Establishment of a Novel Nine-gene Signature for the Prognosis of Cholangiocarcinoma

Qiang Liu
Hangzhou First People's Hospital
Liyun Zheng
Hangzhou First People's Hospital
Dongchao Xu
Hangzhou First People's Hospital
Hangbin Jin
Hangzhou First People's Hospital
Sile Cheng
Hangzhou First People's Hospital
Hongzhang Shen
Hangzhou First People's Hospital
Ye Gu
Hangzhou First People's Hospital
Shenghui Chen
Hangzhou First People's Hospital
Jianpeng Zhu
Zhejiang Chinese Medical University
Xiaofeng Zhang
Hangzhou First People's Hospital
Jianfeng Yang (yjf3303@zju.edu.cn)
Hangzhou First People's Hospital
Ying Bian
Zhejiang Chinese Medical University

Research Article

Keywords: Cholangiocarcinoma, prognostic biomarker, microenvironment, infiltration, comprehensive analysis

Posted Date: May 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1633412/v1

Establishment of a Novel Nine-gene Signature for the Prognosis of Cholangiocarcinoma

Qiang Liu[1)], Liyun Zheng[1)], Dongchao Xu[1)], Hangbin Jin[1)], Sile Cheng[1)], Hongzhang Shen[1)], Ye Gu[1)], Shenghui Chen[1)], Jianpeng Zhu[2)], Ying Bian[2)], Xiaofeng Zhang[1)] & Jianfeng Yang[1)]

1) Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China.

2) The Fourth School of Clinical Medicine, Zhejiang Chinese Medical University, Hangzhou, China.

* These authors contributed equally: Qiang Liu and Liyun Zheng

Authors:

1) Qiang Liu, Ph.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: 11718046@zju.edu.cn

2) Liyun Zheng, M.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: 22160874@zju.edu.cn

3) Dongchao Xu, M.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: xudongchaoxdc@sina.com

5）Hnagbin Jin, Ph.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: kenjhb@163.com

5） Sile Cheng, M.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: slcheng@njmu.edu.cn

6) Hongzhang Shen, Ph.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: shz@zcmu.edu.cn

7) Ye Gu, Ph.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: chinesegu@foxmail.com
8) Shenghui Chen, Ph.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: chensheng0805@163.com
9) Jianpeng Zhu, B.S. The Fourth School of Clinical Medicine, Zhejiang Chinese Medical University, Hangzhou, China. E-mail: 806001555@qq.com
10) Ying Bian, B.S. The Fourth School of Clinical Medicine, Zhejiang Chinese Medical University, Hangzhou, China. E-mail: bianying1202@163.com

Corresponding authors:

1） Jianfeng Yang, Ph.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: yjf3303@zju.edu.cn
2） Xiaofeng Zhang, Ph.D. Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. E-mail: 837837@zju.edu.cn

Abstract

Background: Cholangiocarcinoma (CCA) is a rare malignant carcinoma characterized by high mortality, challenging diagnosis, and poor prognosis. A powerful prediction biomarker is urgently needed for the early diagnosis and individualized treatment of CCA patients.

Methods: A systematic bioinformatics analysis was conducted based on mRNA expression data and clinical information from The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO) and National Genomics Data Center (NGDC) datasets. Differentially expressed genes (DEGs) between tumor tissues and adjacent counterpart controls were identified in the TCGA and GSE107943 datasets. A nine-gene prediction model was constructed, and its effects on CCA prognosis were analyzed using univariate,

multivariate and LASSO Cox proportional hazards regression models, Kaplan-Meier plotter, CIBERSORT and OncoPredict in the discovery and validation cohorts. Additionally, the expression profiles of the target genes were determined via qRT-PCR and DEG analyses in an independent cohort.

Results: A nine-gene signature (*HELLS*, *HOXC6*, *MFSD2A*, *OTX1*, *PTGES*, *PYGB*, *SMOC1*, *TEX30* and *ZBTB12*) displayed excellent predictive performance for the overall survival of CCA. According to the prognostic signature, CCA patients were classified into high-risk and low-risk groups, with obvious differences in overall survival probabilities. The low-risk group had a significantly better prognosis than the high-risk group in both the discovery cohort (n = 66) and replication cohort (n = 255) ($P < 0.0001$, $P < 0.0001$). Additionally, five cancer drugs (Erlotinib, ML323, AGI-6780, Gallibiscoquinazole and AZD3759) presented clearly specific sensitivities for high-risk and low-risk group patients. Moreover, according to tumor microenvironment analyses, high-risk group patients had a higher level of M0 macrophage infiltration than low-risk group patients ($P = 0.025$, $P = 0.0048$). In replicating the expression patterns of the nine genes, eight of the nine genes (except *TEX30*) were found to have significant expression profiles between 15 tumor tissues and adjacent counterpart controls; moreover, the qRT-PCR results validated the abnormal expression pattern of the target genes in CCA.

Conclusions: Collectively, we established an effective prognostic model for different populations of CCA patients based on nine DEGs. These findings may provide potential benefits for the development of new prognostic biomarkers and therapeutic targets for CCA.

Key words: Cholangiocarcinoma, prognostic biomarker, microenvironment, infiltration, comprehensive analysis

Background

CCA is a rare and lethal disease accounting for nearly 3% of digestive neoplasias and 15% of primary liver cancers [1-3]. However, the global incidence and mortality rates of this fatal disease have increased in recent decades [4-6]. Due to the absence of obvious

clinical symptoms or characteristics at the early stages, most CCA patients are diagnosed at advanced stages with limited therapeutic options, resulting in a disappointing prognosis with a median overall survival (OS) duration of less than 18 months [7-10]. Despite advances in curative therapeutics for CCA, surgery is not suitable for patients at advanced stages, and the operative resection rate is no more than 30%, with a high recurrence rate after resection [3, 9].

Imaging techniques, including CT scans and MRI, have been applied for the diagnosis of CCA. However, their accuracy and sensitivity still need to be improved and require additional histological confirmation. A traditional serum biomarker, carbohydrate antigen 19-9 (CA19-9), is used to aid in the diagnosis of intrahepatic and extrahepatic tumors. Confounded by many factors, such as being negative for Lewis antigen and/or positive for bacterial cholangitis, the sensitivity, specificity, and adjusted positive predictive value of CA19-9 are dramatically dampened in the diagnosis of CCA [11-13]. Hence, more efficient biomarkers are urgently needed to improve the clinical outcomes of CCA patients.

Although individualized medicine and precise therapies have been improved for various tumors in decades, the characteristics of the biological mechanisms in CCA remain largely unknown. In addition, CCA is highly heterogeneous at the intertumoral and intratumoral levels. The development of new diagnostic methods, therapeutic options, and personalized medicine and the individual characterization of these tumors at both the genomic and epigenomic levels via multiomics analysis to ascertain their pathogenesis and the subtypes of CCA are essential [4, 8, 14]. Of note, the incidence rate of CCA presents considerable geographical variation, greater than 6 per 100,000 habitants in East Asian countries and 0.3-6 per 100,000 habitants in others [15]. Discrepancies in the incidence rate seemingly indicate the difference in risk factors and genetic predispositions. In addition, Farshidfar and his colleagues reported four possible molecular subtypes of CCA based on specific variations in DNA methylation, gene expression, copy number alterations, and mutation profiles [16]. Benefiting from sequencing methods and bioinformatic analysis development, we aimed to characterize CCA using genetic and epigenetic features for diagnosis, prognosis prediction, and personalized treatments.

In the current study, we comprehensively analyzed the expression profiles in tumor tissues of CCA and adjacent normal specimens from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) datasets. Univariate, least absolute shrinkage and selection operator (LASSO), and multivariate Cox regression analyses were performed to screen OS-related differentially expressed genes (DEGs), which allowed us to extend the knowledge of the biological function of the CCA-specific genes in the prognostic process. The accuracy and sensitivity of the constructed prognostic model were further validated with a large cohort of Chinese CCA patients. In addition, the potential relationships between tumor immune cell infiltration and drug sensitivity and the prognostic signature were also investigated. Furthermore, independent cohort analysis and qRT-PCR were performed to verify the expression patterns of the genes involved in the prognostic model. Our findings presented the prognostic value of a novel nine-gene prediction model and provided novel insights into the potential mechanism as well as therapeutic opportunities in CCA.

Methods

Data and resources

The level three count-based gene expression profiles and associated clinical information of the TCGA-CCA cohort (n = 45), which included 9 adjacent normal tissue samples and 36 tumor tissue samples, were downloaded from the UCSC Xena browser (https://gdc.xenahubs.net) [17]. GSE107943 and GSE119336 were collected from GEO as validation datasets (https://www.ncbi.nlm.nih.gov/geo/). GSE107943 is a high-throughput sequencing dataset that contains 30 CCA tissue samples and 27 adjacent normal tissue samples. To ameliorate potential false-positive results, nonexpressed or lowly expressed genes with average counts less than 5 were excluded from downstream analyses. Considering the heterogeneity of CCA in different populations [4, 8] and the fact that the majority of patients in the GSE107943 and TCGA-CCA cohorts were European and American, we also downloaded GSE119336, an RNA-seq dataset that included paired CCA tumor tissues and adjacent tissues from 15 Chinese CCA patients, for further validation of the abnormal expression of *HELLS*, *HOXC6*, *MFSD2A*, *OTX1*, *PTGES*,

PYGB, *SMOC1*, *TEX30* and *ZBTB12*. However, OS information was not reported in the GSE119336 dataset. Hence, to replicate the accuracy and sensitivity of the novel nine-gene prognostic model in different populations, the mRNA expression profiles and associated clinical information of 255 Chinese CCA patients were acquired from the NGDC dataset (https://www.biosino.org/node/project/detail/OEP001105).

