# An Integrated mRNA-lncRNA Signature for Overall Survival Prediction in Cholangiocarcinoma

**Liping Zeng**
Nanchang University - Qianhu Campus: Nanchang University
**Robert Mukiibi**
University of Edinburgh
**Derong Xu**
Nanchang University - Qianhu Campus: Nanchang University
**Hongbo Xin**
Nanchang University - Qianhu Campus: Nanchang University
**Feng Zhang** ( fengzhang0709@hotmail.com )
Nanchang University https://orcid.org/0000-0003-2723-4037

**Research article**

**Keywords:** Cholangiocarcinoma, prognostic signature, Cox regression model, WGCNA

**Posted Date:** November 2nd, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-99682/v1

# Abstract

Background

The incidence and mortality rate of cholangiocarcinoma (CCA) have been rising globally. Patients with CCA have extremely poor prognosis, partly due to the silent clinical character and hence diagnosed at advantage stage without effective treatments. There is growing evidence showing that aberrant expression of messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs) are involved in tumorigenesis and development of CCA. It is essential to establish an integrated mRNA-lncRNA signature to improve the ability of prognostic prediction in CCA patients.

Methods

We collected a training dataset of 45 patients from The Cancer Genome Atlas dataset and a validation cohort (GSE107943) of 57 patients from Gene Expression Omnibus. An integrated mRNA-lncRNA risk score was established by a univariate and a multivariate Cox regression analyses. Time-dependent receiver operating characteristic (ROC) analysis was used to evaluate prognostic performance. Moreover, we conducted a correlation analysis between the signature and different clinical characteristics, and preformed weighted gene co-expression network analysis (WGCNA) and functional enrichment analysis to investigate functional roles of the integrated signature.

Results

A total of two mRNAs (*CFHR3* and *PIWIL4*) and two 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 = 1.12E-02). The ROC curve suggested the integrated mRNA-lncRNA signature possessed a high specificity and sensitivity of prognostic prediction with an area under the curve (AUC) of 0.872 and 0.790 at 1-year and 3-years, respectively. Subsequently, the signature was validated in GSE107943 cohort and combined dataset, and an area under the ROC curve reached up to 0.750 and 0.819 at 1-year. The signature was not only independent from different clinical features (P-value= 1.12E-02), but also outperformed other clinical characteristics as prognostic biomarkers with AUC of 0.781 at 3 years. These molecules in the integrated signature may associated with metabolic-related biological process and lipid metabolism pathway, which was highly involved in CCA carcinogenesis.

Conclusion

These results showed that the integrated mRNA-lncRNA signature had an independent prognostic value for risk stratification, and further facilitated personalized treatment for CCA patients.


# Background

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 (ECCA) [1]. According to epidemiological reports in the past few

decades, CCA is the second most common primary hepatic neoplasm, and its incidence and mortality rate have been rising globally [2-4]. Surgical resection is the only potentially curative treatment for CCA patients, however, the 5-year survival rate of CCA patients remains poor (<20%) [5]. It is also worth noting that most of the patients diagnosed at advanced stage resulting from asymptomatic conditions have a 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 patients, the commonly used tumor-node-metastasis (TNM) staging system has shown valuable but insufficient accuracy for prognostic evaluation [8]. Therefore, there is an urgent and critical need to develop novel and promising prognostic signatures for CCA patients to distinguish risk stratification and consequently contribute to personalized treatments and follow-up plans.

It is widely acknowledged that messenger RNAs (mRNAs) and long noncoding RNAs (lncRNAs) play important roles in the development and progression of various tumors [9]. Numerous reports have indeed detected molecular signatures to predict survival outcome for different cancers, for example, identified mRNAs signatures for non-small cell lung cancer [10], colorectal cancer [11], and glioblastoma [12] and identified lncRNAs signatures for head and neck squamous cell carcinoma (HNSCC) [13], gastric cancer [14], and hepatocellular carcinoma [15]. Given the functional roles of mRNAs and lncRNAs in carcinogenesis and progression, combining them as a prognosis signature has become a better classifier as reported in triple-negative breast cancer [16] and colon cancer [17]. Regarding CCA, currently no effective clinical biomarker is available for classifying risk stratification. Literature about clinical biomarkers is currently scarce. According to our current knowledge, a three-miRNA signature for prognosis [18] and a 7-mRNA biomarker for recurrence-free survival prediction [19] have been identified recently. However, investigation into the potential of combining both mRNAs and lncRNAs expression across whole transcriptome to predict overall survival in CCA has not been conducted.

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

# Methods

### 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 collected from the TCGA portal on July 3, 2019 [20]. The external validation cohort (GEO Accession: GSE107943) [21] with 30 CCA tumor tissues and 27 adjacent normal tissues was retrieved from Gene Expression Omnibus (GEO) database [22]. All samples from the two 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. All the samples possessed mRNA and lncRNA expression data and complete survival information including survival status, survival time, and classic clinicopathological features in both training and validation datasets. 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.

### Screening of differentially expressed RNAs

The expression profiles of 20271 mRNAs and 14852 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 17010 mRNAs and 6390 lncRNAs for differential expression RNAs screening. The edgeR and DESeq2 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 (FDR) as proposed by the Benjamini-Hochberg procedure to limit the occurrence rate of false positives [23]. The RNAs were considered as differently expressed at a threshold of $|\log_2(\text{fold change})| \geq 1.5$ and FDR < 0.05. The overlapped DERNAs obtained by edgeR and DESeq2 were employed to further analysis.

### The mRNA-lncRNA signature construction and validation

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

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

Where n is the number of molecules used to predict 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 two criteria: firstly, it possessed mRNA or lncRNA expression data of normal or adjacent tissues;

secondly, their sample size was more than 3. The six datasets (GEO Accessions: GSE26566, GSE57555, GSE31370, GSE76297, and GSE32879) with a total of 413 individuals including 228 tumor tissues and 185 normal or adjected tissues were collected. 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 [24]. The inconsistency ($I^2$) test and the Cochran's Q test were utilized to assess heterogeneity. When $I^2$ was greater than 50% or P-value was lower than 0.01, 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 generally expression differences of selected biomarkers cross various CCA groups.

The optimal cut-off 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” of the survMisc package in R. The cut point was chosen by hazard function with the maximal sensitivity and specificity for survival rate [25]. 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 [26] 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 data comprised of TCGA_CHOL training cohort and GSE107943 validation dataset through the KM method and ROC analysis, respectively.

### Weighted gene co-expression network analysis and functional enrichment analysis

To investigate the functional roles behind the integrated mRNA-lncRNA signature, we conducted Pearson’s correlation coefficient between the prognostic RNAs and all mRNAs across whole transcriptome in TCGA_CHOL dataset. P-value < 0.05 and correlation coefficient > 0.6 were chosen as the threshold to filter the co-expression mRNAs with identified mRNAs and lncRNAs in the signature. Subsequently, weighted gene co-expression network analysis (WGCNA) was performed on co-expression mRNAs and prognostic mRNAs to explore co-expression modules associated with the risk model using WGCNA package [27]. Soft power parameter with 12 was used to construct the topological overlap matrix (TOM), and a dynamic hybrid cut method with a minimum module size of 30 genes was implemented to detect co-expression clusters. The relationship between module eigengene (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 submitted to functional enrichment analysis of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) through the clusterProfiler R package [28]. The cut-off of q-value < 0.05 was used to identify both significantly enriched GO terms and KEGG pathways.

