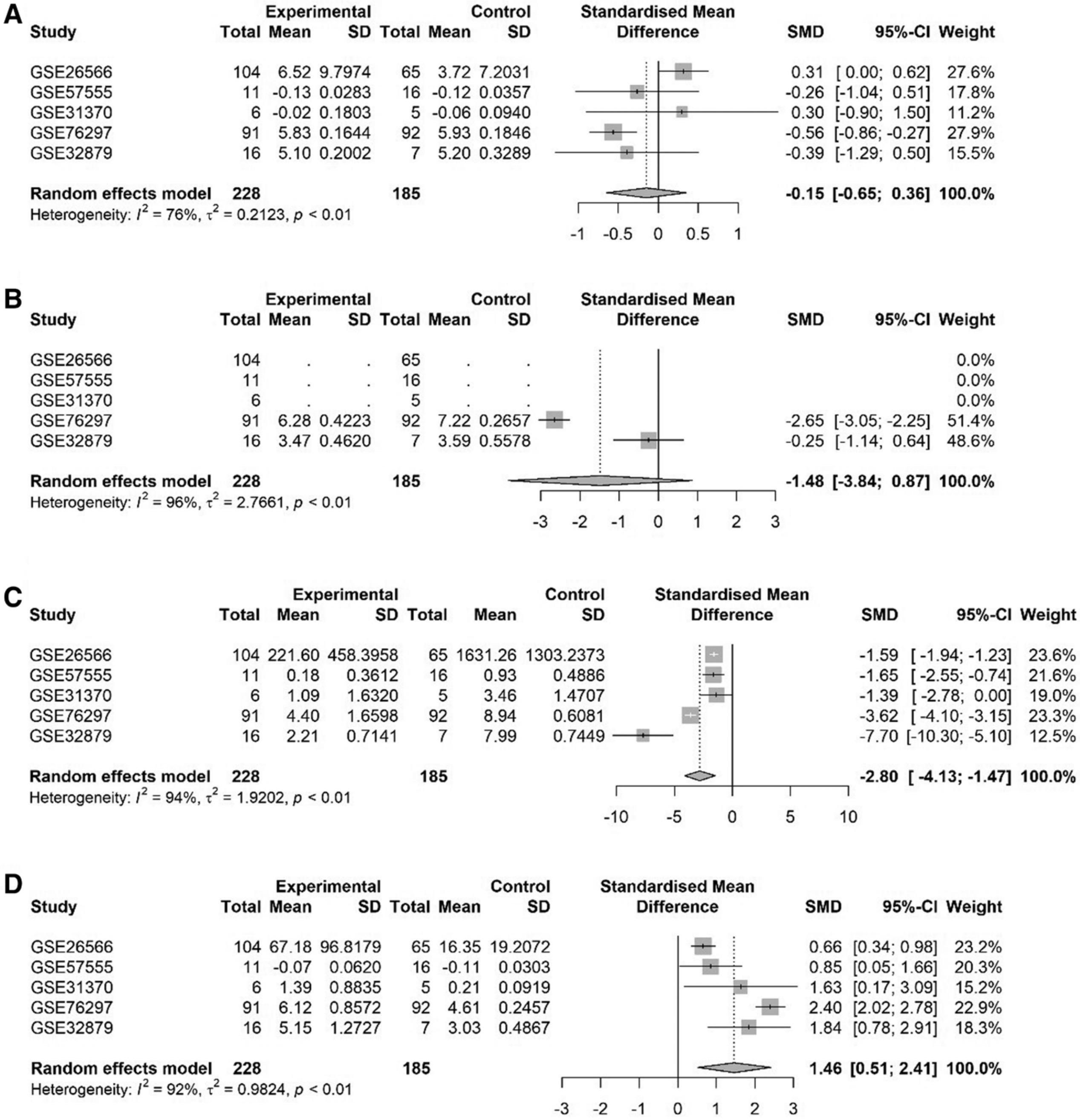

Figure 2. Mata-analysis to evaluate the expression of *TMEM121B* (A), *GOLGA8M* (B), *CFHR3* (C), and *PIWIL4* (D) among GSE26566, GSE57555, GSE31370, GSE76297, and GSE32879.

Table 2

The 4 prognostic RNAs significantly associated with the overall survival in CCA patients.

Ensemble ID	Gene name	Chromosomal position	Gene type	HR	*P*-value	Coefficient
ENSG00000116785	CFHR3	chr1: 196774795–196795406 (+)	Protein_coding	24.13	5.24E–04	3.18
ENSG00000134627	PIWIL4	chr11: 94543840–94621421 (+)	Protein_coding	0.20	3.83E–02	–1.62
ENSG00000227014	AC007285.1	chr7: 29988600–30027543 (+)	Antisense	0.05	6.44E–04	–2.97
ENSG00000261693	AC134682.1	chr8: 142403652–142407028 (+)	Antisense	0.14	1.34E–02	–1.95

CCA = cholangiocarcinoma.

obtained from GEO database and yielded similar results as we obtained from the training dataset. Individuals in the validation dataset were divided into high-risk group (16 patients) and low-risk group (14 patients) according to the threshold determined by the same method as for the training dataset. The survival outcome of patients in high-risk