Gene coexpression network analysis

To better understand the etiology of bile duct neoplasia, we employed the R package WGCNA [18, 19] to identify coexpressed gene subnetworks (modules) using RNA-seq expression data from TCGA-CCA. Specifically, we first normalized the gene count data in reads per kilobase per million mapped reads (RPKM) format. Then, a weighted coexpression network was constructed based on the correlations among all gene pairs. Subsequently, the correlation matrix between gene pairs was converted into a scale-free adjacency matrix. Finally, the unsigned adjacency matrix was translated as a topological overlap matrix with default parameters to assess the modular architectures of module interconnections [19]. Thirty-one coregulated gene modules were generated by average linkage hierarchical clustering according to topological overlap (Fig. 1a). The module sizes of the included genes ranged from 75 to 8,642. Furthermore, we implemented analysis to directly evaluate the relationship between modules and experimental covariates, such as diagnosis, BMI, sex, race, and age.

Differential gene expression analysis

We performed DEG analyses using the DESeq2 package (v. 3.6.3) [20]. Genes with both significant multi-test corrected *P* values (Bonferroni-corrected *P* value < 0.05) and log2-transformed fold changes (absolute value of log2FC > 1) were defined as CCA-associated DEGs. Only DEGs with concordant LFC in the training and validation datasets were retained for further analyses. In addition, the aberrant expression level of the genes included in the prediction model were validated by using the qRT-PCR assay. Reverse transcription reactions were performed with the PrimeScript™ II 1st Strand cDNA Synthesis Kit (TaKaRa) following the manufacturer's protocol. qRT-PCR was performed with UltraSYBR Mixture (CWBIO). The primers are listed in Table S1. The transcriptome expression levels were assessed relative to the expression of *GAPDH* by the

$2^{-\Delta\Delta Ct}$ method and then evaluated by Welch's *t test*.

Pathway enrichment analysis

To illuminate the potential pathologies of CCA, the R package ClusterProfiler (v. 3.14.3) was applied to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis [21]. Items with false discovery rate (FDR) values less than 0.05 were considered significantly enriched pathways. Gene set enrichment analysis (GSEA) was conducted using RPKM format expression data with the R packages gerichplot, ggplot2 and ClusterProfiler (v. 3.14.3) [22].

Identification of survival-related genes and construction of a prognostic signature for CCA

Because of the limited number of patients enrolled in the TCGA-CCA (n = 33) and GSE119336 (n = 30) cohorts, we combined the patients from the two datasets into a discovery cohort to perform further analysis. Expression data from TCGA-CCA and GSE119336 were used to identify the association between the abovementioned DEGs and OS. First, a univariate Cox proportional hazards regression model was conducted to screen prognosis-related genes through the R package glmnet (v 4.1.2) [23]. Genes with a *P* value less than 0.05 were considered statistically significant and were used for further analyses. Second, a LASSO-penalized Cox regression model with 100-fold cross-validation was performed to select the model with the best predictive performance. According to the linear combination of the regression coefficients from LASSO, a risk score was calculated for each patient. The patients were divided into high-risk and low-risk groups based on the median value of all risk scores. Kaplan–Meier analysis (K-M) and time-dependent receiver operating characteristic curves (ROC) were employed to assess the sensitivity and specificity of the constructed model. A *P* value less than 0.05 in K-M analysis was considered significant. To test the prognostic risk score and assess its predictive precision and sensitivity in different populations, we performed the same analyses on 255 Chinese CCA patients from the NGDC dataset.

The K-M and ROC analyses were conducted by using the R packages timeROC (v 0.4), survival (v 3.2-11), survminer (v 0.4.9), and survivalROC (v 1.0.3) in both the training and validation datasets. Additionally, the relationships between the expression

levels of the genes in the prognostic signature, including mRNAs, and the OS rate were analyzed in the training and validation datasets. The K-M survival curves were drawn with a modified drawing function 'ggsurvplot' in survminer (v 0.4.9).

Assessment of immune cell infiltration

The CIBERSORTx algorithm was applied to evaluate the relative abundance of 22 types of tumor-infiltrating immune cells based on the gene expression in the TCGA-CCA dataset [24, 25]. The category of immune cell types included naive B cells, memory B cells, plasma cells, resting memory $CD4^+$ T cells, activated memory $CD4^+$ T cells, naive memory $CD4^+$ T cells, $CD8^+$ T cells, follicular helper T cells, regulatory T cells, gamma delta T cells, resting natural killer cells (NK), activated NK cells, monocytes, M0-M2 macrophages, resting mast cells (MCs), activated MCs, resting dendritic cells, activated dendritic cells, eosinophils, and neutrophils. Because of the limited number of samples in the separate datasets, we combined the two datasets to identify the immune-related cell proportions.

Drug sensitivity differences of the risk score classification

Drug sensitivity to chemotherapeutic drugs and novel therapeutic drugs for CCA patients were evaluated according to the Genomics of Drug Sensitivity in Cancer (GDSC) database (https://www.cancerrxgene.org/). The half-maximal inhibitory concentration (IC50) of each drug for CCA was predicted by the R package OncoPredict (v. 0.2) with the RNA-seq expression matrix [26]. The Wilcoxon test was applied to calculate drug sensitivity differences between the low-risk and high-risk groups in the training and validation cohorts. The threshold for significance was set as a P value < 0.05.

Results

Network analysis identified CCA-related transcriptional signatures

In the present study, the analysis processes were implemented according to the workflow shown in Fig. 1. To gain better insight into the molecular mechanisms involved in CCA, we constructed a coexpression network and investigated the orchestration of the transcriptome in 45 specimens from the TCGA-CCA cohort as described in the Methods. Thirty-one transcriptional modules were identified in CCA, and each module was

annotated with an arbitrary color (Fig. 2a). We found that the “turquoise”, “salmon” and “brown” modules were significantly correlated with disease status (P =2.0 × 10^{-21}, P =5.0 × 10^{-07} and P =3.0 × 10^{-06}, respectively), while there was no significant association with BMI, sex, race, or age (Fig. 2b). Because the turquoise module was the most significant module associated with CCA, we focused on genes in this module for further analyses. A total of 8,642 mRNAs were grouped in the turquoise module.

Highly coexpressed DEGs in CCA

To evaluate the effects of aberrantly expressed transcriptomes in CCA, we implemented DEG analysis in the training and validation datasets. Compared with adjacent normal tissues, DESeq2 identified 2,435 upregulated mRNAs and 2,026 downregulated mRNAs in the training cohort (Table S2) and 4,443 upregulated mRNAs and 2,745 downregulated mRNAs in the replication dataset (Table S3). Furthermore, by intersecting the DEGs in the two datasets, we found a total of 3,414 replicated DEGs, including 1,867 upregulated and 1,547 downregulated genes (Fig. 2c). We further selected the 1,934 overlapping genes between the DEGs and the genes in the turquoise module as potential CCA-associated genes. KEGG pathway enrichment was performed for those 1,934 DEGs to determine whether these highly correlated DEGs might play important roles in the tumorigenesis of CCA. We identified 52 significantly overrepresented pathways (Fig. 2d, and Table S4), including several key pathways that participated in the specific metabolic vulnerabilities during the metastatic processes of various cancers [27-31].

Construction of the prognostic prediction model

To identify prognostic biomarkers for CCA, 1,934 potential targets were further subjected to univariate Cox regression analysis. Our DEG analysis revealed that the 1,934 potential targets were aberrantly expressed with consistent profiles in the TCGA-CCA and GSE119336 datasets, so we performed further analyses with individuals from the two datasets as the training set with the goal of increasing our statistical power (n = 66). A total of 122 DEGs were found to be significantly associated with OS (Supplementary Table 5). LASSO modeling identified a nine-DEG-based signature (Fig. 3a, 3b), which included helicase (*HELLS*), homeobox C6 (*HOXC6*), major facilitator superfamily domain

containing 2A (*MFSD2A*), orthodenticle homeobox 1 (*OTX1*), prostaglandin E synthase (*PTGES*), glycogen phosphorylase B (*PYGB)*, SPARC-related modular calcium binding 1 (*SMOC1*), testis expressed 30 (*TEX30*) and zinc finger and BTB domain containing 12 (*ZBTB12*). The risk score for CCA was calculated as follows: (0.0355341 × expression level of *HELLS*) + (-0.2285914 × expression level of *HOXC6*) + (0.3774950 × expression level of *MFSD2A*) + (0.0024277 × expression level of *OTX1*) + (0.0898670 × expression level of *PTGES*) + (0.1894858 × expression level of *PYGB*) + (1.6696174 × expression level of SMOC1) + (0.7596718 × expression level of *TEX30*) + (0.1891796 × expression level of *ZBTB12*).

All CCA patients were divided into high-risk and low-risk groups for the discovery (high-risk: n = 33, low-risk: n = 33) and validation datasets (high-risk: n = 128, low-risk: n = 127) according to corresponding median values of the risk score. Remarkably, the ROC curve analysis in the discovery cohort showed that the areas under the ROC curve (AUC) values at 6, 12 and 36 months were up to 0.901, 0.886 and 0.880, respectively (Fig. 3c). In the validation dataset, the AUC values at 6, 12 and 36 months were 0.656, 0.643 and 0.654, respectively (Fig. 3d). The K-M analysis indicated that the low-risk group had a better OS rate than the high-risk group in the training dataset (Fig. 3e, $P < 0.0001$). In addition, in line with the K-M analysis in the training dataset, the high-risk group was associated with a worse OS rate in the validation dataset (Fig. 3f, $P < 0.0001$). The association between OS and the risk score was explored through multivariate Cox regression, and the hazard ratios (HRs) and 95% confidence intervals (CIs) are presented as forest plots (Fig. 3g). Because several datasets lacked information on progression-free survival (PFS), disease-free survival (DFS) and disease-specific survival (DSS), the model performance in predicting DSS, PFS, and DFS in CCA patients could not be estimated.

