Identification of a Three-gene Signature Acting as a Novel Prognostic Biomarker in Intrahepatic Cholangiocarcinoma Patients

CURRENT STATUS: UNDER REVIEW

Cancer Cell International  BMC

Zengwei Tang
Zhejiang University School of Medicine First Affiliated Hospital

Xing Hunag
Zhejiang University School of Medicine First Affiliated Hospital

Enliang Li
Zhejiang University School of Medicine First Affiliated Hospital

Yinan Shen
Zhejiang University School of Medicine First Affiliated Hospital

Qi Zhang
Zhejiang University School of Medicine First Affiliated Hospital

Xuele Bai
Zhejiang University School of Medicine First Affiliated Hospital

Tingbo Liang
The First Affiliated Hospital, School of Medicine, Zhejiang University

✉ liangtingbo@zju.edu.cn Corresponding Author
ORCiD: https://orcid.org/0000-0003-0143-3353

DOI: 10.21203/rs.3.rs-20760/v1

SUBJECT AREAS
Cancer Biology  Oncology

KEYWORDS
intrahepatic cholangiocarcinoma; biomarker; survival; gene signature; hub genes
Abstract

Background: Intrahepatic cholangiocarcinoma (iCCA) patients have poor outcomes due to the lack of biomarkers for the selection of treatment options. The present study was conducted to find biomarkers with independent prognostic value in iCCA patients.

Methods: Gene transcriptome profiles of E-MTAB-6389, TCGA-CHOL and GSE26566 were obtained from ArrayExpress, The Cancer Genome Atlas and the Gene Expression Omnibus databases, respectively. Bioinformatic analyses were performed to screen novel biomarkers for predicting the prognosis of iCCA patients. Using multivariate Cox regression analyses, a 3-gene signature (BTD-FER-COL12A1) with potential prognostic value was identified and validated in both a training cohort and two validation cohorts.

Results: A total of 177 iCCA patients were included in this study. From the key gene modules significantly associated with liver cirrhosis and overall survival (OS) of iCCA patients, we identified 89 hub genes for functional analyses. Cox-regression analyses in both the training and validation cohort indicate that FER, COL12A1 and BTD were independent risk factors for iCCA patients. A 3-gene signature (BTD-FER-COL12A1) with independent prognostic value in iCCA patients was validated in the training cohort, as well as in two validation cohorts. In terms of predicting the prognosis of iCCA patients, the receiver operating characteristics (ROC) curves showed that this 3-gene signature had superior prediction power to BTD, FER, and COL12A1 alone, as well as known biomarkers (MUC1, MUC13) of iCCA. Immunohistochemical staining of samples from The Human Protein Atlas showed that FER and COL12A1 were positively expressed in iCCA tissue, although BTD was not, while none of these genes was detected in normal tissue. These findings were consistent with the expression status of BTD, FER and COL12A1 at the transcriptional level. In addition, we found that FER and COL12A1 were significantly associated with the degree of infiltration by tumor-infiltrating immune cells.

Conclusion: We discovered a three-gene signature with independent prognostic value as a novel biomarker for prediction prognosis of iCCA patients. Our findings may help to find novel therapeutic targets for precision treatment of iCCA.

Background
Intrahepatic cholangiocarcinoma (iCCA) is the second most common primary liver cancer after hepatocellular carcinoma and its incidence and mortality rate has increased worldwide [1]. The introduction of new systemic therapies, including immune-based therapies and biomarker-driven therapies, has significantly improved the survival of patients with hepatobiliary cancers, particularly in the advanced stages of disease [2]. However, the prognosis of iCCA patients remains dismal, and surgical resection with negative resection margins and/or liver transplantation is the only curative treatment option [3, 4].

High-throughput sequencing techniques have helped to provide abundant knowledge of the genomic and epigenetic characterization of cholangiocarcinoma (CCA), which has further clarified the pathogenesis of CCA. Previous studies focusing on this field have provided valuable information on molecular classification, oncogenic pathways and potential therapeutic targets for CCA [5-7]. However, CCAs, arising in the epithelium at every point of the biliary tree, represent a highly heterogeneous group of diseases, with significant intra- and inter-tumor differences in iCCA, perihilar CCA and distal CCA [3]. The mechanisms associated with the carcinogenesis and progression of CCA require further elucidation based on its anatomic classification.

Currently, there are lack of effective biomarkers with high-performance for predicting the diagnosis and prognosis of iCCA patients, as well as for assessing the treatment response [2]. Exploring novel biomarkers involved in the carcinogenesis and progression of iCCA may be beneficial in finding potential therapeutic targets for this life-threatening disease. To discover novel potential biomarkers for predicting the prognosis of iCCA patients, we downloaded the transcriptional profiles of a total of 177 iCCA patients from the E-MTAB-6389, TCGA-CHOL and GSE26566 datasets by searching ArrayExpress, The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases, respectively. Weighted gene co-expression network analysis (WGCNA) was conducted to identify key gene modules associated with clinical traits of iCCA patients. After further functional analyses, we identified a three-gene signature (BTD-FER-COL12A1) and validated its independent prognostic value in both the training cohort and two validation cohorts. In addition, using the resources of The Human Protein Atlas, TIMER and MEXPRESS, we validated the expression status of these hub genes at the
translational level, and explored the correlation between the expression of these hub genes and the gene methylation status. We also assessed the association of the expression level of the hub genes with the degree of infiltration by tumor-infiltrating immune cells.