## Results

Clinical characteristics of CCA datasets

Two independent datasets of CCA were collected in this study. The clinical characteristic of all CCA patients was 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.

Differentially expressed mRNAs and lncRNAs in CCA

Based on the expression profiles of 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 (Figure 1A) and 1950 DElncRNAs (Figure 1B) were identified by DESeq2, whereas 4907 DEmRNAs (Figure 1C) and 2216 DElncRNAs (Figure 1D) were detected by edgeR. The 4628 DEmRNAs (Figure 1E) and 1810 DElncRNAs (Figure 1F) identified by both packages were utilized for further downstream analyses. The heatmaps showed that CCA samples were clearly distinguished from normal tissues based on the top 200 DEmRNAs and DElncRNAs (Figure S1).

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 4628 DEmRNAs and 1810 DElncRNAs in the discovery set. A total of six mRNAs (*ACRV1, TMEM121B, PIWIL4, GOLGA8M, CFHR3,* and *FUT4*) and three lncRNAs (*AC134682.1, AC007285.1,* and *AC138430.1*) with P-value < 0.01 were determined to be significantly associated with overall survival (Table S1). The six mRNAs and three lncRNAs were further subjected to a stepwise multivariate Cox regression analysis. The optimally integrated mRNA-lncRNA signature was determined with the lowest value of Akaike information criterion [29], which included two mRNAs (*CFHR3* and *PIWIL4*) and two lncRNAs (*AC007285.1* and *AC134682.1*). The chromosomal position, hazard ratio, P-value, and coefficient of these four prognostic RNAs in CCA are provided in Table 2. Among these four RNAs, only *CFHR3* had a positive coefficient, which suggested that higher expression was related with shorter survival, whereas the remaining three RNAs were protective factors as their negative coefficients implied that higher expression level was associated with higher survival rate. We subsequentially performed a meta-analysis for *CFHR3* and *PIWIL4* using random effect model due to the P-value of heterogeneity test less than 0.05 as shown in Figure S2. 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 as these mRNAs that 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 two mRNAs and the two lncRNAs weighted by their regression coefficients from above multivariate 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 then were classified into high-risk group (23 patients) and low-risk group (13 patients) based on the optimal cut-off point (-0.14) determined by “cutp” function from survMisc package (Figure 2A). The survival status and the expression pattern of the four prognostic RNAs for each CCA patient in the discovery cohort are presented in Figure 2A as well. The Kaplan–Meier 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 (Figure 2B). 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 with the low-risk group for OS. Time-dependent ROC curve for the risk score model in the training cohort is shown in Figure 2C, with an area under the curve (AUC) of 0.872 and 0.790 for 1- and 3-year OS prediction, respectively, which implied that the integrated mRNA-lncRNA signature possessed a high specificity and sensitivity.

Validation for the prognostic prediction value of integrated mRNA-lncRNA signature in the independent validation cohort and the combined dataset

To evaluate robustness of the integrated mRNA-lncRNA signature for prognosis in CCA patients, we validated its prognostic ability in an independent cohort (GSE107943) 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 high-risk group patients was significantly worse than that of low-risk group (P-value = 1.71E-04) as shown in Figure 3B. Notably, there were 14 deaths in patients with high-risk scores and only three death events in low-risk group (Figure 3A). The hazard ratio of high-risk group was 8.04 folds compared with 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- and 3-year overall survival prediction (Figure 3C), representing 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, which was consistent with the findings from the discovery or validation cohort. The principal components analysis (PCA) for the combined samples showed that there was clearly divergence between the training and validation individuals (Figure S3A), which suggests that the batch effect existed in the combined data. To eliminate bias caused by the batch effect, we fitted it as a fixed effect in the formula design generated by DESeq2 package. As a result, the batch effect was adjusted properly (Figure S3B-C). The patients were then divided into high-risk group (n =34) and low-risk group (n = 32) according to the same risk score model and criteria. Kaplan-Meier survival curves between two risk groups were significantly different in the combined dataset with a P-value of 5.51E-06 (Figure 4B). The survival rates at 3- and 5-years were 26.47% and 23.53% for patients in the high-risk group, compared to 87.50% and 75.00% survival rate for patients in the low-risk at 3- and 5-years respectively. Patients with high-risk scores exhibited a 5.27-fold increased risk than patients in low-risk group (Table 3). Results of time-dependent ROC curve analysis similar to those obtained from the training and validation datasets are presented in Figure 4C with AUC equal to 0.819 for 1-year and 0.781 for 3-years OS prediction.

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= 1.12E-02) compared with the other clinical characteristics. Furthermore, after adjusting the age,

gender, and tumor stage, we found 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 dataset, respectively (Table 3).

Besides, the tumor stage was relatively associated with OS in the validation (P-value= 5.65E-02) and combined (P-value= 5.88E-02) dataset (Table 3). The stratification analysis was carried out to estimate the relationship of the mRNA-lncRNA signature with tumor stage. All patients were classified into two subgroups: I/II stage with 49 samples and III/IV stage with 17 individuals. As shown in Figure 5A-B, KM curve observed patients with high-risk scores have significantly shorter survival time than that with low-risk scores in stage I/II (P-value = 4.71E-04) and stage III/IV (P-value = 4.97E-03) subgroup. Accordingly, the multivariate Cox and stratification analysis demonstrated that prognostic power 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 set, 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) (Figure 5C). These results demonstrated that the mRNA-lncRNA signature had a better prognostic power than other factors including tumor stage, age, and gender.

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 grey module with three mRNAs. The turquoise module showed a higher correlation with risk score model than other modules, which had a high correlation coefficient of 0.87 (P-value=8.00E-12) with risk scores (Figure 6A-B). We then carried out GO and KEGG enrichment analyses based on the 522 genes from the turquoise module. These 522 genes significantly enriched in 567 GO terms and 48 KEGG pathways. The top 10 GO biological processes and KEGG pathways are shown in Figures 6C-D. These genes were mostly enriched in catabolic or metabolic biological processes, such as 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, glycine, serine and threonine metabolism, and primary bile acid biosynthesis. In addition, complement and coagulation cascades and PPAR signaling pathways were also enriched by these turquoise module genes.