The risk model predicted the infiltration of immune cells in CCA

To recognize the indicative role of prognostic signals in the tumor microenvironment (TME), we implemented CIBERSORT to evaluate the infiltration of 22 immune-related cells in CCA in the discovery and validation cohorts. CD4 naive T cells were not detected based on RNA-seq in the discovery cohort, and CD4 naive T cells, gamma delta T cells, activated dendritic cells, and resting mast cells were not identified in the validation cohort

(Fig. 4a, 4b). Then, the differential distribution of the detected immune-related cells between the high-risk and low-risk groups was examined by the Wilcoxon rank-sum test. Compared with those in the low-risk group, the proportions of M0 macrophages were significantly increased in the high-risk group in both the discovery and validation datasets (P = 0.025 and P = 0.005, respectively). Neutrophils were significantly increased in the high-risk group in the discovery cohort (P = 0.004), whereas high infiltration was not observed in the validation cohort (P = 0.18). In addition, resting memory CD4 T cells, activated NK cells and M2 macrophages were significantly decreased in the high-risk group compared with the low-risk group in the validation cohort (P = 0.026, P = 0.003, and P = 0.025, respectively), and activated mast cells were increased in the low-risk group (P = 0.0003). However, these results did not meet the statistically significant threshold in either cohort.

Assessment of prognostic factors in CCA

To better understand the relevance and underlying mechanisms of the novel prognostic model in CCA, we applied univariate and multivariate Cox regression analyses to evaluate whether the nine-gene model was an independent prognostic factor of OS for cholangiocarcinoma patients in the discovery and validation datasets. As shown in Tables 1 and 2, both univariate (discovery cohort: $P = 3.40 \times 10^{-06}$; validation cohort: $P = 8.96 \times 10^{-05}$) and multivariate (discovery cohort: $P = 4.76 \times 10^{-06}$; validation cohort: $P = 1.23 \times 10^{-03}$) Cox regression analyses indicated that the risk score was an independent prognostic factor. TNM stage was an independent indicator for predicting CCA patient OS in the validation cohort (univariate analysis: $P = 1.79 \times 10^{-08}$; multivariate analysis: $P = 2.96 \times 10^{-07}$). However, TNM stage as an independent prognostic factor was not replicated in the training cohort (univariate analysis: P = 0.22; multivariate analysis: P = 0.89). These results showed that the nine-gene signature was an independent prognostic factor for different populations of CCA patients.

Furthermore, associations between the clinicopathological features and the high-risk and low-risk groups in the two datasets were assessed. As shown in Table 3, the proportions of intrahepatic metastasis, vascular invasion, CA19-9, CEA and γ-GT were dramatically higher in the high-risk group than in the low-risk group in the replication

cohort. However, without this clinicopathological feature information in the discovery dataset, we were unable to confirm the relationship between these features and the risk score in an independent dataset. In addition, for the different pathological stages of CCA, the proportion of advanced stages (III, IV) in the high-risk group was significantly higher than that in the low-risk group (chi-square test, $P = 0.018$). A statistically significant signature was not identified in the training cohort ($P = 0.234$).

To identify the drug sensitivity differences of chemotherapeutic drugs and novel inhibitors for CCA patients between the high-risk and low-risk groups, the mRNA expression matrix of the training and validation cohorts was used for estimation with OncoPredict (v. 0.2) [26]. As a result, erlotinib, ML323, AGI-6780, Gallibiscoquinazole and AZD3759 revealed more sensitivity in the high-risk group than in the low-risk group in both datasets (training cohort: $P = 0.04$, $P = 0.05$, $P = 0.02$, $P = 0.04$, and $P = 0.03$, respectively; validation cohort: $P = 0.01$, $P = 2.2 \times 10^{-04}$, $P = 9.1 \times 10^{-06}$, $P = 6.9 \times 10^{-05}$, and $P = 5.8 \times 10^{-03}$, respectively) (Fig. 5 a-g).

KEGG enrichment in the risk score phenotype

GSEA was performed to reveal the risk score-associated pathways. Five and seven pathways were significantly associated with the risk score in the discovery cohort and validation cohort, respectively, after FDR correction. In the training set, high-risk score phenotype sets were enriched in bile section, chemical carcinogenesis-receptor activation, complement and coagulation cascades, cytokine-cytokine receptor interaction, and drug metabolism-other enzymes. In the validation set, high-risk score phenotype sets were enriched in complement and coagulation cascades, cytokine-cytokine receptor interaction, IL-17 signaling pathway, mucin type O-glycan biosynthesis, NK-mediated cytotoxicity, neutrophil extracellular trap formation, and TNF signaling pathway (Fig. 5k, 5i). These results indicated that cytokine-cytokine receptor interactions and complement and coagulation cascades, two inflammatory immune response-related pathways, as common pathways, might play an essential role in the molecular pathology of CCA.

Validation of the prediction model including gene expression patterns

To replicate the prediction model that included genes' abnormal expression profiles, we performed DEG analysis on an independent cohort with 15 CCA cancer tissues and 15

adjacent tissues. As shown in Fig. 6 a-c, except for *TEX30* ($P = 0.41$), eight of the nine genes included in the prognostic model were validated, with P values ranging from 2.0×10^{-54} to 7.2×10^{-03}. According to the qRT-PCR experiment results, compared with the normal cell line, the majority of the target genes were significantly aberrantly expressed in each CCA cell line ($P < 0.05$), except *HELLS* and *MFSD2A* in the RBE cell line (Fig. 7 a-i).

Discussion

CCA is a highly fatal digestive disease with poor diagnosis and prognosis. Traditional clinicopathological and physiological characteristics have been examined to diagnose the disease and reflect cancer progression. Unfortunately, the outcomes are still unsatisfactory. In addition, CCA cells potentially stem from both hepatocytes and cholangiocytes, and the differentiated origin of CCA possibly represents specific characteristic features [32, 33]. However, the molecular characterization and clinical features of different subgroups of CCA are still unclear. Hence, molecular prognostic markers, as an effective and sensitive method, may be applied as a beneficial supplement to traditional clinicopathological parameters for bile duct cancer patients to increase prognostic prediction, early diagnostic precision, subgroup identification and personalized treatment. Molecular markers are a dynamic tool that can be quantified by standardized detection procedures that fluctuate with tumor progression to reflect the prognosis of CCA patients in real time. Moreover, the nine-gene prognostic model including genes may also play essential roles in the progression of CCA and serve as novel targets for therapy.

In the current study, we identified 3,414 replicated DEGs in two independent cohorts of CCA, and 122 of them were significantly associated with OS. A novel nine-gene-based signature was constructed by LASSO Cox regression to predict the OS of CCA. The prognostic signature was established in the training dataset and further validated in a dataset with a large number of CCA patients with high sensitivities and accuracies. Heterogeneity among different populations has been widely reported to influence the biological mechanisms, management, and pathological features of CCA patients [4, 8]. Thus, constructing a prognostic biomarker with high precision and sensitivity in different

populations is urgently needed to improve prognosis. The majority of patients enrolled in the TCGA and GSE107943 datasets were European and American. Hence, to detect the precision and stability of the prognostic model in different populations, we evaluated the nine-gene-based model in 255 Chinese CCA patients. Strikingly, the low-risk group separated by the novel prognostic model had a better prognosis than the high-risk group in the training and validation cohorts. In addition, to validate the aberrant expression profile, the prognostic model including genes was subjected to DEG analyses with 15 paired Chinese CCA tumor tissues and adjacent counterpart controls. Among these genes, eight of the nine genes, *HELLS*, *HOXC6*, *MFSD2A*, *OTX1*, *PTGES*, *PYGB*, *SMOC1* and *ZBTB12,* were significantly aberrantly expressed compared with tumor tissues and adjacent controls. Furthermore, consistent with the analysis results, the mRNA expression levels of *HELLS*, *HOXC6*, *OTX1*, *PTGES*, *PYGB* and *ZBTB12* were significantly overexpressed in CCA cell lines compared with normal bile duct cells, and *MFSD2A*, *SMOC1* and *TEX30* were significantly silenced in CCA cell lines. However, *HELLS* and *MFSD2A* did not reach the significance threshold in the RBE cell line ($P > 0.05$).

Five of those nine prognostic genes (*HELLS*, *HOXC6*, *MFSD2A*, *OTX1* and *PYGB*) were previously reported to be associated with various cancers. *HELLS*, as one of the critical genes for chromatin modifiers, plays oncogenic roles in the progression and development of pancreatic cancer, gastric cancer, lymphoma and lung cancer [34-37]. Overexpression of *HELLS* is associated with poor prognosis in pancreatic cancer [38]. *HELLS* expression regulates the hub gene of cytokinesis-related genes, which cooperates the process of cellular proliferation, cytoskeleton organization and cytokinesis. *HOXC6*, *OTX1*, and *PYGB* serve as prognostic biomarkers for cervical cancer, gastric cancer, and lung cancer [39-41]. In several carcinomas, high expression levels of *HOXC6*, *OTX1*, and *PYGB* were significantly associated with increased tumor invasion and migration and poor prognosis. Epigenetic silencing of *MFSD2A* was associated with better prognosis in gastric cancer and lung cancer [42, 43]. However, the biological function of these genes in CCA is still unclear, and further investigation is needed to explore the potential mechanisms of these target genes in the future.

The tumor immune microenvironment is considered a key regulator of carcinogenesis,

angiogenesis, and tumor growth and affects the efficacy of radiotherapy, chemotherapy, and immune checkpoint therapy [44, 45]. Bile duct carcinoma is characterized by an intricate microenvironment in which the stroma is composed of tumor-associated endothelial cells, fibroblasts, macrophages, neutrophils, NK cells, and T cells [46]. According to the infiltration of immune cells in CCA, we detected more M0 macrophage infiltration in the high-risk group in the training and validation datasets. Furthermore, M2 macrophage infiltration was significantly decreased in the low-risk group in the validation cohort, whereas significant infiltration was not identified in the training dataset. M0 macrophages are nonactivated macrophages and differentiate into M1/2 macrophages in the presence of specific conditions. M1 macrophage infiltration promotes an inflammatory microenvironment by increasing the expression of IL-1, IL-6 and IL-12, and M2 macrophage infiltration promotes inflammatory escape by increasing the expression of PD-L1 and IL-10 [47, 48]. A previous study suggested that patients with a higher infiltration of CD8+ T cells and M1 macrophages in the TME have a better prognosis and high immune checkpoint gene expression than those infiltrated by M0/2 macrophages [22, 49]. In our results, the microenvironment of the CCA high-risk group had significantly enriched M0 macrophage infiltration. Therefore, it seems that suppressing the infiltration of M0/2 macrophages might improve the prognosis of CCA patients.