Methods

Data sources
Gene expression profiles of CCA were obtained from TCGA (TCGA-CHOL; platform: RNA-sequencing), ArrayExpress (E-MTAB-6389; platform: Affymetrix Gene Chip HTA-2_0-EXON level-HTA-2_0.r1) and the GEO (GSE26566; platform: Illumina humanRef-8 v2.0 expression bead chip) databases. The R package “TCGAbiolinks” [8] and “oligo” [9] were used to download and process the raw data of TCGA-CHOL and E-MTAB-6389 datasets, respectively. TCGA-CHOL dataset included a total of 32 iCCA samples, E_MTAB-6389 included 78 iCCA samples, and GSE26566 included 68 iCCA samples.

Identification of stable differentially expressed genes
After the expression matrix of TCGA-CHOL, E-MTAB-6389 and GSE26566 datasets were downloaded from TCGA, ArrayExpress and GEO databases, respectively, the R package “limma” was used to screen differentially expressed genes (DEGs) [10]. Venn diagrams were constructed to identify the intersecting DEGs across these three datasets. The identified genes with adjusted $P$ values < 0.05 were defined as significant DEGs.

Weighted gene co-expression network analysis (WGCNA)
The intersecting DEGs in the expression matrix of E-MTAB-6389 were extracted. The R package “WGCNA” [11] was used for calculation of the appropriate soft-threshold (power) for co-expression network construction to identify key gene modules and hub genes in each key module. After transforming the adjacency matrix into a topological overlap matrix (TOM), the hierarchical clustering function was utilized to divide genes with similar expression profiles into different gene modules according to the TOM dissimilarity with a minimum size of 30 for the cluster dendrogram. For hub genes from the module that were significantly associated with clinical features of iCCA patients, we performed Gene Ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses to annotate their biological function using the R package “clusterprofiler” [12]. In
our study, we defined hub genes as those with membership (MM) > 0.8 and gene significance (GS) > 0.3 [13].

**Gene set enrichment analysis (GSEA) and construction of the protein-protein interaction network**

To find the significant enrichment gene set in iCCA, we used the software “GSEA4.0.2” [14] to analyze the expression matrix of the E-MTAB-6389 dataset. In GSEA, we selected “h.all.v7.0.symbol.gmt gene sets” as the gene set database and “HTA-2_0” as the chip platform. The STRING database (version 11.0) and Cytoscape software (version3.7.2) [15] were used to construct the protein-protein interaction network with a minimum interaction score of 0.9.

**Survival analysis and risk score model construction**

Survival analysis was performed for hub genes using the R package “survival” (https://cran.r-project.org/web/packages/survival/index.html) and “survminer” (https://cran.r-project.org/web/packages/survminer/index.html). The R package “pROC” [16] was used to plot receiver operating characteristic (ROC) curves and calculate the area under the ROC curve (AUC), which was used to evaluate the prognostic capability of hub genes. Kaplan–Meier survival curves were generated to assess the overall survival (OS) status of iCCA patients. In risk score model construction, we first used univariate Cox regression analysis to identify hub genes that significantly (P < 0.01) affect the prognosis of iCCA patients. Multivariate Cox regression analysis was then performed to determine the independent prognostic factors among those hub genes and other variables such as age, sex and cirrhosis of the liver. Next, the regression coefficients and expression values of the hub genes found to significantly affect the prognosis of iCCA were used to construct the risk score model, according to the formula $\sum \text{coef}(\text{gene}_i) \times \exp(\text{gene}_i)$. After iCCA patients were divided into high-risk and low-risk groups based on the optimal cut-off value for the risk score, we finally investigated the prognostic value of the risk score model in both the training cohort and two validation cohorts.

**Methylation and gene expression analyses**

We used the online platform MEXPRESS (http://mexpress.be) [17] to investigate the association of the abnormally expressed hub genes with the methylation status of their promoter region.
Analysis of the correlation between the expression of hub genes and tumor-infiltrating immune cells

To analyze the association between the expression of hub genes and tumor-associated infiltrating immune cells (CD4$^+$ T cells, CD8$^+$ T cells, macrophages, dendritic cells, and B cells), we utilized the website tool TIMER (https://cistrome.shinyapps.io/timer/) [18] to explore the association of hub gene expression with the degree of infiltration by tumor-infiltrating immune cells, as well as with tumor proportion.