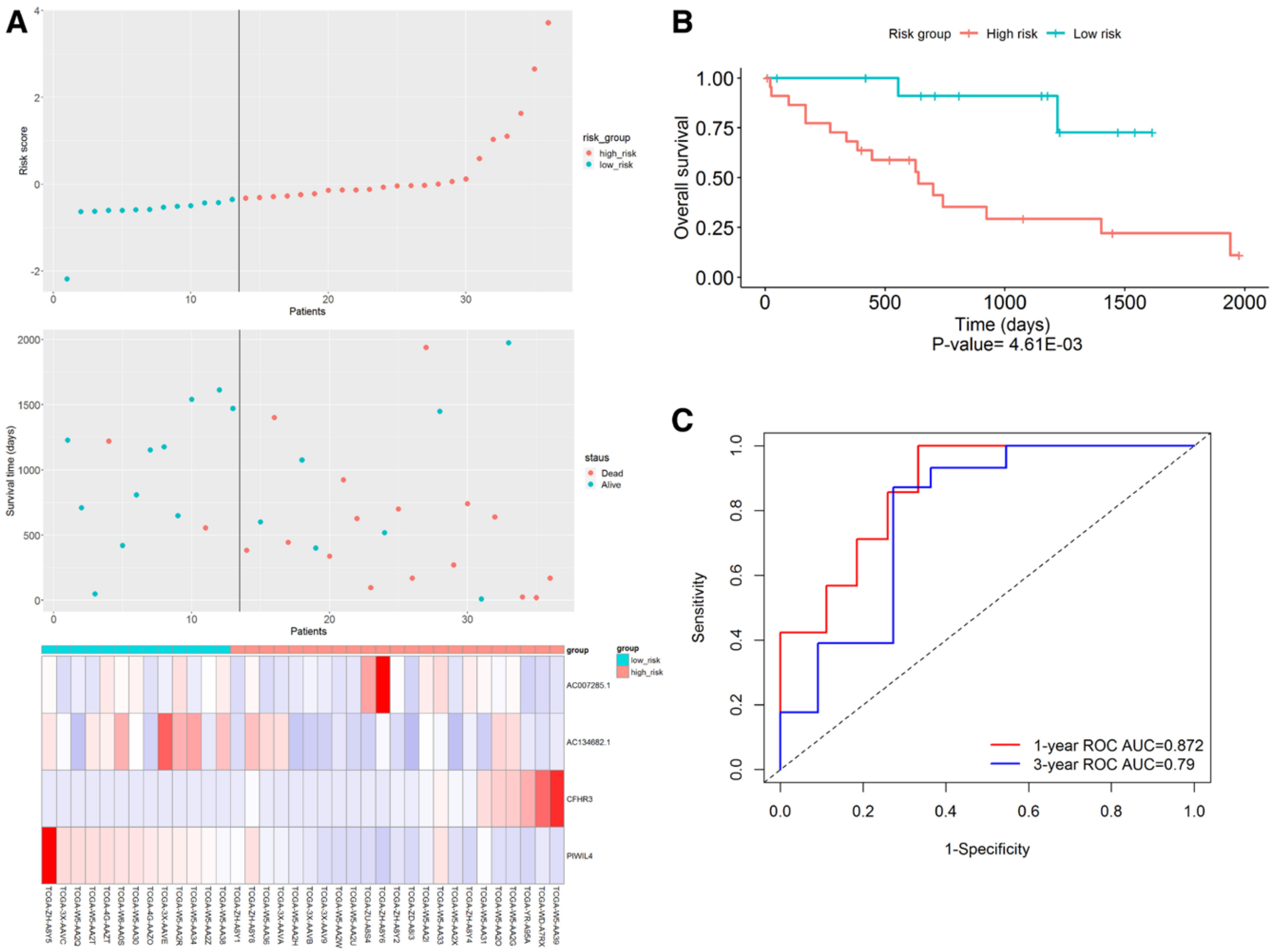

Figure 3. Prognosis assessment of the integrated mRNA–lncRNA signature in the training cohort. (A) The risk distribution, the survival time of patients, expression heatmap of integrated mRNA–lncRNA signature. (B) Kaplan–Meier analysis for overall survival of cholangiocarcinoma patients between low-risk and high-risk groups. (C) Time-dependent receiver operating characteristic (ROC) analysis for overall survival prediction based on the risk scores with 1 and 3 years as the time point.

Table 3

Univariate and multivariate Cox regression analysis of integrated mRNA–lncRNA signature in different dataset.

	Univariate analysis			Multivariate analysis		
Variables	**HR**	**95% CI**	***P*-value**	**HR**	**95% CI**	***P*-value**
Training set (n = 36)						
Risk group (high vs low)	6.46	1.47–28.39	1.35E−02	1.19	1.19–35.08	3.07E−02
Age (≥60 vs <60)	0.73	0.28–1.92	5.19E−01	0.15	0.15–1.58	2.30E−01
Gender (male vs female)	1.39	0.54–3.53	4.94E−01	0.25	0.25–3.12	8.53E−01
Tumor stage (III + IV vs I + II)	1.48	0.52–4.21	4.67E−01	0.24	0.24–7.03	7.69E−01
Residual tumor (R1 vs R0)	1.57	0.44–5.65	4.88E−01	0.48	0.48–7.93	3.52E−01
Histologic grade (G1 + G2 vs G3 + G4)	1.64	0.62–4.32	3.21E−01	0.74	0.74–8.39	1.39E−01
Histologic type (ICCA vs ECCA)	0.84	0.24–2.91	7.78E−01	0.22	0.22–11.91	6.39E−01
Independent validation set (n = 30)						
Risk group (high vs low)	8.04	2.26–28.62	1.29E−03	7.70	1.99–29.77	3.10E−03
Age (>=60 vs <60)	0.93	0.30–2.91	8.96E−01	0.58	0.16–2.10	4.07E−01
Gender (male vs female)	1.37	0.31–6.16	6.80E−01	1.05	0.22–5.13	9.49E−01
Tumor stage (III + IV vs I + II)	5.15	1.60–16.57	5.93E−03	3.22	0.97–10.69	5.65E−02
Combined set (n = 66)						
Risk group (high vs low)	5.27	2.38–11.67	4.23E−05	5.27	2.34–11.86	6.09E−05
Age (≥60 vs <60)	0.86	0.42–1.76	6.76E−01	0.73	0.35–1.55	4.13E−01
Gender (male vs female)	1.37	0.68–2.77	3.77E−01	1.23	0.60–2.51	5.74E−01
Tumor stage (III + IV vs I + II)	2.20	1.08–4.51	3.07E−02	2.04	0.97–4.28	5.88E−02

P-value less than .05 was marked in red.

group was significantly worse than that in low-risk group (*P*-value = 1.71E−04) as shown in Figure 4B. Notably, there were 14 deaths among the patients with high-risk scores, whereas there were only 3 death events in low-risk group (Fig. 4A). The hazard ratio of high-risk group was 8.04 folds compared to that of low-risk group (95% CI = 2.26–28.62, *P*-value = 1.29E–03) in the univariable analysis (Table 3). The AUC of time-dependent ROC curve was 0.750 and 0.729 for 1-year and 3-year overall survival prediction (Fig. 4C), representing that the risk score model has a good performance in CCA patients' OS prediction.