In addition, the risk score for CCA was positively correlated with neutrophils in the training cohort and positively correlated with activated MCs and negatively correlated with activated memory resting NK cells in the validation cohort. Previous studies reported that the abundance of cancer-associated fibroblast cells was positively correlated with tumor growth and poor prognosis [50, 51]. Cancer-associated fibroblasts regulate innate immunity by suppressing NK-cell activation to promote immunosuppression [52]. Moreover, NK-cell-induced antibody-dependent cellular cytotoxicity increased cancer cell death, and infusion of NK cells in CCA mouse xenograft models resulted in tumor regression [53, 54]. Immune checkpoint blockade with monoclonal antibody therapies has been implemented as a novel plan for treating numerous malignancies with remarkable and durable response rates [55], while clinical trials of immunotherapies in CCA have displayed less notable success [8]. NK cells are currently widely targeted in

immunotherapy for cancer patients without major histocompatibility complex (MHC) class I [56]. When the MHC I expression level is downregulated or silenced in cancer cells, NK cells are activated to remove tumor cells. Laura and her colleagues demonstrated that activated MCs could induce biliary hyperplasia, hepatic fibrosis and vascular bed dysfunction, and knockout or inhibition of MCs was likely to be a better therapy for patients with cholangiopathies [57]. Our results extend the knowledge that downregulated MCs could also improve the OS of cholangiocarcinoma patients. In summary, these results strongly highlighted that upregulated activated NK-cell infiltration and downregulated resting MC infiltration in the low-risk group of the prognostic model might exhibit tumor-inhibiting effects in CCA. Furthermore, it is necessary to explore the therapeutic value of the TME of M0 macrophages, NK cells and MCs in immune checkpoint therapy for CCA.

For the drug sensitivity analyses, 5 kinds of therapeutic drugs, erlotinib, ML323, AGI-6780, Gallibiscoquinazole and AZD3759, revealed more sensitivity in the high-risk group than in the low-risk group. Erlotinib, a protein kinase inhibitor, can be targeted to suppress the epidermal growth factor receptor system in several different types of cancer [58]. However, a phase III clinical trial of CCA with erlotinib showed negative results [59]. In our results, the IC50 value of erlotinib in the low-risk group was significantly lower than that in the high-risk group, which suggested that erlotinib may be more effective for the low-risk group than for the high-risk group. Thus, the novel prognostic model probability can serve as a personalized treatment indicator for CCA patients.

Although comprehensive analyses were performed to investigate the potential mechanisms of the prognostic biomarkers for CCA, this study has certain limitations. First, immune cell infiltration in bile duct cancer was assessed on the basis of publicly available TCGA, GSE107943 and NGDC RNA-seq data. Hence, further biological experiments and single-cell and/or spatial transcriptome sequencing need to be implemented in vivo and in vitro to validate the immunotherapeutic value of combining M0 macrophages, NK cells and/or MCs in CCA. The regulation of M0 macrophages, NK cells and MCs might play important roles in developing therapeutics for CCA. Second, even if half of the genes included in the prognostic model have been reported to influence different cancers,

biological experiments should be performed to elucidate the molecular mechanisms of the target genes in the pathogenesis and prognosis of CCA.

Conclusions

In summary, our results identified and verified an accurate and sensitive nine-gene signature to predict the OS of bile duct cancer in different populations. In addition, this prediction model potentially indicated that M0 macrophage, NK-cell and MC infiltration in CCA patients affects the prognostic process. These findings provide a novel prognostic signature to account for the OS of CCA patients and may benefit the timely diagnosis and individualized treatment of CCA.

Figure legends:

Figure 1. Workflow presenting the steps of constructing the novel nine-gene prognostic model of CCA in the study.

Figure 2. Identification of specific DEGs in bile duct carcinoma. (a) Topological overlap matrix plots for CCA modules. (b) Clustering dendrograms of transcriptomes based on topological overlap to distinguish the different modules and every module annotated by an arbitrary color. Correlation analysis was implemented to identify the relationship between different modules and five covariates. Red indicates that the module was positively correlated with the covariate, and green indicates a negative correlation. (c) Venn diagram analysis of DEGs in CCA with the training dataset and validation dataset. DEGs with $P_{bonf} < 0.05$ and absolute value of $\log_2$fold change >1 were considered to be significant. (d) The specific DEG-enriched pathways in CCA. The X axis indicates the –log10 P_{fdr} value.

Figure 3. Effective prognostic signature for CCA patients. (a, b) LASSO Cox regression identified nine genes related to OS. (c) ROC curve analysis of the nine-gene prognostic model for predicting OS in the training dataset (d) and validation dataset. (e) OS curves comparing the high-risk group and low-risk group defined by the prognostic model in CCA in the training dataset (f) and validation dataset. (g) Multivariate Cox analysis of the nine genes.

Figure 4. Analysis of immune cell infiltration in CCA. (a) The proportion of 22 immune-related cells in CCA was assessed through the CIBERSORTx algorithm in the training cohort (n = 66) (b) and validation cohort (n = 255). Significant differences between the two groups were evaluated using the Wilcoxon test. Asterisks indicate significance, and the Y-axis indicates the percentage of immune cells in CCA (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$). Red triangle: continual increase; Blue triangle: continual decrease.

Figure 5. Specific characteristics of the high-risk and low-risk groups of CCA patients. (a-e) Estimated IC50 value differences of chemotherapeutic drugs and novel inhibitors between the high-risk and low-risk score groups in the discovery cohort and (f-g) validation cohort. (i) GSEA results indicating differential enrichment of genes in KEGG terms with high and low risk scores in the training and (k) validation cohorts.

Figure 6. Schematic diagram of identified epigenetic aberrations between tumor tissues and normal tissues in CCA. (a) Box plots indicating differentially expressed profiles between tumor tissues and normal counterparts for constituents of the prognostic model in the TCGA-CCA cohort (n = 45), (b) GSE107943 cohort (n = 57) and (c) GSE119336 cohort (n = 30). The Y axis indicates the –log10 *P* value, and the X axis depicts different gene expression profiles in the training and validation datasets. Red represents adjacent tissues, and blue represents tumor tissues.

Figure 7. Expression patterns of the genes included in the prognostic model at the mRNA level evaluated via qRT-PCR. Relative expression differences of (a) *HELLS*, (b) *HOXC6*, (c) *MFSD2A*, (d) *OTX1*, (e) *ZBTB12*, (f) *PTGES*, (g) *SMOC1*, (h) *TEX30* and (i) *PYGB* between the normal liver cell line (LO2) and four CCA cell lines (Hep-li5, HCCCT1, HCCC9810 and RBE) via qRT-PCR experiments (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ***$P < 0.0001$).

Abbreviations

CCA: Cholangiocarcinoma; TCGA: The Cancer Genome Atlas; GEO: Gene Expression Omnibus; NGDC: National Genomics Data Center; DEGs: Differentially expressed genes; OS: overall survival; CA19-9: carbohydrate antigen 19-9; RPKM: Reads per kilobase per

million mapped reads; KEGG: Kyoto Encyclopedia of Genes and Genomes; GSEA: Gene Set enrichment analysis; K-M: Kaplan–Meier analysis; qRT-PCR: Quantitative real-time PCR; ROC: time-dependent receiver operating characteristic curves; NK: Resting natural killer cells; MCs: GDSC: Resting mast cells;Genomics of Drug Sensitivity in Cancer; IC50: Half-maximal inhibitory concentration.

Declarations

Availability of data and materials

The data set analyzed during the current study are all available in public datasets. The RNA-seq dataset and clinical information from TCGA (https://gdc.xenahubs.net), GSE107943, GSE119336 (https://www.ncbi.nlm.nih.gov/geo/) and NGDC database (https://www.biosino.org/node/project/detail/OEP001105). The immune infiltration-related gene data were downloaded from CIBERSORTX (https://cibersortx.stanford.edu/).

Acknowledgements

We thank Dr. Jianfeng Yang for excellent editing of this manuscript, and we acknowledge TCGA, GEO, and NGDC database for uploading their datasets.

Funding

This study was supported in part by the Zhejiang Medical and Health Science and Technology Plan (Grant Nos. WKJ-ZJ-2136, 2018PY037 and 2022RC056), the Hangzhou Major Science and Technology projects (Grant No. 202004A14), and the Hangzhou Medical and Health Science and Technology Plan (Grant Nos. OO20190610 and A20200174).

Author information

Qiang Liu and Liyun Zheng are co-first authors.

Affiliations

Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China.

Qiang Liu & Liyun Zheng & Dongchao Xu & Hangbin Jin & Sile Cheng & Hongzhang Shen & Ye Gu & Shenghui Chen & Xiaofeng Zhang & Jianfeng Yang

The Fourth School of Clinical Medicine, Zhejiang Chinese Medical University, Hangzhou, China.

Jianpeng Zhu & Ying Bian

Contributions

QL, LZ and HS collected the data, performed the bioinformatic analysis and wrote the manuscript. SC, HJ, LL, GY, JPZ, YB and DX were involved in data collection and reviewed the manuscript. XZ and JY conceived the current study and reviewed the manuscript. All authors have reviewed the subsequent versions and read and approved the final manuscript.

Corresponding author

Correspondence to Jianfeng Yang or Xiaofeng Zhang.

Ethics declarations

Not applicable. All data in this study are publicly available and no permission was required to perform this study. No informed consent was required as experimental data with no personal identifiers was used, which was waived by the ethical committee of the Affiliated Hangzhou First People's Hospital. Our study is based on public source data of CCA, and there are no ethical issues and other conflicts of interest.

Consent for publication

All authors have read and approved the manuscript for submission to "*BMC Cancer*".

Competing interests

The authors declare that they have no conflict of interest.