Validation the expression status of hub genes in protein level

The Human Protein Atlas (https://www.proteinatlas.org/) [19] of gene-encoded protein levels in tissues was used to validate the expression status of our identified hub genes.

Results

Identification of robust and stable DEGs

After screening and checking the detailed sample information and the matrix design of these three downloaded datasets, we included a total of 177 iCCA patients (32 from TCGA-CHOL, 77 from E-MTAB-6389, and 68 from GSE26566) with gene expression data in our study. Figure 1 shows the detailed workflow for identification, validation, and functional analyses of the intersecting DEGs. As shown in Figure 2A and C, we used R package “TCGA-biolinks” to process the raw matrix data of TCGA-CHOL for background correction and normalization. Similarly, we used the R package “oligo” to process the raw matrix data of E-MTAB-6389. As shown in Figure 2B and D, we identified one sample (JFE-058G-HTA) as an outlier in the process of sample by sample correlation after normalization; this outlier was excluded from E-MTAB-6398 in the subsequent analyses. A total of 2,696 intersecting, stable DEGs were identified between the iCCA samples and matched normal tissue samples (Figure 2E). As shown in Figure 2F, 1,470 DEGs were upregulated and 1,226 were downregulated in iCCA patients.

WGCNA and identification of hub genes

After extraction of the intersecting DEGs with expression values from the E-MTAB-6389 dataset, we conducted WGCNA to identify key genes modules significantly associated with clinical traits of iCCA patients. As shown in Figure 3A-B, in the analysis of network topology for selection of the optimal
soft-threshold (power), a soft-threshold value of 3 (the lowest power for scale-free topology that fits an index of 0.9) was identified to construct a hierarchical clustering dendrogram. As shown in Figure 3C, no outliers sample was found in the sample clustering analysis. By setting the dendrogram height cut-off value to 0.25 for module merging, eight modules were generated (Figure 3D–E). From the heatmap of module-trait relationships (Figure 3F), we found that the black module was significantly correlated with alcohol drinking (correlation coefficient = 0.37, \( P = 9e-04 \)) and cirrhosis of the liver (correlation coefficient = 0.32, \( P = 0.004 \)). The brown (correlation coefficient = 0.25, \( P = 0.03 \)) and pink (correlation coefficient = -0.28, \( P = 0.01 \)) modules were significantly correlated with the OS of iCCA patients. **Supplementary Figure 1A**, which depicts the distribution of average GS and errors across modules associated with the Gleason score of iCCA, shows that the black module had the highest average GS among these modules. As shown in **Supplementary Figure 1B–D**, there were 1,307 DEGs in the black module, 232 DEGs in the brown module, and 192 DEGs in the pink module. The turquoise module contained genes that could not be placed in any module; these data were excluded from the subsequent analyses. According to the MM and GS values of each gene, we identified a total of 85 hub genes from the black and pink modules for further functional analyses.

**GSEA and construction of the protein-protein interaction network**
To find significant enrichment gene sets/functional pathways involved in the pathogenesis of iCCA, we conducted GSEA on the E-MTAB-6389 dataset. As shown in Figure 3G, the “hallmark-mitotic-spindle” functional pathway (NES = 1.55, NOM p-val = 0.002) was significantly upregulated in iCCA compared with normal tissue; the corresponding heatmap of this functional pathway is shown in Figure 3H. By using the STRING database (version 11.0) and Cytoscape software, we constructed protein-protein interaction network for the selected hub genes (Figure 4A).

**Functional enrichment analysis**
The GO enrichment analysis of cellular components showed that the top five components were collagen-containing extracellular matrix (ECM), collagen trimer, microbody lumen, microbody part and peroxisomal matrix (**Supplementary Figure 2A**). In terms of biological process associated with those hub genes, alpha-amino acid metabolic process was the most significantly enriched functional
process (Supplementary Figure 2B). In the KEGG pathway analysis, these genes were predominantly associated with glycine, serine and threonine metabolism, retinol metabolism, cysteine and methionine metabolism, beta-alanine metabolism and primary bile acid biosynthesis (Supplementary Figure 2C). Based on the results of PPI and functional analyses, as well as a literature review, we finally selected FTCD, BTD, FER, FETUB, F13B, COL12A1, COL1A2, COL1A1 and COL5A2 for survival analyses. We constructed a heatmap visualizing the expression status of these nine hub genes in 31 normal tissue and 77 iCCA tissue samples from the E-MTAB-6389 dataset. We also visualized the clinical traits (alcohol drinking, OS, cirrhosis of the liver and sample types) in these 108 samples. As shown in Figure 4B, BTD, FTCD, FETUB and F13B were downregulated in iCCA samples, while FER, COL1A1, COL1A2, COL12A1, COL5A2 were upregulated.