Furthermore, we assessed prognostic performance of the integrated mRNA–lncRNA signature in the combined dataset. Indeed, the findings were generally consistent with those from the discovery or validation cohorts. KM survival curves between 2 risk groups were significantly different in the combined dataset with a *P*-value of 5.51E–06. The survival rates at 3-year and 5-year were 26.47% and 23.53% for patients in the high-risk group, as compared to 87.50% and 75.00% survival rate for patients in the low-risk at 3-year and 5-year respectively. Patients with high-risk scores exhibited a 5.27-fold increased risk than patients in low-risk group (Table 3).

3.5. Correlation between the integrated mRNA–lncRNA signature and other clinicopathologic characteristics

To investigate independence of the integrated mRNA–lncRNA signature in survival prediction, a multivariate Cox regression analysis was performed including risk scores, age, gender, tumor stage, residual tumor and histologic grade. In the training cohort, the integrated mRNA–lncRNA signature was the most significant (*P*-value = 3.07E–02) compared with the other clinical characteristics. Furthermore, after adjusting for the age, gender, and tumor stage, the hazard ratios of overall survival in high-risk versus low-risk group were 7.70 and 5.27 in the validation and the combined datasets, respectively (Table 3).

Besides, univariate Cox analysis revealed that the tumor stage was associated with OS in both the validation (*P*-value = 5.93E–03) and the combined (*P*-value = 3.07E–02) datasets (Table 3). The stratification analysis was carried out to estimate the relationship between the mRNA–lncRNA signature and tumor stage. All patients were classified into 2 subgroups: I/II stage with 49 samples and III/IV stage with 17 individuals. As shown in KM curve (Fig. 5A and B), patients with high-risk scores had significantly shorter survival time than those with low-risk scores both in stage I/II (*P*-value = 4.71E–04) and stage III/IV (*P*-value = 4.97E–03) subgroups. Accordingly, the multivariate Cox analysis and stratification analysis demonstrated that prognostic capability of the integrated mRNA–lncRNA signature was independent from other clinical features.

We also compared prognostic performance of the mRNA–lncRNA signature with other clinical features by calculating the AUC of time-dependent ROC. In the combined dataset, the AUC of mRNA–lncRNA risk scores at 3 years was 0.781, which was higher than that of tumor stage (AUC = 0.673), gender (AUC = 0.541), and age (AUC = 0.505) as shown in Figure 5C. These results demonstrated superior prognostic performance of the identified mRNA–lncRNA signature as compared to any of