References

1. Charbel, H. and F.H. Al-Kawas, *Cholangiocarcinoma: epidemiology, risk factors, pathogenesis, and diagnosis.* Curr Gastroenterol Rep, 2011. **13**(2): p. 182-7.
2. Banales, J.M., 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): p. 261-80.
3. DeOliveira, M.L., et al., *Cholangiocarcinoma: thirty-one-year experience with 564 patients at a single institution.* Ann Surg, 2007. **245**(5): p. 755-62.
4. Razumilava, N. and G.J. Gores, *Cholangiocarcinoma.* Lancet, 2014. **383**(9935): p. 2168-79.
5. Everhart, J.E. and C.E. Ruhl, *Burden of digestive diseases in the United States Part III: Liver, biliary tract, and pancreas.* Gastroenterology, 2009. **136**(4): p. 1134-44.
6. Saha, S.K., et al., *Forty-Year Trends in Cholangiocarcinoma Incidence in the U.S.: Intrahepatic Disease on the Rise.* Oncologist, 2016. **21**(5): p. 594-9.
7. Khan, S.A., et al., *Guidelines for the diagnosis and treatment of cholangiocarcinoma: consensus document.* Gut, 2002. **51 Suppl 6**: p. VI1-9.
8. Banales, J.M., et al., *Cholangiocarcinoma 2020: the next horizon in mechanisms and management.* Nat Rev Gastroenterol Hepatol, 2020. **17**(9): p. 557-588.
9. Spolverato, G., et al., *Management and Outcomes of Patients with Recurrent Intrahepatic Cholangiocarcinoma Following Previous Curative-Intent Surgical Resection.* Ann Surg Oncol, 2016. **23**(1): p. 235-43.
10. Yang, J., et al., *Endoscopic radiofrequency ablation plus a novel oral 5-fluorouracil compound versus radiofrequency ablation alone for unresectable extrahepatic cholangiocarcinoma.* Gastrointest Endosc, 2020. **92**(6): p. 1204-1212 e1.
11. Venkatesh, P.G., et al., *Increased serum levels of carbohydrate antigen 19-9 and outcomes in primary sclerosing cholangitis patients without cholangiocarcinoma.* Dig Dis Sci, 2013. **58**(3): p. 850-7.
12. Sinakos, E., et al., *Many patients with primary sclerosing cholangitis and increased serum levels of carbohydrate antigen 19-9 do not have cholangiocarcinoma.* Clin Gastroenterol Hepatol, 2011. **9**(5): p. 434-9 e1.
13. Patel, A.H., et al., *The utility of CA 19-9 in the diagnoses of cholangiocarcinoma in patients without primary sclerosing cholangitis.* Am J Gastroenterol, 2000. **95**(1): p. 204-7.
14. Kendall, T., et al., *Anatomical, histomorphological and molecular classification of cholangiocarcinoma.* Liver Int, 2019. **39 Suppl 1**: p. 7-18.

15. Bertuccio, P., et al., *Global trends in mortality from intrahepatic and extrahepatic cholangiocarcinoma.* J Hepatol, 2019. **71**(1): p. 104-114.
16. Farshidfar, F., et al., *Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles.* Cell Rep, 2017. **19**(13): p. 2878-2880.
17. Blum, A., P. Wang, and J.C. Zenklusen, *SnapShot: TCGA-Analyzed Tumors.* Cell, 2018. **173**(2): p. 530.
18. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.
19. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
20. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
21. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* OMICS, 2012. **16**(5): p. 284-7.
22. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
23. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* J Stat Softw, 2010. **33**(1): p. 1-22.
24. Newman, A.M., et al., *Robust enumeration of cell subsets from tissue expression profiles.* Nat Methods, 2015. **12**(5): p. 453-7.
25. Newman, A.M., et al., *Determining cell type abundance and expression from bulk tissues with digital cytometry.* Nat Biotechnol, 2019. **37**(7): p. 773-782.
26. Maeser, D., R.F. Gruener, and R.S. Huang, *oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data.* Brief Bioinform, 2021. **22**(6).
27. Bergers, G. and S.M. Fendt, *The metabolism of cancer cells during metastasis.* Nat Rev Cancer, 2021. **21**(3): p. 162-180.
28. Zhu, J., et al., *High Expression of PHGDH Predicts Poor Prognosis in Non-Small Cell Lung Cancer.* GTransl Oncol, 2016. **9**(6): p. 592-599.
29. Song, Z., et al., *PHGDH is an independent prognosis marker and contributes cell proliferation, migration and invasion in human pancreatic cancer.* Gene, 2018. **642**: p. 43-50.
30. Phannasil, P., et al., *Mass spectrometry analysis shows the biosynthetic pathways supported by pyruvate carboxylase in highly invasive breast cancer cells.* Biochim Biophys Acta Mol Basis Dis, 2017. **1863**(2): p. 537-551.
31. Zhang, D., et al., *Metabolic regulation of gene expression by histone lactylation.* Nature, 2019. **574**(7779): p. 575-580.
32. Zhu, Y. and L.N. Kwong, *Insights Into the Origin of Intrahepatic Cholangiocarcinoma From Mouse Models.* Hepatology, 2020. **72**(1): p. 305-314.
33. Dong, L., et al., *Proteogenomic characterization identifies clinically relevant subgroups of intrahepatic cholangiocarcinoma.* Cancer Cell, 2022. **40**(1): p. 70-87 e15.
34. Hou, X., et al., *HELLS, a chromatin remodeler is highly expressed in pancreatic cancer and downregulation of it impairs tumor growth and sensitizes to cisplatin by reexpressing the tumor suppressor TGFBR3.* Cancer Med, 2021. **10**(1): p. 350-364.
35. Yang, R., et al., *MiR-365a-3p-Mediated Regulation of HELLS/GLUT1 Axis Suppresses Aerobic Glycolysis and Gastric Cancer Growth.* Front Oncol, 2021. **11**: p. 616390.

36. Zhu, W., et al., *Identification and validation of HELLS (Helicase, Lymphoid-Specific) and ICAM1 (Intercellular adhesion molecule 1) as potential diagnostic biomarkers of lung cancer.* PeerJ, 2020. **8**: p. e8731.
37. Tameni, A., et al., *The DNA-helicase HELLS drives ALK(-) ALCL proliferation by the transcriptional control of a cytokinesis-related program.* Cell Death Dis, 2021. **12**(1): p. 130.
38. Wang, F.J., et al., *HELLS serves as a poor prognostic biomarker and its downregulation reserves the malignant phenotype in pancreatic cancer.* BMC Med Genomics, 2021. **14**(1): p. 189.
39. Wang, Y., et al., *HOXC6 promotes cervical cancer progression via regulation of Bcl-2.* FASEB J, 2019. **33**(3): p. 3901-3911.
40. Qin, S.C., et al., *Dowregulation of OTX1 attenuates gastric cancer cell proliferation, migration and invasion.* Oncol Rep, 2018. **40**(4): p. 1907-1916.
41. Xiao, L., et al., *PYGB facilitates cell proliferation and invasiveness in non-small cell lung cancer by activating the Wnt-beta-catenin signaling pathway.* Biochem Cell Biol, 2020. **98**(5): p. 565-574.
42. Shi, X., et al., *MFSD2A expression predicts better prognosis in gastric cancer.* Biochem Biophys Res Commun, 2018. **505**(3): p. 699-704.
43. Spinola, M., et al., *MFSD2A is a novel lung tumor suppressor gene modulating cell cycle and matrix attachment.* Mol Cancer, 2010. **9**: p. 62.
44. Anderson, N.M. and M.C. Simon, *The tumor microenvironment.* Curr Biol, 2020. **30**(16): p. R921-R925.
45. Kumagai, S., et al., *An Oncogenic Alteration Creates a Microenvironment that Promotes Tumor Progression by Conferring a Metabolic Advantage to Regulatory T Cells.* Immunity, 2020. **53**(1): p. 187-203 e8.
46. Tamma, R., et al., *Inflammatory cells infiltrate and angiogenesis in locally advanced and metastatic cholangiocarcinoma.* Eur J Clin Invest, 2019. **49**(5): p. e13087.
47. Dufresne, A., et al., *Specific immune landscapes and immune checkpoint expressions in histotypes and molecular subtypes of sarcoma.* Oncoimmunology, 2020. **9**(1): p. 1792036.
48. Tamura, R., et al., *Dual role of macrophage in tumor immunity.* Immunotherapy, 2018. **10**(10): p. 899-909.
49. Zhu, N. and J. Hou, *Assessing immune infiltration and the tumor microenvironment for the diagnosis and prognosis of sarcoma.* Cancer Cell Int, 2020. **20**(1): p. 577.
50. Kalluri, R., *The biology and function of fibroblasts in cancer.* Nat Rev Cancer, 2016. **16**(9): p. 582-98.
51. Chuaysri, C., et al., *Alpha-smooth muscle actin-positive fibroblasts promote biliary cell proliferation and correlate with poor survival in cholangiocarcinoma.* Oncol Rep, 2009. **21**(4): p. 957-69.
52. Ziani, L., S. Chouaib, and J. Thiery, *Alteration of the Antitumor Immune Response by Cancer-Associated Fibroblasts.* Front Immunol, 2018. **9**: p. 414.
53. Morisaki, T., et al., *Combining cetuximab with killer lymphocytes synergistically inhibits human cholangiocarcinoma cells in vitro.* Anticancer Res, 2012. **32**(6): p. 2249-56.
54. Jung, I.H., et al., *In Vivo Study of Natural Killer (NK) Cell Cytotoxicity Against Cholangiocarcinoma in a Nude Mouse Model.* In Vivo, 2018. **32**(4): p. 771-781.
55. Wei, S.C., C.R. Duffy, and J.P. Allison, *Fundamental Mechanisms of Immune Checkpoint Blockade Therapy.* Cancer Discov, 2018. **8**(9): p. 1069-1086.
56. Purdy, A.K. and K.S. Campbell, *Natural killer cells and cancer: regulation by the killer cell Ig-like receptors (KIR).* Cancer Biol Ther, 2009. **8**(23): p. 2211-20.
57. Hargrove, L., et al., *Bile duct ligation-induced biliary hyperplasia, hepatic injury, and fibrosis are*

reduced in mast cell-deficient Kit(W-sh) mice. Hepatology, 2017. **65**(6): p. 1991-2004.
58. Jensen, L.H., *Clinical aspects and perspectives of erlotinib in the treatment of patients with biliary tract cancer.* Expert Opin Investig Drugs, 2016. **25**(3): p. 359-65.
59. Lee, J., et al., *Gemcitabine and oxaliplatin with or without erlotinib in advanced biliary-tract cancer: a multicentre, open-label, randomised, phase 3 study.* Lancet Oncol, 2012. **13**(2): p. 181-8.