Establishment of a three-gene signature for predicting the prognosis of iCCA in the training cohort

The characteristics of iCCA patients from the TCGA-CHOL and E-MTAB-6389 datasets are summarized in Table 1. We randomly divided the E-MTAB-6389 dataset into the training cohort (n = 38) and validation cohort 1 (n = 39). Univariate Cox proportional hazard regression analysis was performed to identify the hub genes significantly associated with the OS of iCCA patients from the training cohort. As a result, a total of six hub genes (SULF1, LAMA4, COL12A1, FER, BTD and CTH) were found to be significantly associated with the OS of iCCA patients (P < 0.05). Subsequently, multivariate Cox regression was performed for these six hub genes as well as the clinical characteristics of iCCA patients. As shown in Supplementary Figure 3A–C, BTD, FER and COL12A1, as well as the cirrhosis of liver were identified as independent prognostic factors for iCCA patients. As shown in Figure 5A–C, BTD was identified as a favorable prognostic factor for iCCA patients, while COL12A1 and FER were unfavorable prognostic factors. Based on the results of the multivariate Cox regression analysis, we constructed a risk model for predicting the OS of iCCA patients. The risk score (RS) formula = BTD × 1.927 + COL12A1 ×(-2.429) + FER ×(-1.313). As shown in Supplementary Figure 3D, the RS was identified as an independent prognostic factor for iCCA patients. In addition, as shown in Figure 5D, iCCA patients with high RS had an unfavorable outcome compared with that of patients with low RS (P
Validation of the independent prognostic value of the three-gene signature in iCCA patients from two validation cohorts

The prognostic capability of this three-gene signature (RS model) was validated in both the internal (validation cohort 1 from E-MTAB-6389) and external (validation cohort 2 from TCGA-CHOL) validation cohorts. Based on the expression value of BTD, COL12A1 and FER, the RS was calculated according to the formula described previously. In validation cohort 1, we found that iCCA patients with high RS had a poor outcome \( (P = 0.006) \) compared with that of patients with low RS (Figure 6A). As shown in Figure 6B, we also found that iCCA patients with high RS in validation cohort 2 had an unfavorable outcome compared with that of patients with low RS \( (P = 0.019) \). In addition, the results of multivariate Cox regression applied in both validation cohort 1 (Figure 6C; \( P = 0.038 \)) and validation cohort 2 (Figure 6D; \( P = 0.015 \)) indicated that this three-gene signature (RS model) was an independent risk factor for OS of iCCA patients.

Evaluation of the predictive capability of the three-gene signature using ROC analysis

We constructed ROC curves to assess the performance of the three-gene signature in predicting survival of iCCA patients in validation cohort 1. As shown in Figure 6E, the AUC of the three-gene signature was 70.8% (53.7%-80.9%), indicating that the three-gene signature had higher prognostic accuracy compared with that of FER (60.2% [41.2%-79.3%]), BTD (64.6% [44.7%-84.5%]), and COL12A1 (60.2% [40.5%-80.0%]) alone. Furthermore, as shown in Figure 6F, compared with the AUC values for biomarkers such as MUC1 (48.1% [28.5%-67.7%]) and MUC13 (68.0% [50.75%-85.5%]), which have been identified as risk prognostic factors of iCCA patients [20, 21], this three-gene signature (70.8% [53.7%-80.9%]) had higher prognostic accuracy.

Validation of the expression status of the hub genes at the translational level

As shown in Supplementary Figure 4A-F, the immunohistochemical (IHC) staining analysis of iCCA and normal tissue samples obtained from The Human Protein Atlas database (https://www.proteinatlas.org/) showed that FER and COL12A1 were positively detected in iCCA tissue, whereas BTD was not. None of these three hub genes were detected in the bile duct cells in normal
liver tissue. The expression status of these three hub genes at the protein level was consistent with that at the transcriptional level. These results demonstrate the validity of the hub genes identified in this study.

The relationship between the expression level of the hub genes and their methylation status

To clarify the potential mechanism of these three hub genes that were found to be abnormally expressed in iCCA tissue (FER and COL12A1 were upregulated, while BTD was downregulated in iCCA), we screened the MEXPRESS (http://mexpress.be) database to investigate the association of the expression level of these three hub genes with the methylation status of the corresponding promoter region. As shown in Supplementary Figures 5–7, the DNA sequences of FER, BTD and COL12A1 contained numerous methylation sites, which were negatively correlated with their expression level.

The association of the expression level of the hub gene with the degree of infiltration by tumor-infiltrating immune cells and tumor purity

We used TIMER (https://cistrome.shinyapps.io/timer/) to explore the potential associations of the expression level of these three hub genes with tumor purity, as well as the degree of infiltration by tumor-infiltrating immune cells (B cells, CD8+ T cells, CD4+ T cells, macrophages, neutrophils and dendritic cells). Unfortunately, no significant associations were found between these three hub genes and tumor purity (Supplementary Figures 8A–C). Conversely, moderate or strong positive correlations were found between FER and the degree of infiltration by CD4+ T cells, macrophages, and neutrophils (Supplementary Figures 8A). Weak or moderate associations were found between COL12A1 and the degree of infiltration by B cells, macrophages, neutrophils and dendritic cells (Supplementary Figures 8B). However, no significant associations were observed between the infiltration of tumor-infiltrating immune cells and BTD expression level (Supplementary Figures 8C). These results indicated that BTD and FER regulate the degree of infiltration by tumor-infiltrating immune cells in iCCA.

