Hepatology Research 2019; **49**: 1195–1206 doi: 10.1111/hepr.13386

Original Article

Weighted gene coexpression network analysis reveals hub genes involved in cholangiocarcinoma progression and prognosis

Aiping Tian,[1] Ke Pu,[2] Boxuan Li,[3,4] Min Li,[1] Xiaoguang Liu,[5] Liping Gao[6] and Xiaorong Mao[1,6]

[1]Departments of Infectious Diseases, [2]Key Laboratory for Gastrointestinal Diseases of Gansu Province, [3]School of Pharmacy, [4]Pharmacy, and [5]Rheumatology, The First Hospital of Lanzhou University and [6]The First Clinical Medical College, Lanzhou University, Lanzhou, China

Aim: Cholangiocarcinoma (CCA) is a highly malignant tumor found in the bile duct epithelial cells, and the second most common primary tumor of the liver. However, the pivotal roles of molecular biomarkers in oncogenesis of CCA are unclear. Therefore, we aim to explore the underlying mechanisms of progression and screen for novel prognostic biomarkers and treatment targets.

Method: The data of mRNA sequencing and clinical information of CCA patients in The Cancer Genome Atlas was analyzed by weighted gene coexpression network analysis (WGCNA). Modules and clinical traits were constructed according to Pearson's correlation analysis, and Gene Ontology and pathway enrichment analysis were applied. Hub genes of these modules were screened by intramodule analysis; Cytoscape with Search Tool for the Retrieval of Interacting Genes was utilized to visualize protein–protein interaction of these modules; hub genes of these modules were validated afterwards. Furthermore, the significance of these genes was confirmed by survival analysis.

Results: Genes *MRPS18A*, *CST1*, and *SCP2* were identified as candidate genes in the module, which was associated with clinical traits including pathological stage, histological grade, and liver function and which also affected overall survival of CCA patients. Nineteen hub genes were analyzed together and were associated with progression and prognosis of CCA. Survival analyses found that several of the multiple genes could serve as biomarkers to stratify CCA patients into low- and high-risk groups.

Conclusion: These candidate genes could be involved in progression of CCA, which could serve as novel prognostic markers and treatment targets. Moreover, most of them were first reported in CCA and deserve further research.

Key words: carcinogenesis, cholangiocarcinoma, weighted gene coexpression network analysis (WGCNA)

INTRODUCTION

CHOLANGIOCARCINOMA (CCA) IS a highly malignant tumor located in the bile duct epithelial cells; it is the second most common primary tumor of the liver. Cholangiocarcinoma are subdivided into intrahepatic cholangiocarcinoma (iCCA), originating from the biliary tree within the liver, and extrahepatic cholangiocarcinoma (eCCA), outside the liver parenchyma; the latter is further subclassified into perihilar cholangiocarcinoma (pCCA) and distal cholangiocarcinoma, with a proportion of 10–20% iCCA, 50% pCCA, and 30–40% eCCA.[1] Most cholangiocarcinoma are well, moderately, and poorly differentiated adenocarcinomas with other histological subtypes encountered rarely.[2,3] The incidence of CCA has increased to 18% of all liver cancers during the past 40 years and account for 2% of cancer-related deaths worldwide per year.[3] The fact that CCA has a slow evolution, atypical symptoms, and limited therapeutic measures makes it difficult for early diagnosis and hence most patients are detected only in advanced or metastatic stages. The overall 5-year survival after resection is usually lower than 40%; in non-operable CCAs the overall 5-year survival is less than 5%.[4] Therefore, to establish reliable biomarkers that are specific to the early stages of the disease is needed urgently to overcome the poor prognosis and delayed treatment.

Correspondence: Dr Xiaorong Mao, Department of Infectious Diseases, The First Hospital of Lanzhou University, 1 West Donggang, Chengguan District, Lanzhou, Gansu 730000, China. Email: mxr2013@126.com
Received 10 August 2018; revision 9 April 2019; accepted 14 May 2019.