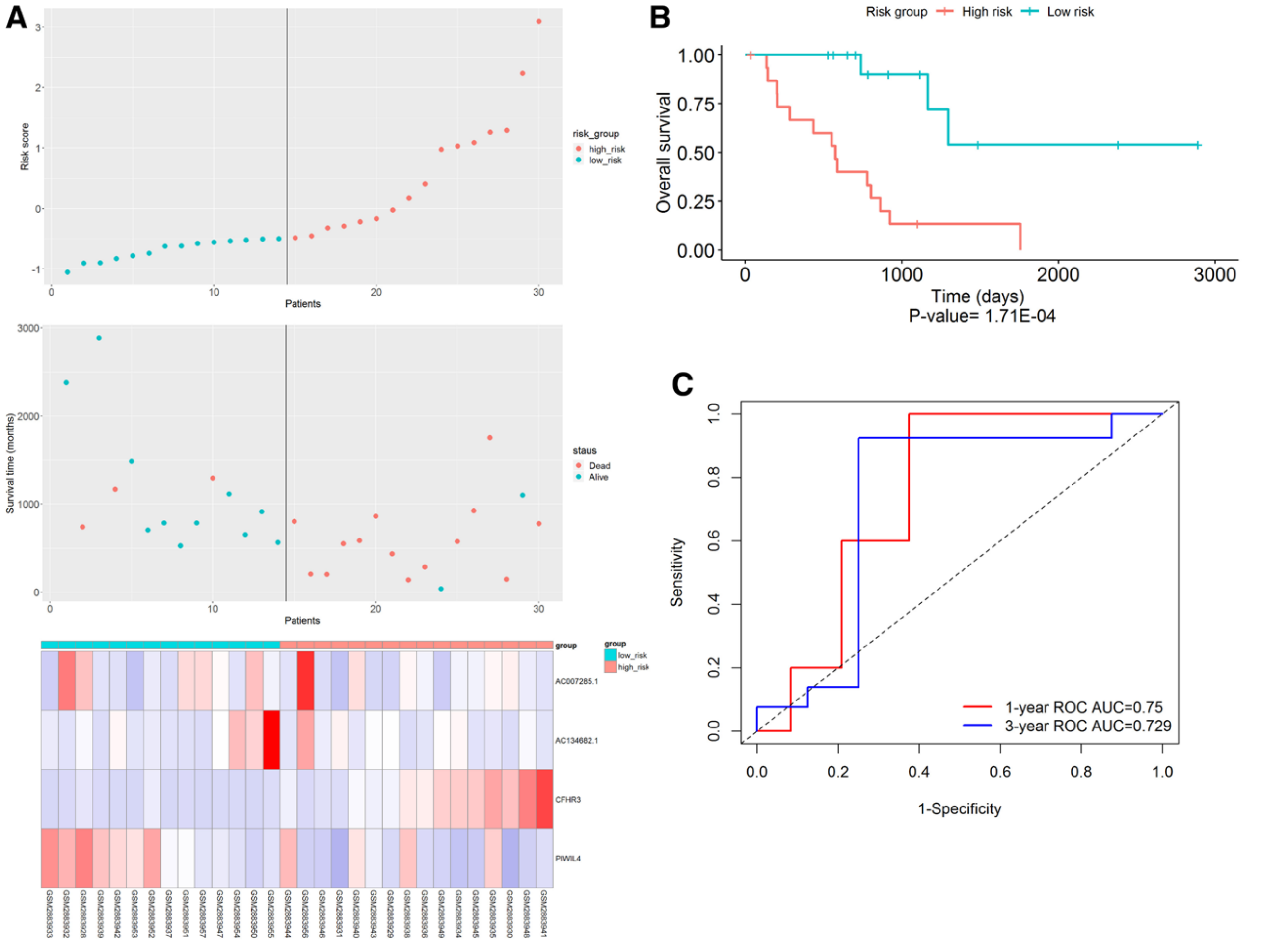

Figure 4. Prognosis validation of the integrated mRNA–lncRNA signature in the independent cohort. (A) The risk distribution, the survival time of patients, expression heatmap of integrated mRNA–lncRNA signature. (B) Kaplan–Meier analysis of overall survival of between low- and high-risk cholangiocarcinoma patients. (C) Time-dependent receiver operating characteristic (ROC) analysis for overall survival prediction based on the risk score with 1 and 3 years as the time point.

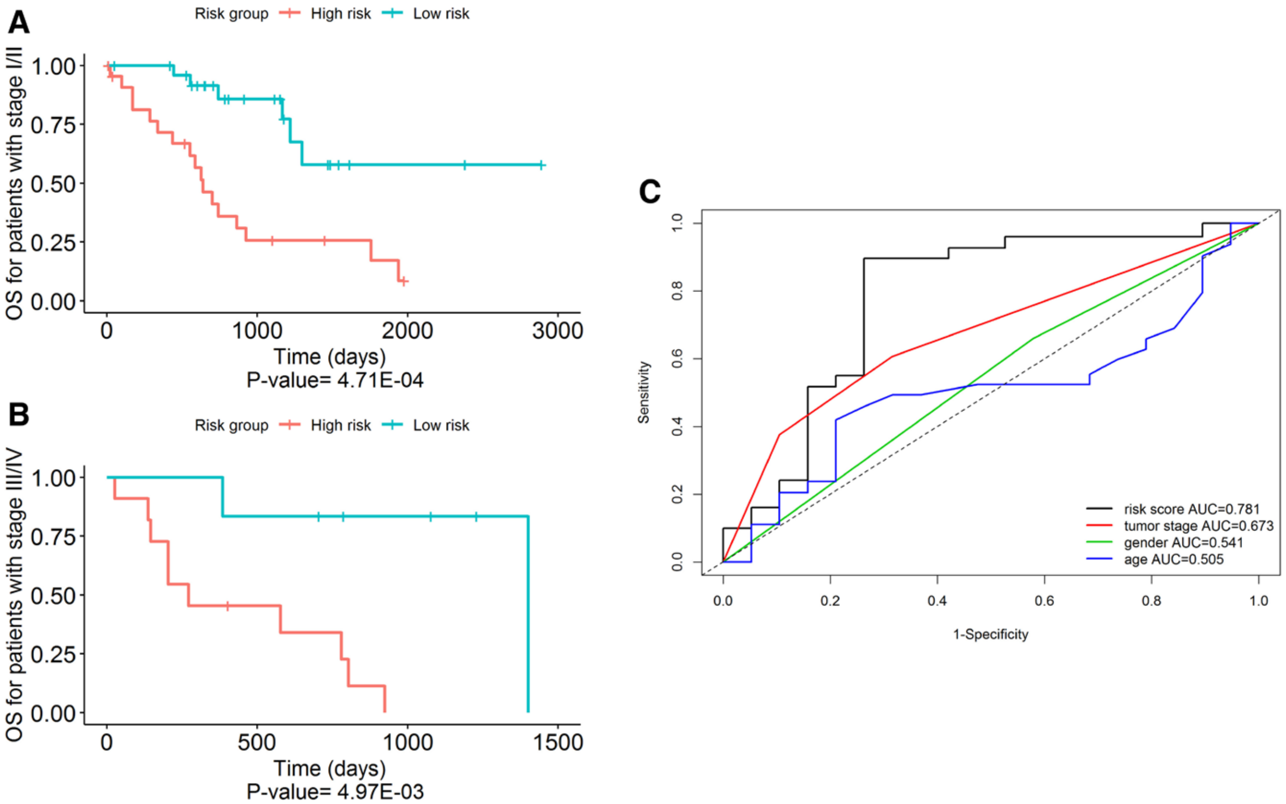

Figure 5. Correlation between the integrated mRNA–lncRNA signature and other clinicopathologic characteristics. Kaplan–Meier curve for patients with stage I/II (A) and stage III/IV (B); (C) comparison of sensitivity and specificity for overall survival prediction between the mRNA–lncRNA signature and other clinical factors in the combined dataset.

the other factors investigated in the present study (i.e., tumor stage, age, and gender).

3.6. Functional roles of the integrated mRNA–lncRNA signature in the CCA biology

We performed WGCNA for 1067 co-expressed mRNAs to cluster genes that highly correlated with the risk scores. A total of 5 modules were identified including a turquoise module with 522 mRNAs, a blue module with 272 mRNAs, a brown module with 139 mRNAs, a yellow module with 131 mRNAs and a gray module with 3 mRNAs. The turquoise module showed a higher correlation ($R = 0.87$, P-value = 8.00E–12) with risk score model than the other modules (Fig. 6A and B). We then carried out GO term and KEGG pathway enrichment analyses based on the 522 genes from the turquoise module. A total of 567 GO terms and 48 KEGG pathways were identified as significantly enriched by these genes. The top 10 GO biological processes and KEGG pathways are shown in Figure 6C and D. Some of the strongly enriched metabolic biological processes included small molecule catabolic process, organic acid catabolic process, carboxylic acid catabolic process and lipid related metabolic processes. The enriched KEGG pathways included metabolic-related pathways involved in cholesterol metabolism, drug metabolism—cytochrome P450, amino acid (glycine, serine, and threonine) metabolism and primary bile acid biosynthesis. Additionally, genes in the turquoise module also enriched in complement and coagulation cascades and the PPAR signaling pathways.