# Discussion

The TNM 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 [30]. Up to the time of conducting this current study, only a few studies have performed using high-throughput sequencing data to identify more powerful molecular biomarkers for CCA prognosis. For example, miRNAs have been identified as prognostic markers by Cao et al. [18]. They discovered three miRNAs (*miR-10b, miR-22,* and *miR-551b*) that showed a relatively precise prediction with an AUC of 0.715

for 1-year and 0.723 for 3-years (Figure S4). 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 two mRNAs (*CFHR3* and *PIWIL4*) and two lncRNAs (*AC007285.1* and *AC134682.1*) that could potentially be used for CCA patients' prognosis. The signature was further confirmed in the independent validation and complete dataset, and performed well in 1- and 3- year survival prediction according to time-dependent ROC curve (Figure 2C, 3C, and 4C). Compared with the previous miRNA prognostic signature [18], prediction accuracy of our model showed improvement compared to that of [18] with higher AUC of 1-year (0.872 vs. 0.715) and 3-year (0.790 vs. 0.723) survival prediction in CCA patients from TCGA_CHOL cohort (Figure S4). The multivariable Cox regression and stratified analysis revealed that our integrated mRNA-lncRNA signature had the independent prognostic ability from other clinical features.

The relationship between these prognosis 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 mRNAs identified in our signature indicate that overexpressed *CFHR3* (Complement Factor H-Related Protein 3) would suppress proliferation and promote apoptosis of hepatocellular carcinoma (HCC) cells [31] and is a potential prognosis biomarker for HCC [32]. *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 [33], retinoblastoma [34], and hepatocellular carcinoma [35]. As for the two antisense lncRNAs identified in this study, their functional roles have not been elucidated in any cancer. 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 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 [36]. Various studies based on multiple CCA independent cohorts [37-39] also detected that metabolic-related biological processes including small molecule and lipid metabolic processes relate to energy metabolism were pivotal for CCA development. Increasing evidence has demonstrated that fatty acids synthesis related genes (*FASN* and *SLC27A1*) [40, 41], fatty acid transport proteins (*FATP2*, *FATP1*, *FATP5,* and *CD36*), and fatty acid binding proteins (*FABP1*, *FABP4*, and *FABP5*) [42] contribute to CCA carcinogenesis. Additionally, PPARγ ligands suppressed cholangiocarcinoma cell growth [43, 44] and induced the cholangiocarcinoma cell apoptosis [45], 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 genome-wide 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, the RNA expression level was quantified based on CCA tissues, which may not precisely predict prognosis when body fluids (saliva, serum, urine, or stool) are commonly used in clinical application. Hence, collecting more CCA samples and verifying prognostic value of the risk score model in samples from body fluids are necessary for further research endeavors.

# Conclusion

In conclusion, we performed a comprehensive analysis to develop an integrated mRNA-lncRNA signature for CCA patients' prognosis. The signature consisting of two mRNAs (*CFHR3* and *PIWIL4*) and two 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 may serve as a valuable and alternative biomarker for CCA patients.

# Declarations

Availability of data and materials

The raw data involved in the current study are publicly available in TCGA (https://portal.gdc.cancer.gov/) and GEO (https://www.ncbi.nlm.nih.gov/geo/).

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This study was supported by the National Natural Science Foundation of China (No.31601034 to XDR).

Author Contributions

ZF conceived and designed the study; ZLP, MR, and XDR analyzed the data; ZF, ZLP, MR and XHB wrote and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We are deeply grateful to all the subjects for their participation in the study.

# References

1. Khan SA, Thomas HC, Davidson BR, Taylor-Robinson SD: **Cholangiocarcinoma**. *Lancet (London, England)* 2005, **366**(9493):1303-1314.
2. Kirstein MM, Vogel A: **Epidemiology and Risk Factors of Cholangiocarcinoma**. *Visceral medicine* 2016, **32**(6):395-400.
3. Blechacz B: **Cholangiocarcinoma: Current Knowledge and New Developments**. *Gut Liver* 2017, **11**(1):13-26.

4. Saha SK, Zhu AX, Fuchs CS, Brooks GA: **Forty-Year Trends in Cholangiocarcinoma Incidence in the U.S.: Intrahepatic Disease on the Rise**. *Oncologist* 2016, **21**(5):594-599.
5. Squadroni M, Tondulli L, Gatta G, Mosconi S, Beretta G, Labianca R: **Cholangiocarcinoma**. *Crit Rev Oncol Hematol* 2017, **116**:11-31.
6. Valle J, Wasan H, Palmer DH, Cunningham D, Anthoney A, Maraveyas A, Madhusudan S, Iveson T, Hughes S, Pereira SP *et al*: **Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer**. *The New England journal of medicine* 2010, **362**(14):1273-1281.
7. Okusaka T, Nakachi K, Fukutomi A, Mizuno N, Ohkawa S, Funakoshi A, Nagino M, Kondo S, Nagaoka S, Funai J *et al*: **Gemcitabine alone or in combination with cisplatin in patients with biliary tract cancer: a comparative multicentre study in Japan**. *British journal of cancer* 2010, **103**(4):469-474.
8. Blechacz B, Komuta M, Roskams T, Gores GJ: **Clinical diagnosis and staging of cholangiocarcinoma**. *Nat Rev Gastroenterol Hepatol* 2011, **8**(9):512-522.
9. Valastyan S, Weinberg RA: **Tumor metastasis: molecular insights and evolving paradigms**. *Cell* 2011, **147**(2):275-292.
10. Zuo S, Wei M, Zhang H, Chen A, Wu J, Wei J, Dong J: **A robust six-gene prognostic signature for prediction of both disease-free and overall survival in non-small cell lung cancer**. *Journal of translational medicine* 2019, **17**(1):152.
11. Zuo S, Dai G, Ren X: **Identification of a 6-gene signature predicting prognosis for colorectal cancer**. *Cancer cell international* 2019, **19**:6.
12. Yang J, Wang L, Xu Z, Wu L, Liu B, Wang J, Tian D, Xiong X, Chen Q: **Integrated Analysis to Evaluate the Prognostic Value of Signature mRNAs in Glioblastoma Multiforme**. *Frontiers in genetics* 2020, **11**:253.
13. Wang P, Jin M, Sun CH, Yang L, Li YS, Wang X, Sun YN, Tian LL, Liu M: **A three-lncRNA expression signature predicts survival in head and neck squamous cell carcinoma (HNSCC)**. *Bioscience reports* 2018, **38**(6):BSR20181528.
14. Song P, Jiang B, Liu Z, Ding J, Liu S, Guan W: **A three-lncRNA expression signature associated with the prognosis of gastric cancer patients**. *Cancer medicine* 2017, **6**(6):1154-1164.
15. Liao X, Yang C, Huang R, Han C, Yu T, Huang K, Liu X, Yu L, Zhu G, Su H *et al*: **Identification of Potential Prognostic Long Non-Coding RNA Biomarkers for Predicting Survival in Patients with Hepatocellular Carcinoma**. *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology* 2018, **48**(5):1854-1869.
16. Liu YR, Jiang YZ, Xu XE, Hu X, Yu KD, Shao ZM: **Comprehensive Transcriptome Profiling Reveals Multigene Signatures in Triple-Negative Breast Cancer**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2016, **22**(7):1653-1662.
17. Dai W, Feng Y, Mo S, Xiang W, Li Q, Wang R, Xu Y, Cai G: **Transcriptome profiling reveals an integrated mRNA-lncRNA signature with predictive value of early relapse in colon cancer**. *Carcinogenesis* 2018, **39**(10):1235-1244.
18. Cao J, Sun L, Li J, Zhou C, Cheng L, Chen K, Yan B, Qian W, Ma Q, Duan W: **A novel three-miRNA signature predicts survival in cholangiocarcinoma based on RNASeq data**. *Oncology reports* 2018, **40**(3):1422-1434.