Discussion

New systemic therapies, including immune-based therapies and biomarker-driven therapies, have
been applied to treat iCCA [4]; however, the prognosis of this intractable disease remains dismal. Identification of novel biomarkers for iCCA is necessary for the discovery of potential therapeutic targets and elucidation of the molecular mechanisms of tumorigenesis in iCCA. In our present study, we found a total of 2,696 abnormally expressed DEGs through bioinformatic analyses of iCCA and normal tissues. Of these, 1,470 DEGs were upregulated and 1,226 DEGs were downregulated in iCCA tissue, indicating the existence of complex molecular mechanisms in the pathogenesis of iCCA. Previous studies using high-throughput sequencing of both iCCA and other biliary tract cancer have provided valuable information on molecular classification, molecular alteration and oncogenic pathways associated with iCCA [5-7, 22]. Nevertheless, powerful biomarkers for predicting prognosis and treatment response for iCCA patients are still lacking. In the present study, we performed WGCNA to identify key gene modules significantly associated with the clinical traits of iCCA patients. We then identified a total of 85 hub genes from the black and pink gene models based on the GS and MM values. The GO enrichment analysis of cellular components showed that these hub genes were mainly involved in collagen-containing ECM, and the abnormalities in ECM can result in desmoplastic stroma which is marked characteristic of iCCA [23].

Univariate Cox regression analyses showed that SULF1, LAMA4, COL12A1, FER, BTD, CTH and cirrhosis of the liver were significantly associated with the OS of iCCA patients. In contrast, in the multivariate Cox regression analysis, BTD, FER and COL12A1, as well as cirrhosis of liver, were identified as independent prognostic factors in iCCA patients. Given that BTD was found to be a favorable prognostic factor in iCCA patients, whereas COL12A1 and FER were unfavorable prognostic factors, a three-gene signature (BTD-FER-COL12A1) was established to predict the survival of iCCA patients. We further validated the independent prognostic value of this three-gene signature in two validation cohorts (Figure 6C–D). In addition, we found that the prognostic accuracy of this three-gene signature was superior to that of BTD, COL12A1 and FER alone (Figure 6E). In comparison with other known prognostic biomarkers such as MUC1 [20] and MCU13 [21], we found that this three-gene signature provides distinctly higher accuracy in terms of prognosis prediction for iCCA patients. These results indicate that the application of multiple biomarkers is a power tool for predicting the
prognosis of iCCA patients.

Searches of The Human Protein Atlas to validate the expression status of FER, COL12A1 and BTD at the translational level showed that FER and COL12A1 were positively expressed in iCCA tissue, whereas BTD was not. While none of these three hub genes were detected in bile duct cells of normal liver tissue, the transcriptional expression status of these genes was consistent with that at the translational level. Thus, these findings indicate that these three abnormally expressed hub genes in iCCA are reliable biomarkers involved in the pathogenesis and tumorigenesis of iCCA. In addition, using MEXPRESS, we found that many methylated sites in the promoter regions of the hub genes correlated negatively with expression levels. This finding indicates that regulation of the expression of these hub genes by demethylation intervention as a therapeutic strategy for iCCA patients.

BTD encodes the biotinidase protein, which recycles protein-bound biotin by cleaving biotin (biotin-epsilon-lysine) that is a normal product of carboxylase degradation, resulting in regeneration of free biotin [24]. BTD-deficiency is associated with various diseases [25]. In our current study, we found that BTD was downregulated in iCCA, and was positively associated with the OS of iCCA patients. FER, a cancer-related gene, encodes FER tyrosine kinase, which is a member of the FPS/FES family of non-transmembrane receptor tyrosine kinases. FER regulates cell-cell adhesion, promotes cell migration, and mediates signaling from the cell surface to the cytoskeleton via growth factor receptors [26]. A recent study also showed that targeting FER kinase can inhibit melanoma growth and metastasis [27]. In our present study, we discovered that FER was upregulated at both transcriptional and translational levels. In addition, we found significant correlations between the expression of FER and the degree of infiltration by CD4⁺ T cells, macrophages, and neutrophils (Supplementary Figure 8A), thus implicating FER as a new therapeutic target for iCCA treatment.