Figures

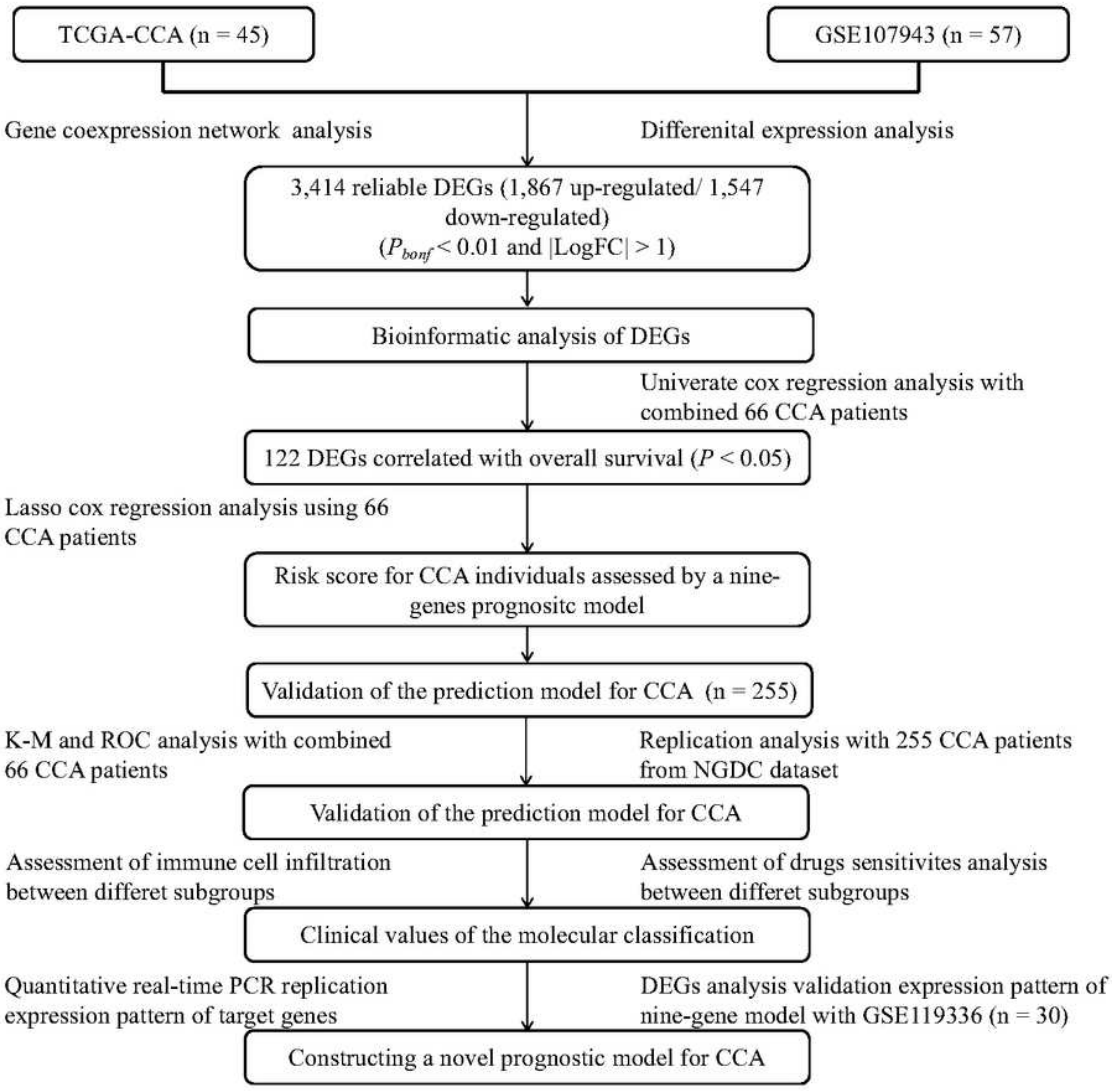

Figure 1

Workflow presenting the steps of constructing the novel nine-gene prognostic model of CCA in the study

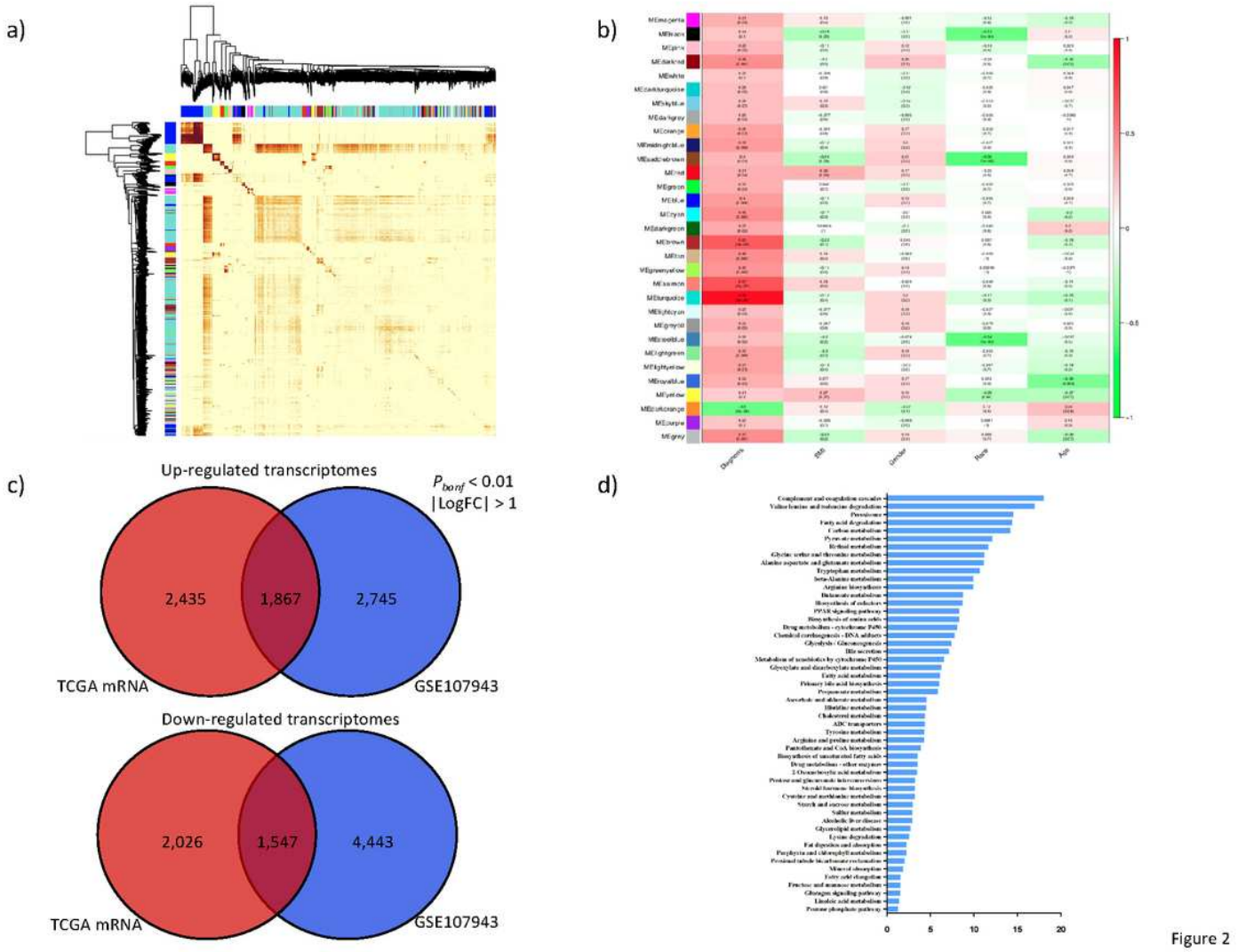

Figure 2

Figure 2

Identification of specific DEGs in bile duct carcinoma. (a) Topological overlap matrix plots for CCA modules. (b) Clustering dendrograms of transcriptomes based on topological overlap to distinguish the different modules and every module annotated by an arbitrary color. Correlation analysis was implemented to identify the relationship between different modules and five covariates. Red indicates that the module was positively correlated with the covariate, and green indicates a negative correlation. (c) Venn diagram analysis of DEGs in CCA with the training dataset and validation dataset. DEGs with Pbonf < 0.05 and absolute value of log2fold change >1 were considered to be significant. (d) The specific DEG-enriched pathways in CCA. The X axis indicates the −log10 Pfdr value.