19. Guo H, Cai J, Wang X, Wang BR, Wang F, Li X, Qu XY, Kong XM, Gao YQ, Wu HL *et al*: **Prognostic values of a novel multi-mRNA signature for predicting relapse of cholangiocarcinoma**. *International journal of biological sciences* 2020, **16**(5):869-881.
20. Balbin OA, Malik R, Dhanasekaran SM, Prensner JR, Cao X, Wu YM, Robinson D, Wang R, Chen G, Beer DG *et al*: **The landscape of antisense gene expression in human cancers**. *Genome research* 2015, **25**(7):1068-1079.
21. Ahn KS, Kang KJ, Kim YH, Kim TS, Song BI, Kim HW, O'Brien D, Roberts LR, Lee JW, Won KS: **Genetic features associated with (18)F-FDG uptake in intrahepatic cholangiocarcinoma**. *Annals of surgical treatment and research* 2019, **96**(4):153-161.
22. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M: **NCBI GEO: archive for functional genomics data sets–update**. *Nucleic Acids Research* 2013, **41**(D1):D991-D995.
23. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995, **57**(1):289-300.
24. Schwarzer G, Carpenter J, Rücker G: **Meta-Analysis with R**. Switzerland: Springer International Publishing; 2015.
25. Piti CHU, Cedex P, Diego S: **An application of changepoint methods in studying the effect of age on survival in breast cancer**. *Computational Statistics & Data Analysis* 1999, **30**(3):253-270.
26. 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**. *Statistics in Medicine* 2013, **32**(30):5381-5397.
27. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC bioinformatics* 2008, **9**:559.
28. Yu GC, Wang LG, Han YY, He QY: **clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters**. *Omics-a Journal of Integrative Biology* 2012, **16**(5):284-287.
29. Venables WN, Ripley BD: **Modern Applied Statistics with S**, 4th ed edn. New York: Springer Publishing Company, Incorporated; 2010.
30. Zheng B, Jeong S, Zhu Y, Chen L, Xia Q: **miRNA and lncRNA as biomarkers in cholangiocarcinoma(CCA)**. *Oncotarget* 2017, **8**(59):100819-100830.
31. Liu H, Zhang L, Wang P: **Complement factor Hrelated 3 overexpression affects hepatocellular carcinoma proliferation and apoptosis**. *Molecular medicine reports* 2019, **20**(3):2694-2702.
32. Liu J, Li WL, Zhao HT: **CFHR3 is a potential novel biomarker for hepatocellular carcinoma**. *J Cell Biochem* 2020, **121**(4):2970-2980.
33. Heng ZSL, Lee JY, Subhramanyam CS, Wang C, Thanga LZ, Hu Q: **The role of 17betaestradiolinduced upregulation of Piwilike 4 in modulating gene expression and motility in breast cancer cells**. *Oncology reports* 2018, **40**(5):2525-2535.

34. Sivagurunathan S, Arunachalam JP, Chidambaram S: **PIWI-like protein, HIWI2 is aberrantly expressed in retinoblastoma cells and affects cell-cycle potentially through OTX2**. *Cellular & molecular biology letters* 2017, **22**:17.
35. Zeng G, Zhang D, Liu X, Kang Q, Fu Y, Tang B, Guo W, Zhang Y, Wei G, He D: **Co-expression of Piwil2/Piwil4 in nucleus indicates poor prognosis of hepatocellular carcinoma**. *Oncotarget* 2017, **8**(3):4607-4617.
36. Pastore M, Lori G, Gentilini A, Taddei ML, Di Maira G, Campani C, Recalcati S, Invernizzi P, Marra F, Raggi C: **Multifaceted Aspects of Metabolic Plasticity in Human Cholangiocarcinoma: An Overview of Current Perspectives**. *Cells* 2020, **9**(3):596.
37. Tian A, Pu K, Li B, Li M, Liu X, Gao L, Mao X: **Weighted gene coexpression network analysis reveals hub genes involved in cholangiocarcinoma progression and prognosis**. *Hepatol Res* 2019, **49**(10):1195-1206.
38. Likhitrattanapisal S, Tipanee J, Janvilisri T: **Meta-analysis of gene expression profiles identifies differential biomarkers for hepatocellular carcinoma and cholangiocarcinoma**. *Tumour Biol* 2016, **37**(9):12755-12766.
39. Huang QX, Cui JY, Ma H, Jia XM, Huang FL, Jiang LX: **Screening of potential biomarkers for cholangiocarcinoma by integrated analysis of microarray data sets**. *Cancer Gene Ther* 2016, **23**(2-3):48-53.
40. Li L, Che L, Tharp KM, Park HM, Pilo MG, Cao D, Cigliano A, Latte G, Xu Z, Ribback S *et al*: **Differential requirement for de novo lipogenesis in cholangiocarcinoma and hepatocellular carcinoma of mice and humans**. *Hepatology* 2016, **63**(6):1900-1913.
41. Li L, Pilo GM, Li X, Cigliano A, Latte G, Che L, Joseph C, Mela M, Wang C, Jiang L *et al*: **Inactivation of fatty acid synthase impairs hepatocarcinogenesis driven by AKT in mice and humans**. *J Hepatol* 2016, **64**(2):333-341.
42. Nakagawa R, Hiep NC, Ouchi H, Sato Y, Harada K: **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**(1):42-49.
43. Kobuke T, Tazuma S, Hyogo H, Chayama K: **A Ligand for peroxisome proliferator-activated receptor gamma inhibits human cholangiocarcinoma cell growth: potential molecular targeting strategy for cholangioma**. *Dig Dis Sci* 2006, **51**(9):1650-1657.
44. Han C, Demetris AJ, Michalopoulos GK, Zhan Q, Shelhamer JH, Wu T: **PPARgamma ligands inhibit cholangiocarcinoma cell growth through p53-dependent GADD45 and p21 pathway**. *Hepatology* 2003, **38**(1):167-177.
45. Okano H, Shiraki K, Inoue H, Kawakita T, Deguchi M, Sugimoto K, Sakai T, Murata K, Nakano T, Enjoji M: **The PPARgamma ligand, 15-Deoxy-Delta12,14-PGJ2, regulates apoptosis-related protein expression in cholangio cell carcinoma cells**. *Int J Mol Med* 2003, **12**(6):867-870.