COL12A1 (collagen type XII alpha 1 chain), also known as COL12A1L, encodes the alpha chain of type XII collagen that is a member of the FACIT (fibril-associated collagens with interrupted triple helices) collagen family. Abnormal COL12A1 expression can cause myopathy and is correlated with poor survival in gastric cancer patients [28, 29]. Similar to the expression status of FER in iCCA, we found that COL12A1 was upregulated in iCCA tissue compared with normal tissue. Further studies are
required to clarify the exact mechanism of the increased expression of COL12A1 and FER in iCCA. Some limitations of our present study should be noted. First, the small sample size in both the training cohort and the two validation cohorts may influence the stability of our results. Second, some clinical characteristics, such as the surgical resection margin status and lymph node status, were not included in our multivariate analysis model since the original data were unavailable. Finally, all our work is based on bioinformatic analyses and data mining from online databases, and further experimental studies are required to further confirm our findings.

Conclusion
We identified a three-gene signature that can be considered as a novel independent prognostic biomarker in terms of predicting the OS of iCCA patients. This information may aid in the discovery of novel therapeutic targets for precision treatment of iCCA patients.

List Of Abbreviations
iCCA: intrahepatic cholangiocarcinoma iCCA;
CCA: cholangiocarcinoma; OS: overall survival; DEGs: differential expressed genes;
TCGA: The Cancer Genome Atlas; GEO: Gene Expression Omnibus;
WGCNA: Weighted gene co-expression network analysis;
GO: gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes;
GS: gene significance; GSEA: gene set enrichment analysis;
ROC: receiver operating characteristic curve; AUC: area under the ROC curve;
IHC: immunohistochemical;

Declarations

**Ethics approval and consent to participate**
This study did not require ethical approval and consent to participate as the data used have been published previously or have been uploaded on online databases, and hence are already in the public domain.

**Consent for publication**
This study did not contain any individual person’s data in any form.

**Data availability statement**
The datasets analyzed in this study can be found in ArrayExpress (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6389/), Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26566), and The Cancer Genome Atlas (https://portal.gdc.cancer.gov/) (Level 3, TCGA-CHOL).

**Conflicts of interest**

The authors declare no conflicts of interest.

**Funding**

This study was funded by the grants from the National Key Research and Development Program (2019YFC1316000), National Natural Science Foundation of China (31970696 and 81502975), and Zhejiang Provincial Program for the Cultivation of High-level Innovative Health Talents. The funding sources played no roles in the study design, experiment method development, data collection, organization, analysis and interpretation, nor in the manuscript drafting, reviewing or approval.

**Author contributions**

Zengwei Tang: study concept and design, performing statistical analysis and bioinformatics analysis, interpretation of data, and drafting the manuscript; Enliang Li: statistical analysis and interpretation of data; Xing Huang & Qi Zhang: interpretation of data and critical revision of the manuscript for significant intellectual content; Xueli Bai & Tingbo Liang: study supervision and funding application; critical revision of the manuscript for significant intellectual content.

**References**

1. Rizvi S, Khan SA, Hallemeier CL, Kelley RK, Gores GJ. **Cholangiocarcinoma - evolving concepts and therapeutic strategies.** *Nat Rev Clin Oncol* 2018, **15**(2):95-111.

2. Nault JC, Villanueva A. **Biomarkers for hepatobiliary cancers.** *Hepatology* 2020, doi:10.1002/hep.31175 (In press).

3. Banales JM, Cardinale V, Carpino G, Marzioni M, Andersen JB, Invernizzi P, *et al.* **Expert consensus document: Cholangiocarcinoma: current knowledge and future perspectives consensus statement from the European Network for the**
Study of Cholangiocarcinoma (ENS-CCA). Nat Rev Gastroenterol Hepatol 2016, 13(5):261-280.

4. Kelley RK, Bridgewater J, Gores GJ, Zhu AX. Systemic therapies for intrahepatic cholangiocarcinoma. Journal of hepatology 2020, 72(2):353-363.

5. Jiao Y, Pawlik TM, Anders RA, Selaru FM, Streppel MM, Lucas DJ, et al. Exome sequencing identifies frequent inactivating mutations in BAP1, ARID1A and PBRM1 in intrahepatic cholangiocarcinomas. Nature genetics 2013, 45(12):1470-1473.

6. Sia D, Hoshida Y, Villanueva A, Roayaie S, Ferrer J, Tabak B, et al. Integrative molecular analysis of intrahepatic cholangiocarcinoma reveals 2 classes that have different outcomes. Gastroenterology 2013, 144(4):829-840.

7. Farshidfar F, Zheng S, Gingras MC, Newton Y, Shih J, Robertson AG, et al. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. Cell Rep 2017, 18(11):2780-2794.

8. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic acids research 2016, 44(8):e71.

9. Carvalho BS. Working with Oligonucleotide Arrays. Methods in molecular biology (Clifton, NJ) 2016, 1418:145-159.

10. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research 2015, 43(7):e47.

11. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics 2008, 9:559.

12. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing
biological themes among gene clusters. *Omics: a journal of integrative biology* 2012, 16(5):284-287.