Serum biomarkers such as carbohydrate antigen 199 and cancer antigen 125 are now used routinely as laboratory tests for CCA screening.[5] However, the diagnostic sensitivity and specificity are not satisfactory and are insufficient for early detection. A study showed that benign bile duct obstruction diseases also show moderately elevated levels; these serum biomarkers have wide ranges of sensitivity (50–90%) and specificity (54–98%).[6]

In the last decade, numerous studies indicated that different genomic alterations are involved in the pathogenesis of CCA and reported their potential value for diagnosis and prognosis. Among these genes, *TP53* was presumed to affect DNA repair,[4,7] some are involved in cell growth pathways (*KRAS*, *SMAD4*, *PTPN3*, *BRAF*, and *FGFR2*),[4,7,8] some are worked during chromatin remodeling (*PBRM1*, *KMT2C*, *ARID1A*, and *BAP1*),[7–9] and others, such as the Notch and Wnt signaling pathways, play important roles in the development of CCA.[10,11] The *FGFR2* fusion genes are of particular interest, as they could not be detected in other liver malignancies, could be used for therapeutic target purposes, and also have diagnostic value.[12,13] It was also found that *IDH1* and *IDH2* alter the methylation status of CCA cells.[13] The genomic variability of CCAs could be the reflection of different etiologies or stages of tumor development, and thus be used as biomarkers and for targeted therapy. Aberrant transcriptomic change would occur subsequent to genetic and epigenetic modifications. Previous studies have confirmed the abnormal microRNA profiles in both CCA tumor tissues and cell lines.[14–16] Mass spectrometry (MS)-based proteomics has become a useful tool for the analysis of different biofluids to find accurate and specific protein biomarkers for risk stratification, diagnosis, and prognosis.[17] Studies focused on proteomics have identified a specific peptide, SSP411 protein, in bile and urine showed better diagnostic value than the general non-specific tumor markers used in serum.[18,19] However, MS proteomics are difficult to implement in these samples due to the abundance of high dynamic range proteins, such as albumin or immunoglobulins, making the discrimination of less abundance aimed proteins difficult.

Although widely investigated in the field of tumorigenesis, the analysis of gene expression data by the weighted gene coexpression network analysis (WGCNA) systems biology approach has not yet, so far as we know, been applied to CCA-derived data. The WGCNA allows a global interpretation of gene expression data by constructing gene networks based on similarities in expression profiles among samples.[20] Highly coexpressed genes are connected in the network and highly connected network regions can be grouped into modules. As these modules often consist of functionally related genes, different modules are involved in individual functions. Within the modules, WGCNA also allows the identification of the most central and connected genes, called hub genes.[21] These modules and their hub genes could be involved in pathogenesis development and, therefore, might have important clinical applicability as potential prognostic biomarkers or as therapeutic targets. In this study, in order to improve our understanding of the biological mechanisms underlying CCA, we analyzed CCA gene expression datasets by constructing a coexpression network analysis strategy to detect key genes potentially involved in the carcinogenesis of CCA.

METHODS

Analysis of gene expression data

RNA SEQUENCING DATASETS and clinical information of CCA patients were downloaded from The Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov/). RNA sequencing data were derived from Illumina (San Diego, CA, USA) HiSeq V2 RSEM genes. The gene expression level was measured as fragments per kilobase of transcript per million mapped reads. Clinical follow-up data of CCA patients in TCGA were retrieved for prognostic analysis. Clinical information, including American Joint Committee on Cancer pathological tumor–node–metastasis (TNM) stage, which was carried out according to the TNM 2010 system and is specific for every subtype of CCA,[22] gender, age at initial pathological diagnosis, and tumor type, were all extracted for WGCNA analysis.

Gene coexpression network construction

Coexpression networks were constructed according to the protocol of the WGCNA package in R language environment.[20] The similarity between the gene expression profiles were calculated based on a matrix of pairwise Pearson's correlation coefficients, which represented a measure of the degree of concordance between gene expression profiles. Then we used the WGCNA function adjacency to transform the similarity matrix into an adjacency matrix.