# Tables

**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%) |
| **residual tumor** | R0 | 28 (77.78%) | / | / |
| | R1/RX | 8 (22.22%) | / | / |
| **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.

**Table 2. The 4 prognostic RNAs significantly associated with the overall survival in CCA patients**

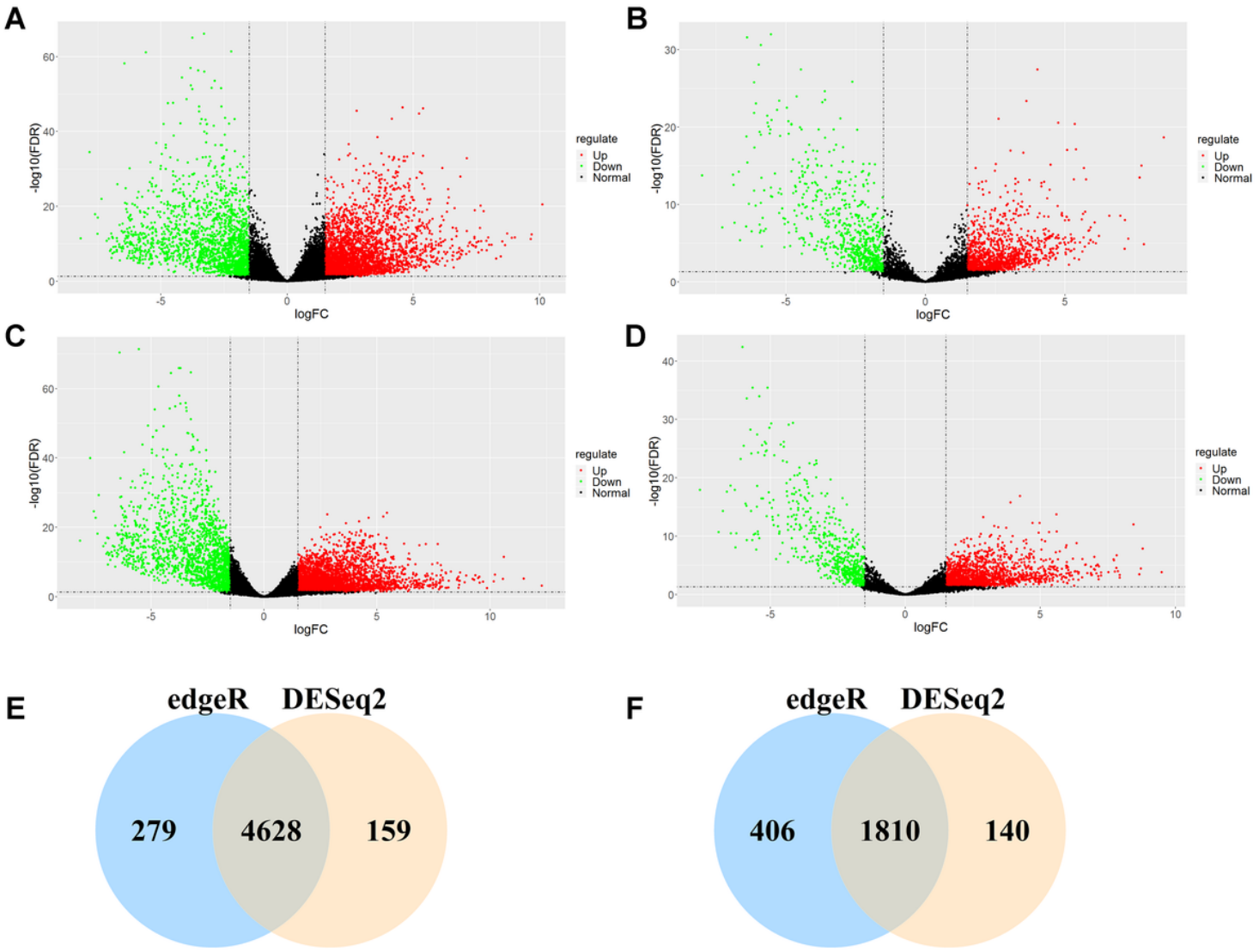
| Ensemble ID | Gene name | Chromosomal position | Gene type | HR | P-value | Coefficient |
|---|---|---|---|---|---|---|
| ENSG00000116785 | CFHR3 | chr1: 196774795-196795406 (+) | protein_coding | 24.13 | 5.24E-04 | 3.18 |
| ENSG00000134627 | PIWIL4 | chr11: 94543840-94621421 (+) | protein_coding | 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.

**Table 3. Univariate and multivariate Cox regression analysis of integrated mRNA–lncRNA signature in different dataset**