13. Song ZY, Chao F, Zhuo Z, Ma Z, Li W, Chen G. **Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis.** *Aging* 2019, 11(13):4736-4756.

14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, 102(43):15545-15550.

15. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome research* 2003, 13(11):2498-2504.

16. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. **pROC: an open-source package for R and S+ to analyze and compare ROC curves.** *BMC bioinformatics* 2011, 12:77.

17. Koch A, Jeschke J, Van Criekinge W, van Engeland M, De Meyer T. **MEXPRESS update 2019.** *Nucleic acids research* 2019, 47(W1):W561-W565.

18. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, et al. **TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells.** *Cancer research* 2017, 77(21):e108-e110.

19. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. **Proteomics. Tissue-based map of the human proteome.** *Science (New York, NY)* 2015, 347(6220):1260419.

20. Tang Z, Yang Y, Wang X, Meng W, Li X. **Meta-analysis of the diagnostic value of Wisteria floribunda agglutinin-sialylated mucin1 and the prognostic role of**
mucin1 in human cholangiocarcinoma. *BMJ open* 2019, 9(1):e021693.

21. Tiemin P, Fanzheng M, Peng X, Jihua H, Ruipeng S, Yaliang L, et al. MUC13 promotes intrahepatic cholangiocarcinoma progression via EGFR/PI3K/AKT pathways. *Journal of hepatology* 2020, 72(4):761-773.

22. Andersen JB, Spee B, Blechacz BR, Avital I, Komuta M, Barbour A, et al. Genomic and genetic characterization of cholangiocarcinoma identifies therapeutic targets for tyrosine kinase inhibitors. *Gastroenterology* 2012, 142(4):1021-1031 e1015.

23. Sirica AE, Gores GJ. Desmoplastic stroma and cholangiocarcinoma: clinical implications and therapeutic targeting. *Hepatology* 2014, 59(6):2397-2402.

24. Borsatto T, Sperb-Ludwig F, Blom HJ, Schwartz IVD. Effect of BTD gene variants on in vitro biotinidase activity. *Molecular genetics and metabolism* 2019, 127(4):361-367.

25. Liu Z, Zhao X, Sheng H, Cai Y, Yin X, Chen X, et al. Clinical features, BTD gene mutations, and their functional studies of eight symptomatic patients with biotinidase deficiency from Southern China. *American journal of medical genetics Part A* 2018, 176(3):589-596.

26. Yamamoto H, Kondo A, Itoh T. A curvature-dependent membrane binding by tyrosine kinase Fer involves an intrinsically disordered region. *Biochemical and biophysical research communications* 2018, 495(1):1522-1527.

27. Ivanova IA, Arulanantham S, Barr K, Cepeda M, Parkins KM, Hamilton AM, et al. Targeting FER Kinase Inhibits Melanoma Growth and Metastasis. *Cancers* 2019, 11(3).

28. Hicks D, Farsani GT, Laval S, Collins J, Sarkozy A, Martoni E, et al. Mutations in the collagen XII gene define a new form of extracellular matrix-related
myopathy. Human molecular genetics 2014, 23(9):2353-2363.

29. Jiang X, Wu M, Xu X, Zhang L, Huang Y, Xu Z, et al. COL12A1, a novel potential prognostic factor and therapeutic target in gastric cancer. Molecular medicine reports 2019, 20(4):3103-3112.

Table
Table 1. Clinical Characteristics of iCCA patients included in our study

| The training cohort (n=38) |
|--------------------------|
| Gender (Male/Female)     | 19/19     |
| Vascular invasion (Yes/No)| 20/18     |
| Alcohol drinking (Yes/No)  | 4/29      |
| Cirrhosis of liver (Yes/No) | 7/30      |
| Necrosis detected (Yes/No) | 15/22     |

| The validation cohort 1 (n=39) |
|-------------------------------|
| Gender (Male/Female)          | 18/21     |
| Vascular invasion (Yes/No)    | 30/9      |
| Alcohol drinking (Yes/No)     | 5/30      |
| Cirrhosis of liver            | 5/32      |
| Necrosis detected (Yes/No)    | 14/24     |

| The validation cohort 2 (n=32) |
|-------------------------------|
| Mean age at diagnosis (yrs)   | 63.2(29.2-82.3) |
| Gender (Male/Female)          | 18/14      |
| N0/N1/NX                      | 25/4/3     |
| M0/M1/MX                      | 27/3/2     |
| Tumor stage (i/ii/iii/iv)     | 18/8/1/5   |

Supplementary Figure Legends
Supplementary Figure 1. Bar plot of the average gene significance and errors in the modules associated with Gleason score of iCCA (A). Scatter plot of module eigengenes in the brown module (B). Scatter plot of module eigengenes in the black module (C). Scatter plot of module eigengenes in the pink module (D).
Supplementary Figure 2. Gene ontology (GO) enrichment showing the top five cellular components of hub genes (A). GO enrichment showing the top five biological processes of hub genes (B). Dot plot showing the top five pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis (C).