3.7. Comparison of the integrated mRNA–lncRNA signature with other CCA prognostic signatures

To further evaluate the prognostic value of the novel signature, we compared it with previous molecular signature identified from transcriptome profiles, including 3-miRNA signature (*miR-10b*, *miR-22*, and *miR-551b*)[17] and 7-mRNA signature (*CD36*, *GGCX*, *UBASH3B*, *DBN1*, *PTTG1*, *CCNA2*, and *SPATS2*).[18] We performed multivariate Cox regression and the ROC analyses for 7-mRNA signatures in the TCGA-CHOL and GSE107943 dataset, and the 3-miRNA signature in the TCGA-CHOL since *miR-22* is not available in the validation dataset. The 3 miRNAs showed a relatively precise prediction with an AUC of 0.715 at 1-year and 0.723 at 3-year (Fig. 7). Although Guo et al reported that the 7-mRNA was a valuable signature for relapse of cholangiocarcinoma,[18] it presented a relatively worse accuracy for predicting overall survival time with an AUC < 0.70 at 1-year and 3-year in TCGA-CHOL and GSE107943 cohorts.

4. Discussion

The tumor-node-metastasis staging system is the most common indicator to predict survival time of patients with malignancy worldwide. Unfortunately, due to high molecular heterogeneity in CCA patients, it is difficult to predict OS by clinical features.[38] Up to the time of conducting this current study, only a few studies have performed using high-throughput sequencing data to identify powerful molecular biomarkers for CCA prognosis. For example, Cao et al[17] discovered 3 miRNAs (*miR-10b*, *miR-22*, and *miR-551b*) prognostic signature that showed a relatively precise prediction with an AUC of 0.715 for 1-year and 0.723 for 3-year. However, no study has endeavored to investigate candidate mRNAs and lncRNAs as an integrated prognostic signature for CCA. In our study, we identified an integrated prognostic signature consisting of 2 mRNAs (*CFHR3* and *PIWIL4*) and 2 lncRNAs (*AC007285.1* and *AC134682.1*) that could be used for CCA patients' prognostic prediction. The signature was further confirmed in the independent validation