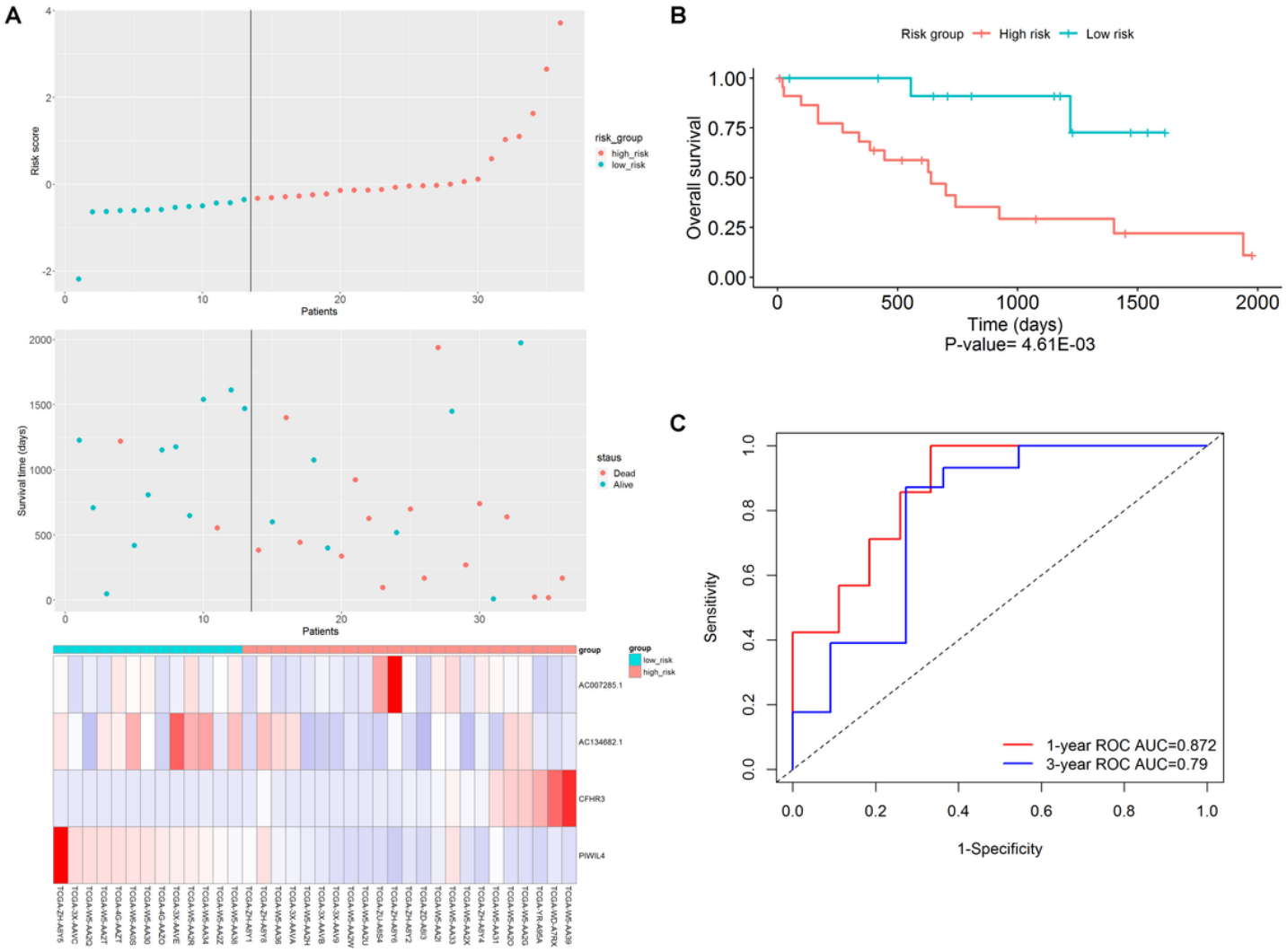
| | Favorable/Unfavorable | Univariate analysis | | | Multivariate analysis | | | |
|---|---|---|---|---|---|---|---|---|
| Variables | | HR | 95% CI | P-value | HR | 95% CI | P-value | coef |
| **Training set (n = 36)** | | | | | | | | |
| risk group | low/high | 6.46 | 1.47 - 28.39 | 1.35E-02 | 8.99 | 1.65 - 49.05 | 1.12E-02 | 2.20 |
| age | <60/>=60 | 0.73 | 0.28 - 1.92 | 5.19E-01 | 0.81 | 0.28 - 2.33 | 6.90E-01 | -0.22 |
| gender | female/male | 1.39 | 0.54 - 3.53 | 4.94E-01 | 0.84 | 0.27 - 2.60 | 7.67E-01 | -0.17 |
| tumor stage | I+II/III+IV | 1.48 | 0.52 - 4.21 | 4.67E-01 | 0.91 | 0.27 - 3.13 | 8.83E-01 | -0.09 |
| residual tumor | R0/R1+RX | 2.75 | 1.01 - 7.49 | 4.72E-02 | 3.99 | 1.28 - 12.50 | 1.74E-02 | 1.38 |
| histologic grade | G3+G4/G1+G2 | 1.64 | 0.62 - 4.32 | 3.21E-01 | 1.89 | 0.66 - 5.40 | 2.35E-01 | 0.64 |
| **Independent validation set (n = 30)** | | | | | | | | |
| risk group | low/high | 8.04 | 2.26 - 28.62 | 1.29E-03 | 7.70 | 1.99 - 29.77 | 3.10E-03 | 2.04 |
| age | <60/>=60 | 0.93 | 0.30 - 2.91 | 8.96E-01 | 0.58 | 0.16 - 2.10 | 4.07E-01 | -0.54 |
| gender | female/male | 1.37 | 0.31 - 6.16 | 6.80E-01 | 1.05 | 0.22 - 5.13 | 9.49E-01 | 0.05 |
| tumor stage | I+II/III+IV | 5.15 | 1.60 - 16.57 | 5.93E-03 | 3.22 | 0.97 - 10.69 | 5.65E-02 | 1.17 |
| **Combined set (n=66)** | | | | | | | | |
| risk group | low/high | 5.27 | 2.38 - 11.67 | 4.23E-05 | 5.27 | 2.34 -11.86 | 6.09E-05 | 1.66 |
| age | <60/>=60 | 0.86 | 0.42 - 1.76 | 6.76E-01 | 0.73 | 0.35 - 1.55 | 4.13E-01 | -0.31 |
| gender | female/male | 1.37 | 0.68 - 2.77 | 3.77E-01 | 1.23 | 0.60 - 2.51 | 5.74E-01 | 0.21 |
| tumor stage | I+II/III+IV | 2.20 | 1.08 - 4.51 | 3.07E-02 | 2.04 | 0.97 - 4.28 | 5.88E-02 | 0.71 |