Supplementary Figure 3. Forest plot showing the multivariate analysis results from the training cohort. Forest plot showing that BTD and the cirrhosis of liver were independent prognostic factors for iCCA (A). Forest plot showing that FER and the cirrhosis of liver were independent prognostic factors for iCCA (B). Forest plot showing that COL12A1 and cirrhosis of liver were independent prognostic factors for iCCA (C). Forest plot showing that the three-gene signature was an independent prognostic factor for iCCA (D).

Supplementary Figure 4. Validation of the expression of FER, BTD and COL12A1 at the translational level using the resources of The Human Protein Atlas. FER was positively expressed in iCCA (A), but not in normal bile duct cells (B). BTD was not detected iCCA samples (C) or normal bile duct cells (D). COL12A1 was positively expressed in iCCA (E), but not in normal bile duct cells (F).

Supplementary Figure 5. The correlation between FER expression and methylation status of its DNA sequences.

Supplementary Figure 6. The correlation between BTD expression and methylation status of its DNA sequences.

Supplementary Figure 7. The correlation between COL12A1 expression and methylation status of its DNA sequences.

Supplementary Figure 8. The correlation between the FER expression level and the level of tumor-infiltrating immune cells (A). The correlation between the BTD expression level and the level of tumor-infiltrating immune cells (B). The correlation between the COL12A1 expression level and the level of tumor-infiltrating immune cells (C).

Figures
Figure 1

Diagram showing the detailed workflow for identification, validation, and functional analyses of the intersecting differentially expressed genes. Notes: TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; WGCNA, weighted gene co-expression network analysis; GSEA, gene set enrichment analysis; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.
Figure 2

Heatmap showing the Array-Array intensity after using the “RMA” method for background correction in TCGA-CHOL dataset (A). Boxplot of sample by sample correlation after normalization in the E-MTAB-6389 dataset (B). Boxplot of sample by sample correlation after normalization in TCGA-CHOL dataset (C). Boxplot of sample by sample correlation after normalization after excluding one outlier from the entire dataset of the E-MTAB-6389 dataset (D). Venn diagram showing the intersecting differentially expressed genes across the GSE26566, TCGA-CHOL, and E_MTAB_6398 datasets (E). Volcano plot showing the downregulated and upregulated differentially expressed genes in iCCA (F).
Figure 3

Analysis of the scale-free fit index (A) and the mean connectivity for various soft-thresholding powers (B). Clustering of samples for detecting outliers in the E-MTAB-6389 dataset (C). Cluster dendrogram of the intersecting differentially expressed genes in the E-MTAB-6389 dataset. Each branch in the plot represents one gene, and every color below represents one co-expression module (D). Network heatmap showing the interaction relationship of co-expressed genes in the E-MTAB-6389 dataset. Different colors in the vertical and horizontal axes represent different modules. The intensity of the yellow color in the middle represents the degree of connectivity of different modules. No significant difference was found among different modules, which indicates a high-scale independence degree among these modules (E). Heatmap showing relationships between module eigengenes and the clinical traits of iCCA patients (F). Enrichment plot showing the functional pathway of “hallmark_mitotic_spindle” significantly upregulated in iCCA (G). Heatmap of genes from the functional pathway of “hallmark_mitotic_spindle”, showing that these genes were significantly upregulated in iCCA compared with normal tissue (H).
Protein-protein interaction network of the identified hub genes. Node size represents the connectivity of the gene by degree and edge size represents the betweenness centrality of genes. Nodes in green represent downregulated genes, and nodes in red represent upregulated genes (A). Heatmap visualizing the overall expression of BTD, FTCD, FETUB, F13B, FER, COL1A1, COL1A2, COL12A1 and COL5A2 in iCCA and normal tissues, as well as the distribution of the clinical characteristics of the sample sources (B).
The Kaplan–Meier curves for overall survival of iCCA patients in the training cohort. BTD was significantly associated with favorable overall survival of iCCA patients (A), while COL12A1 (B), FER (C), and risk score (D) were significantly associated with the poor survival of iCCA patients.
The Kaplan–Meier curves for overall survival of high- and low-risk groups of patients in (internal) validation cohort 1 (A) and validation cohort 2 (B). Forest plot showing the multivariate analysis results from the validation cohort 1 (C) and validation cohort 2 (D). Receiver operating characteristic (ROC) curves showing that the performance of this three-gene signature in predicting the prognosis of iCCA patients in validation cohort 1 was superior to that of FER, BDT, and COL12A1 alone (E). ROC curves showing that the performance of this three-gene signature was superior to that of known biomarkers, such as MUC1 and MUC13 (F).

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

Supplementary Figure 5.png
Supplementary Figure 6.png
Supplementary Figure 7.png
Supplementary Figure 3.tif