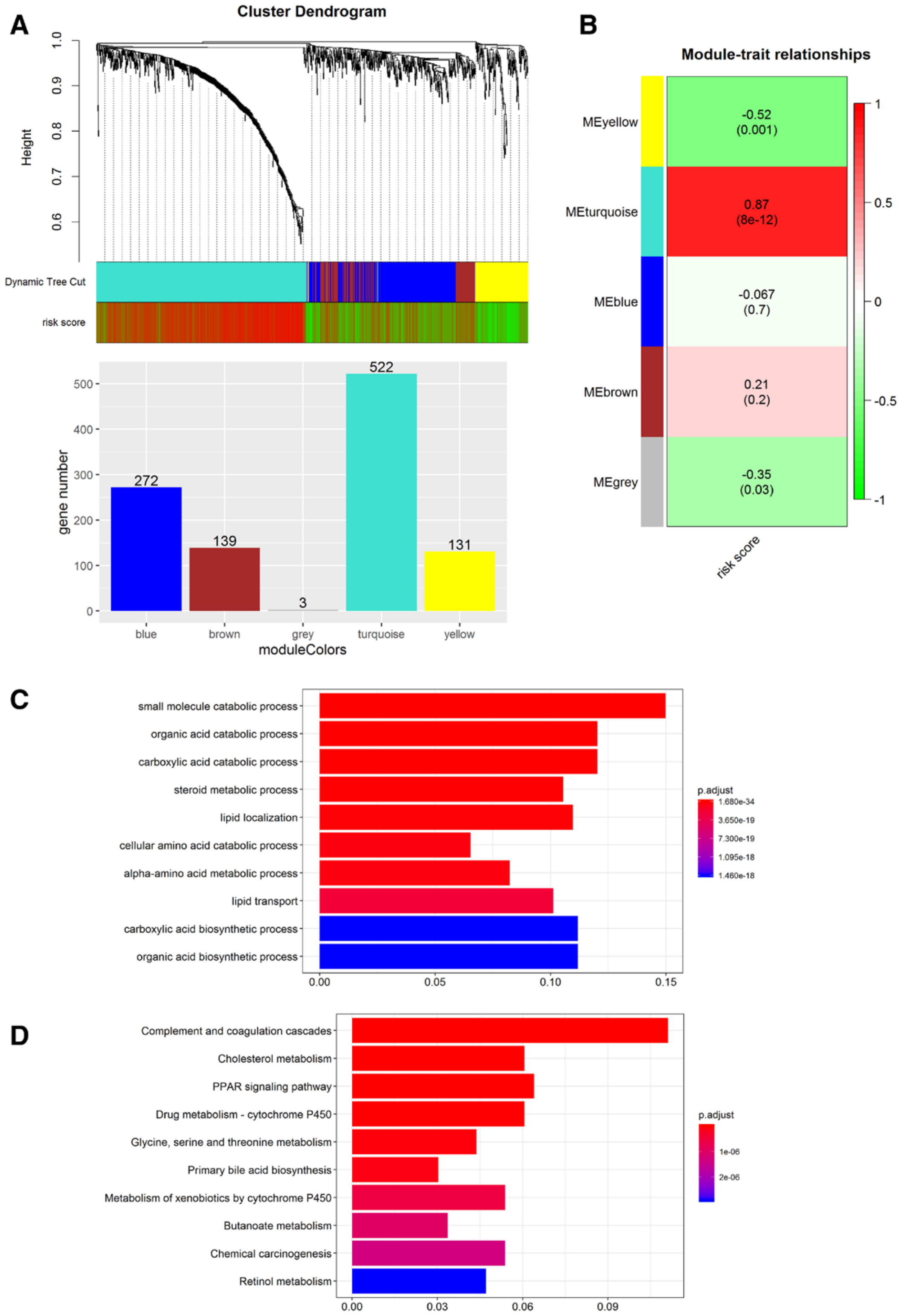

Figure 6. Functional enrichment analysis of the integrated mRNA–lncRNA signature related functional genes. (A) Clustering dendrogram and bar chart of gene number in the 5 modules that were generated. The color bar labeled as "Dynamic Tree Cut" beneath the dendrogram represents the module assignment of each gene. The other color bar labeled as "risk score" represents the correlation of genes with risk score. Red means a gene is positively correlated with risk score and green means a negative correlation. (B) Heatmap of the correlation between module eigengenes (ME) and risk score. Red indicates positive correlation and green indicates negative. The numbers in the brackets are *P*-value of the correlation. (C) GO term enrichment results of the turquoise module (522 genes). (D) KEGG enrichment results of the turquoise module (522 genes).GO = gene ontology; KEGG = Kyoto Encyclopedia of Genes and Genomes; WGCNA = weighted gene co-expression network analysis.

as well as in the complete dataset. Indeed, the identified signature performed well in 1-year and 3-year survival prediction according to time-dependent ROC curves (Figures 2C, 3C, and 4C). The multivariable Cox regression and the stratified analysis revealed that our integrated mRNA–lncRNA signature had independent prognostic ability from other clinical features.

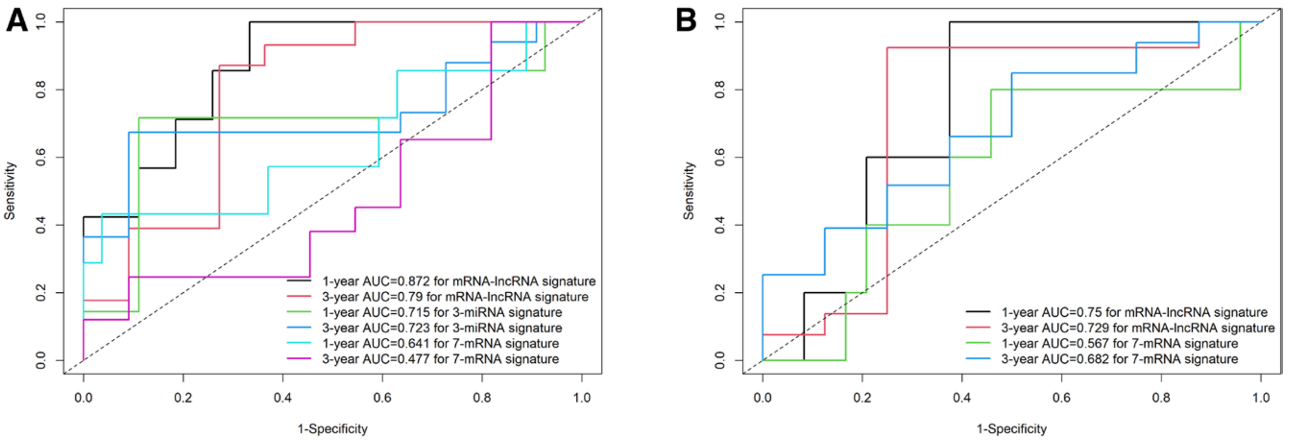

Figure 7. The time-dependent receiver operating characteristic analysis for the integrated mRNA–lncRNA, 3-miRNA, and 7-mRNA prognostic signatures. The time-dependent receiver operating characteristic (ROC) curves for overall survival prediction between the integrated mRNA–lncRNA, 3-miRNA and 7-mRNA signatures in the training dataset (A) and the validation dataset (B).

The relationship between these prognostic biomarkers and OS of CCA patients implied the signature's potentially vital roles in underlying mechanism of carcinogenesis and progression of CCA. Published records of the mRNAs identified in our signature indicate that overexpression of *CFHR3* (Complement Factor H-Related Protein 3) would suppress proliferation and promote apoptosis of hepatocellular carcinoma (HCC) cells.[39] It has also been reported as a potential prognostic biomarker for HCC.[40] *PIWIL4* (piwi like RNA-mediated gene silencing 4) belongs to Piwi-like (Piwil) proteins and is aberrantly expressed in various human cancers, including breast cancer,[41] retinoblastoma,[42] and hepatocellular carcinoma.[43] As for the 2 antisense lncRNAs identified in this study, their functional roles in any cancer have not been reported. Therefore, to infer potential biological roles of the integrated mRNA–lncRNA signature, we performed WGCNA analysis for mRNAs strongly co-expressed with the discovered prognostic mRNAs and lncRNAs. The 522 genes from turquoise module, which was significantly correlated with risk score model, were mainly enriched in metabolic-related biological pathways and PPAR signaling pathway. These pathways are well documented as participating in the carcinogenesis and progression of CCA.[44] Various studies based on multiple CCA independent cohorts[45–47] also detected that metabolic-related biological processes including small molecule and lipid metabolic processes related to energy metabolism were pivotal for CCA development. Increasing evidence has demonstrated that fatty acid synthesis related genes (*FASN* and *SLC27A1*),[48,49] fatty acid transport proteins (*FATP2*, *FATP1*, *FATP5*, and *CD36*), and fatty acid binding proteins (*FABP1*, *FABP4*, and *FABP5*)[50] contribute to CCA carcinogenesis. Additionally, PPARγ ligands suppressed cholangiocarcinoma cell growth[51,52] and induced the cholangiocarcinoma cell apoptosis,[53] which suggested potentially vital roles of the PPAR signaling pathway in CCA pathogenesis.