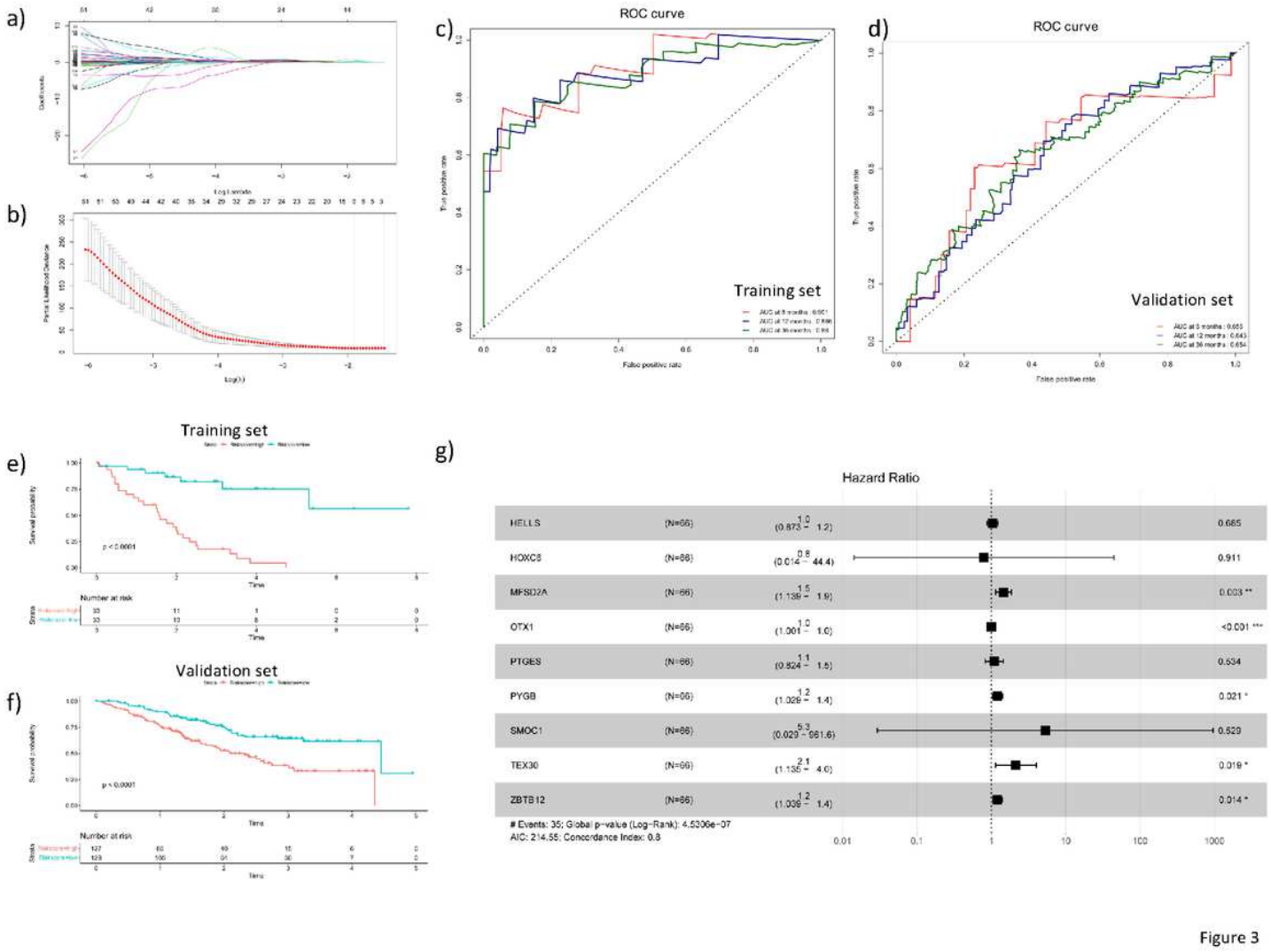

Figure 3

Figure 3

Effective prognostic signature for CCA patients. (a, b) LASSO Cox regression identified nine genes related to OS. (c) ROC curve analysis of the nine-gene prognostic model for predicting OS in the training dataset (d) and validation dataset. (e) OS curves comparing the high-risk group and low-risk group defined by the prognostic model in CCA in the training dataset (f) and validation dataset. (g) Multivariate Cox analysis of the nine genes.

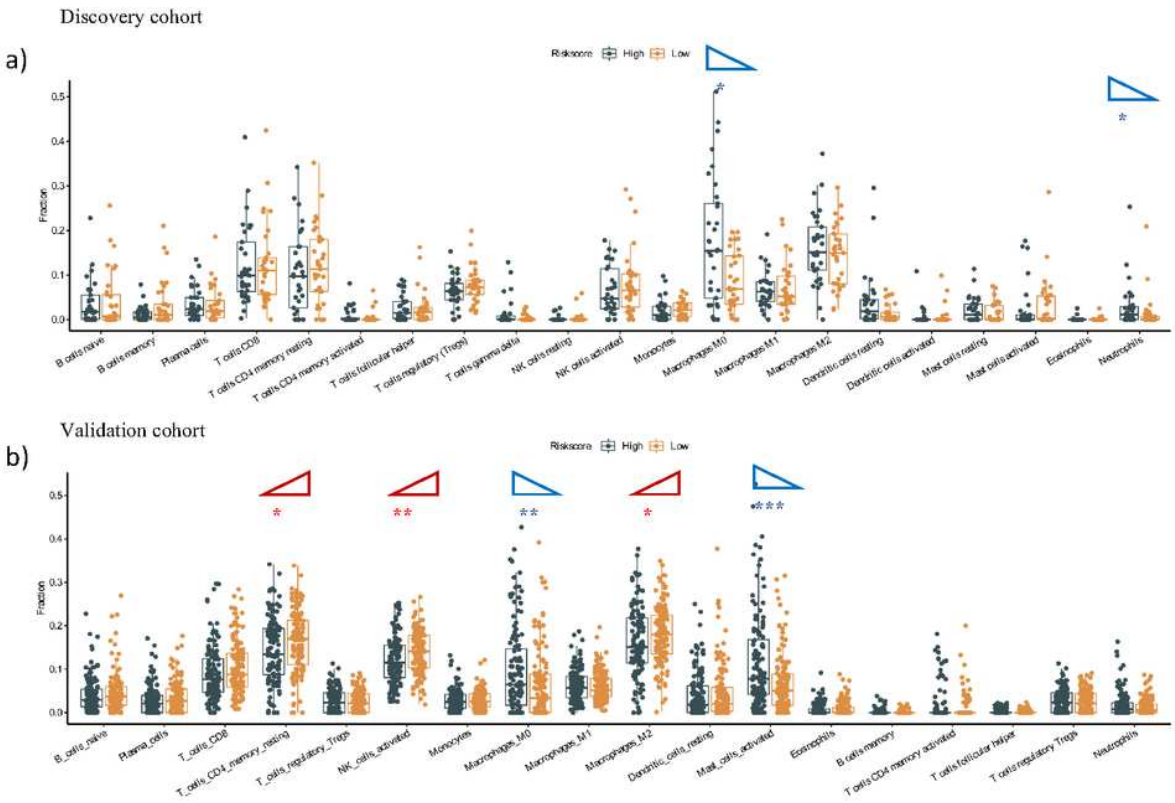

Figure 4

Figure 4

Analysis of immune cell infiltration in CCA. (a) The proportion of 22 immune-related cells in CCA was assessed through the CIBERSORTx algorithm in the training cohort (n = 66) (b) and validation cohort (n = 255). Significant differences between the two groups were evaluated using the Wilcoxon test. Asterisks indicate significance, and the Y-axis indicates the percentage of immune cells in CCA (*P < 0.05; **P < 0.01; ***P < 0.001). Red triangle: continual increase; Blue triangle: continual decrease

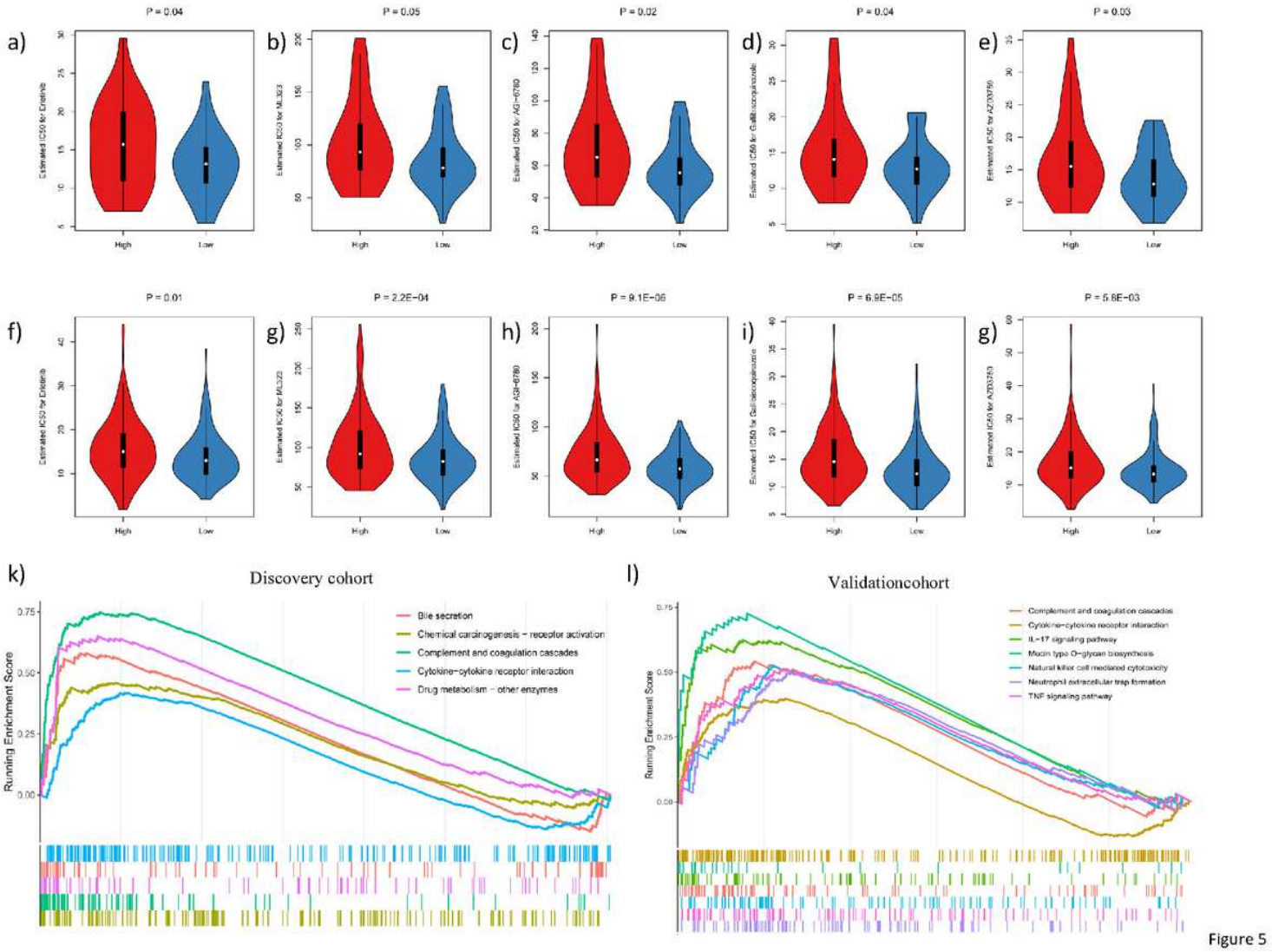

Figure 5

Specific characteristics of the high-risk and low-risk groups of CCA patients. (a-e) Estimated IC50 value differences of chemotherapeutic drugs and novel inhibitors between the high-risk and low-risk score groups in the discovery cohort and (f-g) validation cohort. (i) GSEA results indicating differential enrichment of genes in KEGG terms with high and low risk scores in the training and (k) validation cohorts.