# Figures



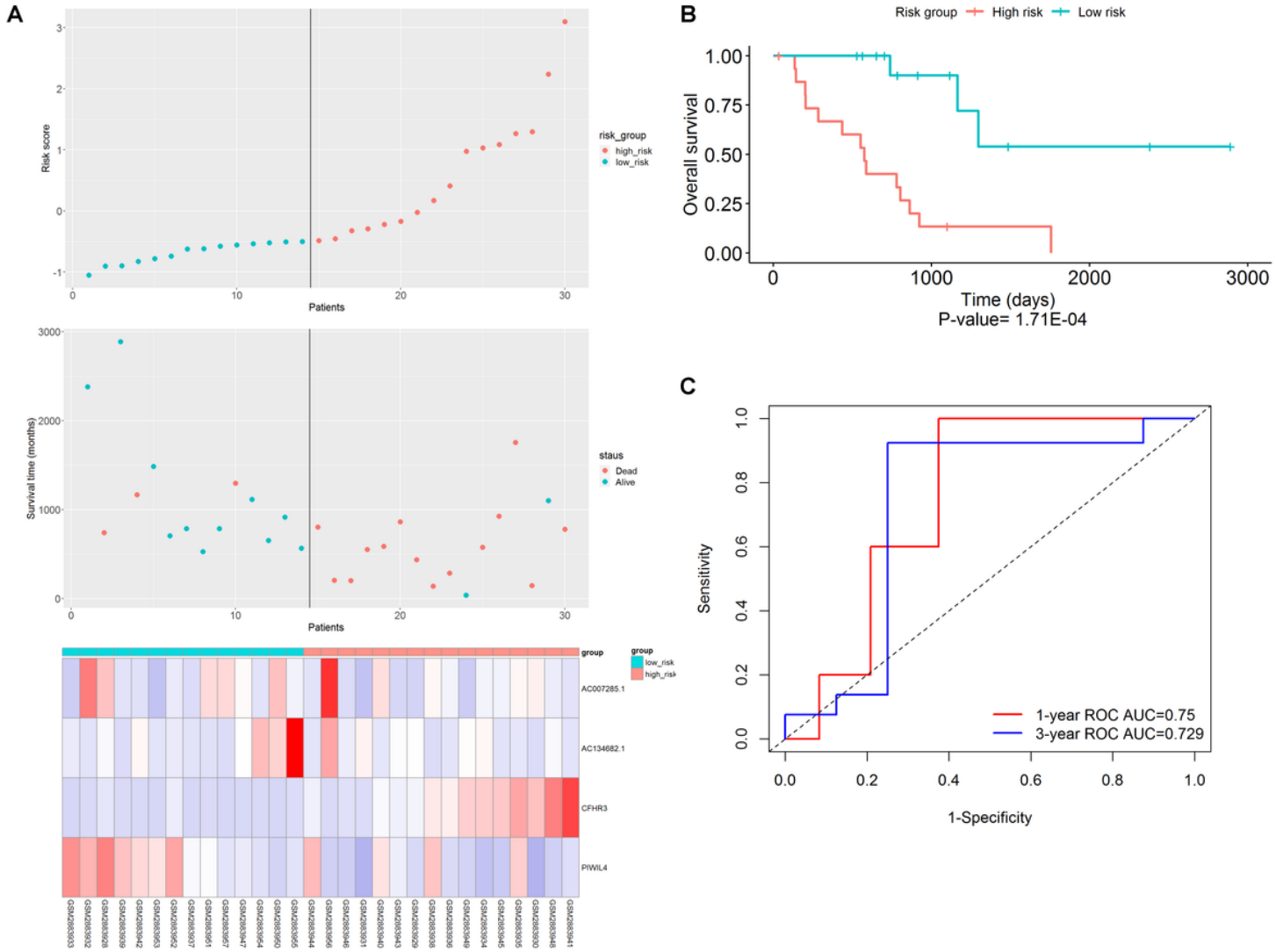

## Figure 1

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.



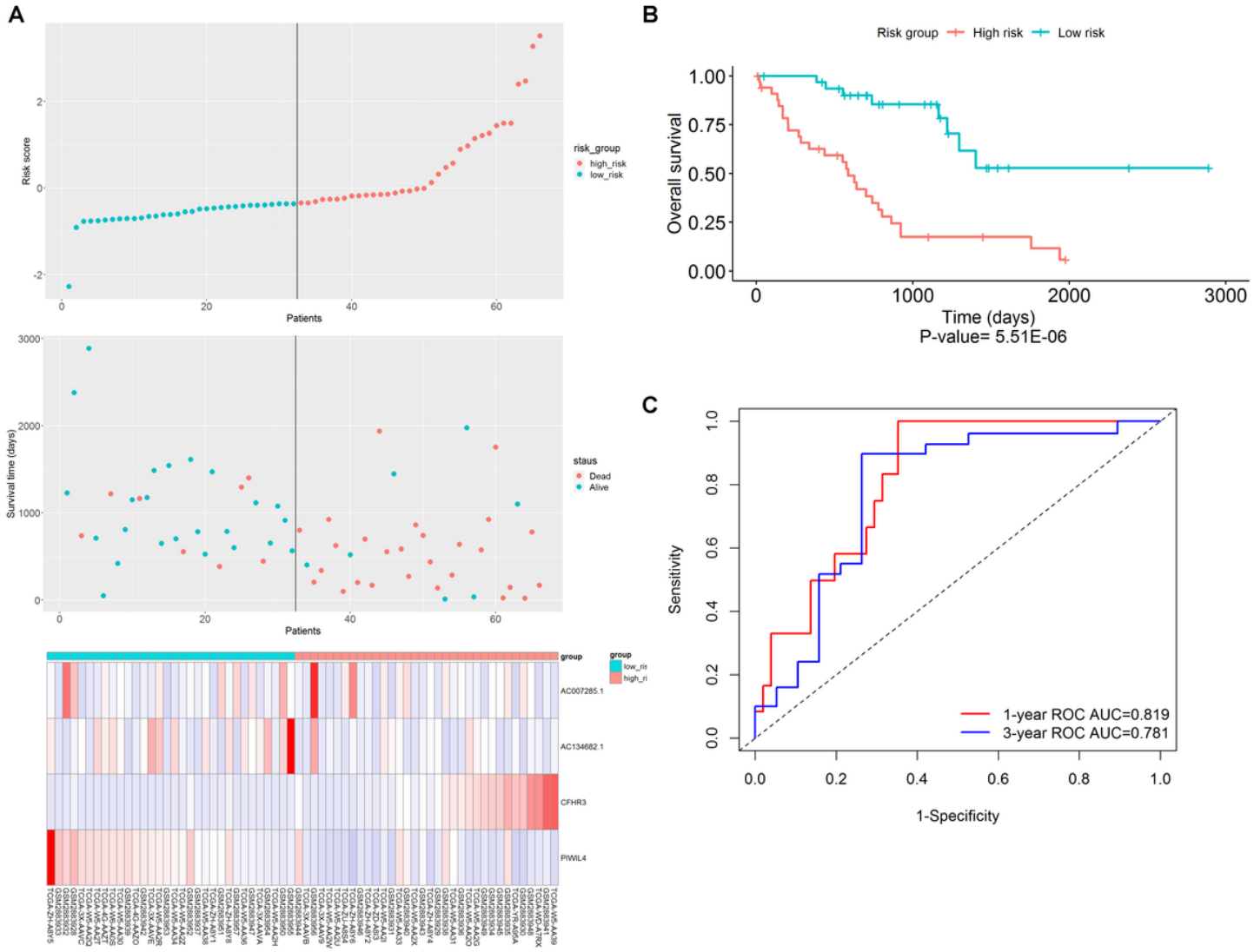

## Figure 2

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 CCA patients between low and high risk groups. (C) Time-dependent receiver operating characteristic (ROC) analysis for overall survival prediction based on the risk scores with one and three years as the time point.