To our knowledge, this study is the first attempt to develop a prognostic signature for CCA patients through combining the expression profiles of both mRNA and lncRNA at whole gene expression level. However, there are a few limitations to the current study. Firstly, the small size sample and the mismatched number of individuals between tumor and normal group may cause the false positive rate of DERNAs. Secondly, since the RNA expression in the current study was quantified using CCA tissues, the prognostic capability might not be reproduced when body fluids (such as saliva, serum, urine, and stool) commonly used in clinical application are utilized. Hence, collecting more CCA samples and verifying prognostic value of the risk score model using samples from body fluids are necessary for further research endeavors.

In conclusion, we performed a comprehensive analysis to develop an integrated mRNA–lncRNA signature for CCA patients' prognosis. The identified signature consisting of 2 mRNAs (*CFHR3* and *PIWIL4*) and 2 lncRNAs (*AC007285.1* and *AC134682.1*) was independent of other clinical characteristics including age, gender, tumor stage, residual tumor, and histologic grade. WGCNA indicated that the mRNAs strongly co-expressed with our signature were enriched in numerous metabolic processes and pathways, some of which have been reported to be involved in different cancers. Our findings revealed that the integrated mRNA–lncRNA signature might serve as a valuable and alternative prognostic biomarker for CCA patients.

Acknowledgments

The authors would like to thank all the subjects for their participation in the study. We gratefully acknowledge the support funding received from the National Natural Science Foundation of China.

Author contributions

Conceptualization: Hongbo Xin, Feng Zhang.
Data curation: Derong Xu, Liping Zeng.
Formal analysis: Liping Zeng.
Funding acquisition: Derong Xu.
Investigation: Derong Xu, Lili Wei.
Methodology: Liping Zeng, Robert Mukiibi.
Writing – original draft: Robert Mukiibi, Hongbo Xin, Feng Zhang.
Writing – review & editing: Lili Wei, Feng Zhang.

References

[1] Khan SA, Thomas HC, Davidson BR, et al. Cholangiocarcinoma. Lancet. 2005;366:1303–14.

[2] Kirstein MM, Vogel A. Epidemiology and risk factors of cholangiocarcinoma. Visc Med. 2016;32:395–400.

[3] Blechacz B. Cholangiocarcinoma: current knowledge and new developments. Gut Liver. 2017;11:13–26.

[4] Saha SK, Zhu AX, Fuchs CS, et al. Forty-year trends in cholangiocarcinoma incidence in the U.S.: intrahepatic disease on the rise. Oncologist. 2016;21:594–9.

[5] Banales JM, Marin JJG, Lamarca A, et al. Cholangiocarcinoma 2020: the next horizon in mechanisms and management. Nat Rev Gastroenterol Hepatol. 2020;17:557–88.

[6] Valle J, Wasan H, Palmer DH, et al. Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. N Engl J Med. 2010;362:1273–81.

[7] Okusaka T, Nakachi K, Fukutomi A, et al. Gemcitabine alone or in combination with cisplatin in patients with biliary tract cancer: a comparative multicentre study in Japan. Br J Cancer. 2010;103:469–74.

[8] Blechacz B, Komuta M, Roskams T, et al. Clinical diagnosis and staging of cholangiocarcinoma. Nat Rev Gastroenterol Hepatol. 2011;8:512–22.

[9] Lowery MA, Ptashkin R, Jordan E, et al. Comprehensive molecular profiling of intrahepatic and extrahepatic cholangiocarcinomas: potential targets for intervention. Clin Cancer Res. 2018;24:4154–61.

[10] Fiste O, Ntanasis-Stathopoulos I, Gavriatopoulou M, et al. The emerging role of immunotherapy in intrahepatic cholangiocarcinoma. Vaccines (Basel). 2021;9:422.

[11] Merdrignac A, Papoutsoglou P, Coulouarn C. Long noncoding RNAs in cholangiocarcinoma. Hepatology. 2021;73:1213–26.

[12] Huang CS, Zhu YQ, Xu QC, et al. YTHDF2 promotes intrahepatic cholangiocarcinoma progression and desensitises cisplatin treatment by increasing CDKN1B mRNA degradation. Clin Transl Med. 2022;12:e848.

[13] Tiemin P, Fanzheng M, Peng X, et al. MUC13 promotes intrahepatic cholangiocarcinoma progression via EGFR/PI3K/AKT pathways. J Hepatol. 2020;72:761–73.

[14] Lei S, Cao W, Zeng Z, et al. JUND/linc00976 promotes cholangiocarcinoma progression and metastasis, inhibits ferroptosis by regulating the miR-3202/GPX4 axis. Cell Death Dis. 2022;13:967.

[15] Ni Q, Zhang H, Shi X, et al. Exosomal lncRNA HCG18 contributes to cholangiocarcinoma growth and metastasis through mediating miR-424-5p/SOX9 axis through PI3K/AKT pathway. Cancer Gene Ther. 2023;30:582–95.

[16] Ruys AT, Groot Koerkamp B, Wiggers JK, et al. Prognostic biomarkers in patients with resected cholangiocarcinoma: a systematic review and meta-analysis. Ann Surg Oncol. 2014;21:487–500.

[17] Cao J, Sun L, Li J, et al. A novel three-miRNA signature predicts survival in cholangiocarcinoma based on RNASeq data. Oncol Rep. 2018;40:1422–34.

[18] Guo H, Cai J, Wang X, et al. Prognostic values of a novel multi-mRNA signature for predicting relapse of cholangiocarcinoma. Int J Biol Sci. 2020;16:869–81.

[19] Li F, Ma J, Yan C, et al. ER stress-related mRNA–lncRNA co-expression gene signature predicts the prognosis and immune implications of esophageal cancer. Am J Transl Res. 2022;14:8064–84.

[20] Chen ZA, Tian H, Yao DM, et al. Identification of a ferroptosis-related signature model including mRNAs and LncRNAs for predicting prognosis and immune activity in hepatocellular carcinoma. Front Oncol. 2021;11:738477.