Scale-free gene coexpression networks were constructed by the WGCNA package.[20] To ensure that the results of network construction were reliable, outlier samples whose connectivity was less than −2.5 were removed.[23] Function pickSoftThreshold was used to calculate scale-free topology fitting indices R^2 for several soft threshold powers. When the power value is approximately 0.8, it means that the topology of the network is scale-free, and there are no

batch-effects. The adjacency matrix was transformed into a topological overlap matrix (TOM) and the corresponding dissimilarity (dissTOM). The resulting topological overlap is a biologically meaningful measure of gene similarity based on coexpression relationships between two genes.[24] Module identification was undertaken with the dynamic tree cut method for branch cutting to generate modules on the basis of hierarchically clustering genes using dissTOM. During the process, the distance measure with a deepSplit value of 2 to branch splitting and a minimum size cut-off of 30 (minClusterSize = 30) for the resulting dendrogram were chosen to avoid generating abnormal modules. Highly similar modules were identified by clustering and then merged together with a height cut-off of 0.25.

Construction of module–trait relationships of CCA

The module eigengene (ME), defined as the first principal component of a given module, was calculated by function module eigengenes. The ME can be considered to be representative of gene expression profiles and to capture the maximal amount of variation in the module. Modules would be merged if their correlation of MEs was greater than 0.75, which means they have similar expression profiles. The correlation between MEs and clinical traits, including pathologic stage, histologic type and grade, liver function (Child–Pugh classification), and Ishak fibrosis staging were evaluated by Pearson's correlation tests. $P < 0.05$ was considered to be significantly correlated.

Finding meaningful modules and functional annotations

The correlation between modules and clinical features was evaluated by Pearson's correlation tests, by which we can obtain biologically meaningful modules. The module and clinical feature that had the highest correlation were selected as the module of interest and clinical feature to be studied. In order to explore the potential mechanism of how module genes affect related clinical features, all genes of the module of interest were mapped into the DAVID database and subjected to GO functional and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. A *P*-value <0.01 and false discovery rate <0.01 were set as the cut-off criteria.[25]

Identifying hub genes and correlation analysis

To quantify module–trait associations, given that we had a summary profile (eigengene) for each module, we correlated each eigengene with external traits and looked for the most significant associations. This calculation was referred to as the module–trait relationship.[20] For intramodular analysis, we evaluated the gene significance (GS) and module membership (MM), the latter of which is also known as eigengene-based connectivity. The GS is the absolute value of the correlation between a specific gene and a trait. The MM is the correlation between the module eigengene and the gene expression profile. Using the GS and MM, we can identify genes that are significantly associated with clinical traits and important MM.[20] Node centrality has been shown to be useful to identify functionally critical genes; the node degree, number connections associated with a gene, was calculated and graphically visualized.

In addition, Search Tool for the Retrieval of Interacting Genes (STRING) is an online tool designed to evaluate protein–protein interaction (PPI) information.[26] To detect the potential relationship among the hub genes, we used the STRING application in Cytoscape and mapped the hub genes into STRING. A confidence score ≥0.4 was set as the cut-off criterion. In the PPI network, genes with a connectivity degree ≥8 were also defined as hub genes. The common hub genes in both the coexpression network and PPI network were regarded as "real" hub genes for further analysis.

Hub gene validation

The role of hub genes was validated by survival analysis in the transcriptional levels between CCA and normal samples from overall survival data of TCGA database. Moreover, Gene Expression Profiling Interactive Analysis (http://gepia.cancer-pku.cn) and SurvExpress (http://bioinformatica.mty.itesm.mx/SurvExpress) online tools to undertake validation of cancer-specific expression and prognosis of hub genes.[27,28]

Survival analyses were undertaken using log–rank tests and Kaplan–Meier survival curves. The SurvExpress tool divides samples into two groups (high-risk and low-risk) through the median of the prognostic index obtained by a Cox regression model. It then generates risk hazard ratios (HR), relative confidence intervals (CI), and *P*-values.

RESULTS

Gene coexpression network of CCA

USING THE WGCNA approach, we analyzed gene coexpression patterns based on mRNA expression profiles of CCA in TCGA database to identify key and candidate mRNAs that regulate the pathologic stage, histopathological identity, and pathology of fibrosis in the process of bile duct carcinogenesis. Generally, WGCNA

calculates correlations among genes that are analyzed across samples, and the correlations are weighted using a power function to determine the connection strengths between genes. Co-regulated genes are grouped into modules based on similarities in their expression patterns. Finally, each module is summarized, and hub genes are identified based on intramodular statistical analysis and node centrality properties, linking network topological features to biological information.