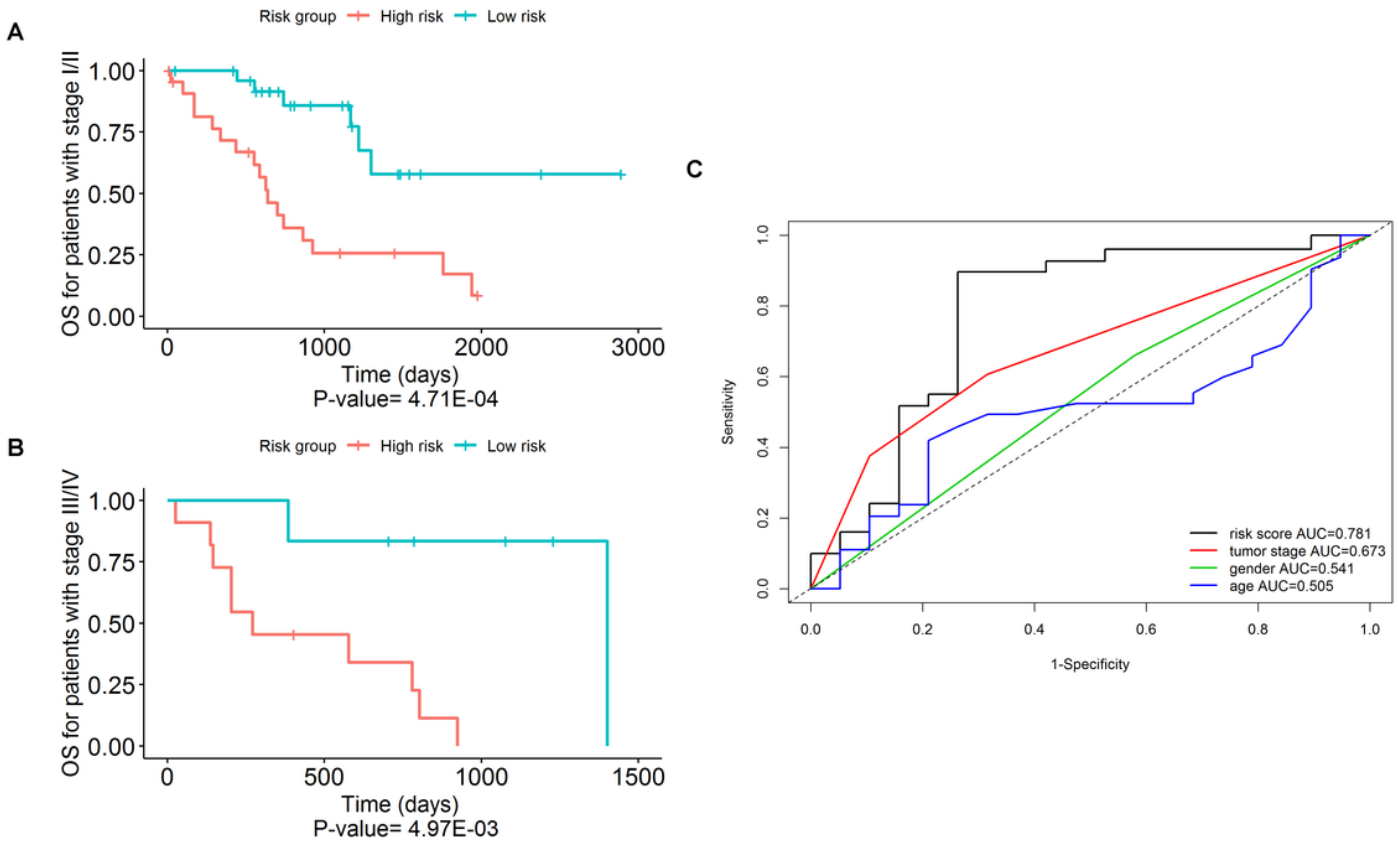

## Figure 3

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 CCA patients. (C) Time-dependent receiver operating characteristic (ROC) analysis for overall survival prediction based on the risk score with one and three years as the time point.



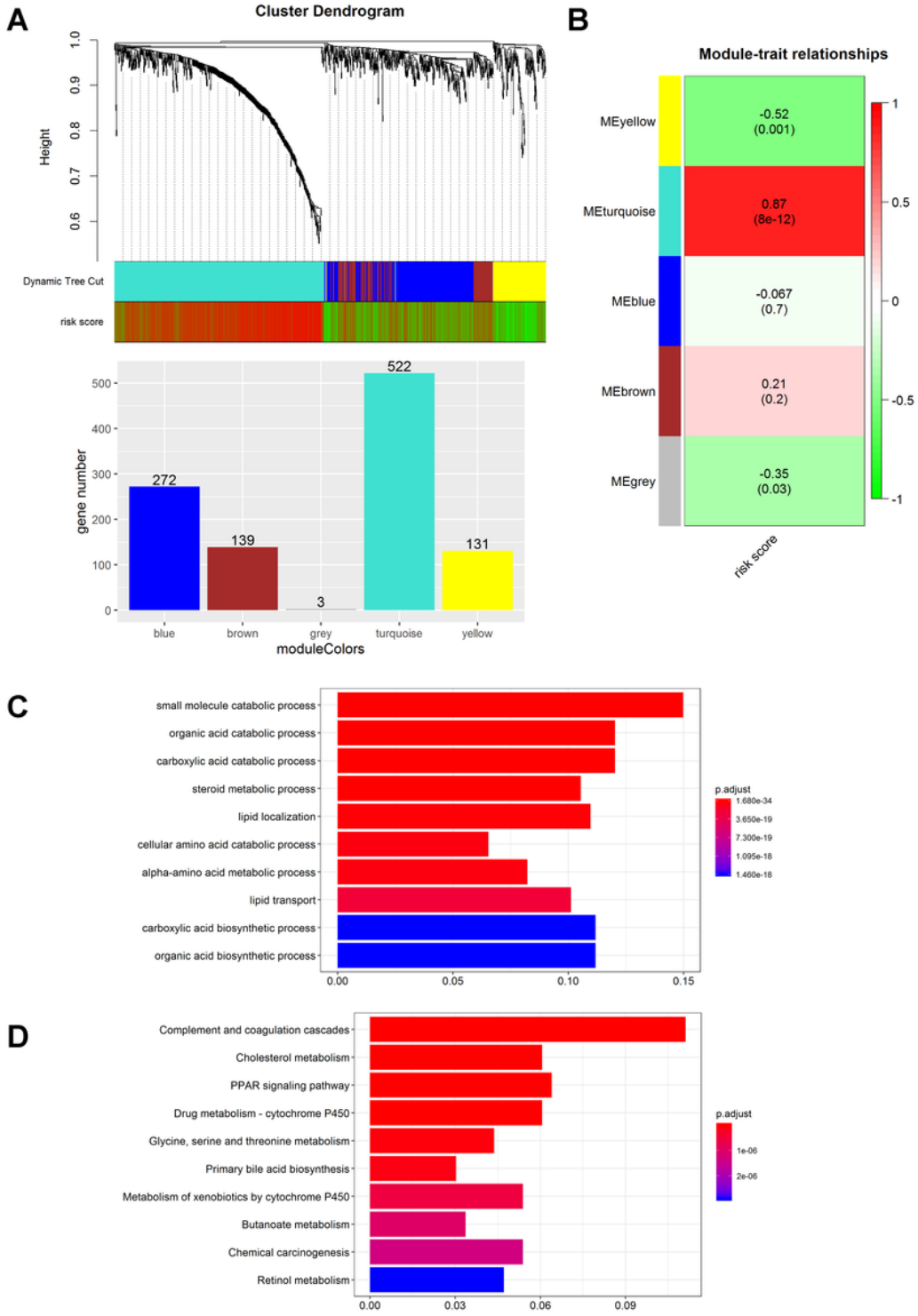

## Figure 4

Prognosis validation of the integrated mRNA-lncRNA signature in the combined dataset. (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 CCA patients. (C) Time-dependent receiver operating characteristic (ROC) analysis for overall survival prediction based on the risk score with one and three 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 by the mRNA-lncRNA signature and other clinical factors in the combined dataset.




## 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 five 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). WGCNA, weighted gene co-expression network analysis; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

## Supplementary Files

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

- Addtionalfiles.docx