[21] Balbin OA, Malik R, Dhanasekaran SM, et al. The landscape of antisense gene expression in human cancers. Genome Res. 2015;25:1068–79.

[22] Ahn KS, Kang KJ, Kim YH, et al. Genetic features associated with (18) F-FDG uptake in intrahepatic cholangiocarcinoma. Ann Surg Treat Res. 2019;96:153–61.

[23] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41:D991–5.

[24] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

[25] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

[26] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodological). 1995;57:289–300.

[27] Andersen JB, Spee B, Blechacz BR, et al. Genomic and genetic characterization of cholangiocarcinoma identifies therapeutic targets for tyrosine kinase inhibitors. Gastroenterology. 2012;142:1021–1031.e15.

[28] Murakami Y, Kubo S, Tamori A, et al. Comprehensive analysis of transcriptome and metabolome analysis in intrahepatic cholangiocarcinoma and hepatocellular carcinoma. Sci Rep. 2015;5:16294.

[29] Seok JY, Na DC, Woo HG, et al. A fibrous stromal component in hepatocellular carcinoma reveals a cholangiocarcinoma-like gene expression trait and epithelial-mesenchymal transition. Hepatology. 2012;55:1776–86.

[30] Chaisaingmongkol J, Budhu A, Dang H, et al. Common molecular subtypes among Asian hepatocellular carcinoma and cholangiocarcinoma. Cancer Cell. 2017;32:57–70.e3.

[31] Oishi N, Kumar MR, Roessler S, et al. Transcriptomic profiling reveals hepatic stem-like gene signatures and interplay of miR-200c and epithelial-mesenchymal transition in intrahepatic cholangiocarcinoma. Hepatology. 2012;56:1792–803.

[32] Schwarzer G, Carpenter J, Rücker G. Meta-Analysis with R. Switzerland: Springer International Publishing; 2015.

[33] Piti CHU, Cedex P, Diego S. An application of change point methods in studying the effect of age on survival in breast cancer. Comput Stati Data Analy. 1999;30:253–70.

[34] Blanche P, Dartigues JF, Jacqmin Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013;32:5381–97.

[35] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. 2008;9:559.

[36] Yu GC, Wang LG, Han YY, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS J Integr Biol. 2012;16:284–7.

[37] Venables WN, Ripley BD. Modern Applied Statistics with S. New York: Springer Publishing Company, Incorporated; 2010.

[38] Zheng B, Jeong S, Zhu Y, et al. miRNA and lncRNA as biomarkers in cholangiocarcinoma(CCA). Oncotarget. 2017;8:100819–30.

[39] Liu H, Zhang L, Wang P. Complement factor Hrelated 3 overexpression affects hepatocellular carcinoma proliferation and apoptosis. Mol Med Rep. 2019;20:2694–702.

[40] Liu J, Li WL, Zhao HT. CFHR3 is a potential novel biomarker for hepatocellular carcinoma. J Cell Biochem. 2020;121:2970–80.

[41] Heng ZSL, Lee JY, Subhramanyam CS, et al. The role of 17betaestradiolinduced upregulation of Piwilike 4 in modulating gene expression and motility in breast cancer cells. Oncol Rep. 2018;40:2525–35.

[42] Sivagurunathan S, Arunachalam JP, Chidambaram S. PIWI-like protein, HIWI2 is aberrantly expressed in retinoblastoma cells and affects cell-cycle potentially through OTX2. Cell Mol Biol Lett. 2017;22:17.

[43] Zeng G, Zhang D, Liu X, et al. Co-expression of Piwil2/Piwil4 in nucleus indicates poor prognosis of hepatocellular carcinoma. Oncotarget. 2017;8:4607–17.

[44] Pastore M, Lori G, Gentilini A, et al. Multifaceted aspects of metabolic plasticity in human cholangiocarcinoma: an overview of current perspectives. Cells. 2020;9:596.

[45] Tian A, Pu K, Li B, et al. Weighted gene coexpression network analysis reveals hub genes involved in cholangiocarcinoma progression and prognosis. Hepatol Res. 2019;49:1195–206.

[46] Likhitrattanapisal S, Tipanee J, Janvilisri T. Meta-analysis of gene expression profiles identifies differential biomarkers for hepatocellular carcinoma and cholangiocarcinoma. Tumour Biol. 2016;37:12755–66.

[47] Huang QX, Cui JY, Ma H, et al. Screening of potential biomarkers for cholangiocarcinoma by integrated analysis of microarray data sets. Cancer Gene Ther. 2016;23:48–53.

[48] Li L, Che L, Tharp KM, et al. Differential requirement for de novo lipogenesis in cholangiocarcinoma and hepatocellular carcinoma of mice and humans. Hepatology. 2016;63:1900–13.

[49] Lu D, Akanno EC, Crowley JJ, et al. Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes. J Anim Sci. 2016;94:1342–53.

[50] Nakagawa R, Hiep NC, Ouchi H, et al. Expression of fatty-acid-binding protein 5 in intrahepatic and extrahepatic cholangiocarcinoma: the possibility of different energy metabolisms in anatomical location. Med Mol Morphol. 2020;53:42–9.

[51] Kobuke T, Tazuma S, Hyogo H, et al. A Ligand for peroxisome proliferator-activated receptor gamma inhibits human cholangiocarcinoma cell growth: potential molecular targeting strategy for cholangioma. Dig Dis Sci. 2006;51:1650–7.

[52] Han C, Demetris AJ, Michalopoulos GK, et al. PPARgamma ligands inhibit cholangiocarcinoma cell growth through p53-dependent GADD45 and p21 pathway. Hepatology. 2003;38:167–77.

[53] Okano H, Shiraki K, Inoue H, et al. The PPARgamma ligand, 15-Deoxy-Delta12,14-PGJ2, regulates apoptosis-related protein expression in cholangio cell carcinoma cells. Int J Mol Med. 2003;12:867–70.