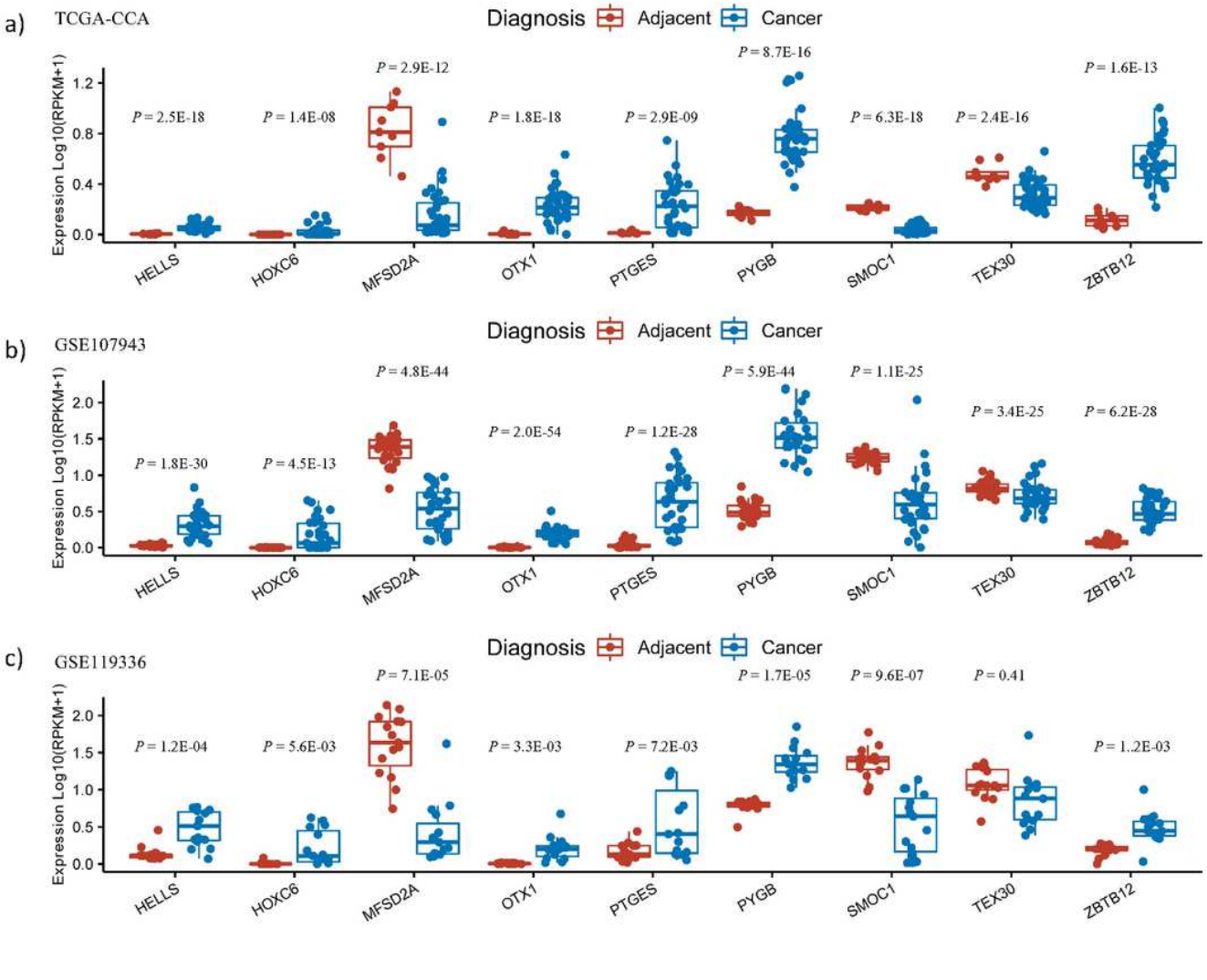

Figure 6

Figure 6

Schematic diagram of identified epigenetic aberrations between tumor tissues and normal tissues in CCA. (a) Box plots indicating differentially expressed profiles between tumor tissues and normal counterparts for constituents of the prognostic model in the TCGA-CCA cohort (n = 45), (b) GSE107943 cohort (n = 57) and (c) GSE119336 cohort (n = 30). The Y axis indicates the –log10 P value, and the X axis depicts different gene expression profiles in the training and validation datasets. Red represents adjacent tissues, and blue represents tumor tissues.

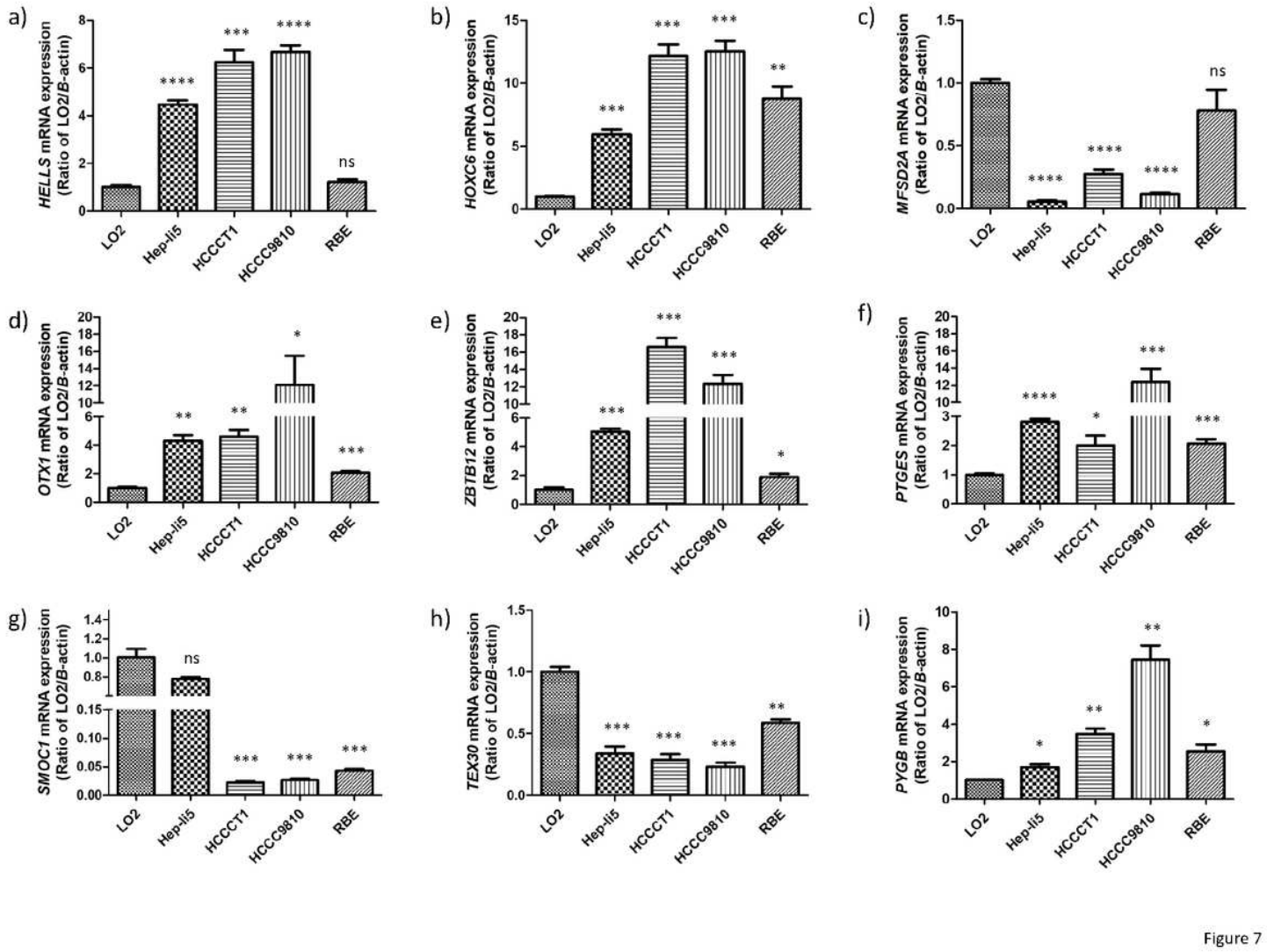

Figure 7

Figure 7

Expression patterns of the genes included in the prognostic model at the mRNA level evaluated via qRT-PCR. Relative expression differences of (a) HELLS, (b) HOXC6, (c) MFSD2A, (d) OTX1, (e) ZBTB12, (f) PTGES, (g) SMOC1, (h) TEX30 and (i) PYGB between the normal liver cell line (LO2) and four CCA cell lines (Hep-li5, HCCCT1, HCCC9810 and RBE) via qRT-PCR experiments (*P < 0.05; **P < 0.01; ***P < 0.001; ***P < 0.0001).

Supplementary Files

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

- Table.pdf
- CHOLSupplementaryTable.xlsx