In TCGA, expression values of CCA were used to construct the coexpression network. The cluster analysis on these samples, undertaken with the flashClust package, are shown in Figure S1. After discarding the outlier samples, the soft threshold power value was set to 5 (Fig. S2), in accordance with the standard scale-free network distribution, with which adjacencies between all differential genes were calculated by power function. Then 28 gene coexpression modules, clustered in size from 55 to 1409 genes, were identified by hierarchical clustering and dynamic branch cutting (Fig. 1a). Each module was assigned a unique color as an identifier and is listed in Table S1. Interaction between the 28 coexpression modules was analyzed with the TOM among all genes (Fig. S3). Among these modules, two were merged into others because of similar MEs. Twenty-eight modules were finally generated (Fig. 1b). The gray module represented a gene set that was not assigned to any of the modules.

Modules with clinical significance

The clinic traits dataset was also downloaded from TCGA; some useless clinic traits information were removed in this research. To explore the clinical significance of each module, correlations between MEs and clinical traits, including clinical TNM staging, pathologic stage, histologic grade, Child–Pugh classification, and Ishak fibrosis classification, were analyzed. This observation is shown by the r-value of correlations shown in Figure 2. There were eight modules positively correlated with pathological stage and four modules positively correlated with histological grade, while there were only two modules positively correlated with Child–Pugh classification and one module correlated with Ishak fibrosis score. The highest association in the module–feature relationship was between the turquoise module and pathological stage III ($r = 0.95$, $P = 3 \times 10^{-18}$), and between the dark green module and pathological grade G1 ($r = 0.95$, $P = 3 \times 10^{-19}$), which were selected as modules of interest and clinical features to be studied in subsequent analyses. The second-highest association in the module–trait relationship was found between the orange module and pathological stage IVa ($r = 0.72$, $P = 7 \times 10^{-7}$), and between the turquoise module and

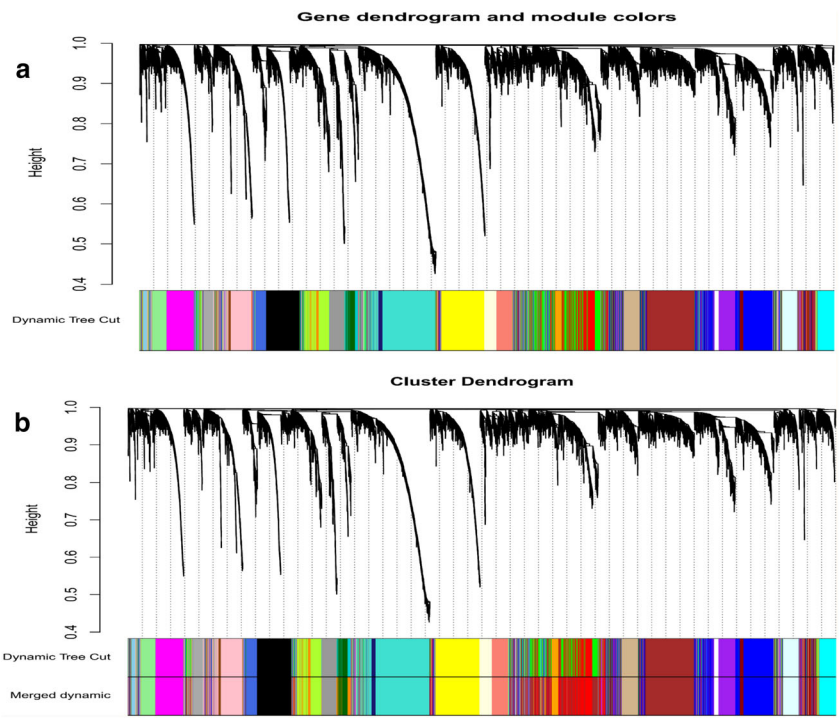

Figure 1 Cluster dendrogram generated by hierarchical clustering of genes involved in cholangiocarcinoma progression and prognosis, based on dissimilarity measure (topological overlap matrix) of genes. The branches correspond to modules of highly interconnected groups of genes. Two colored bars below the dendrogram represent the original modules and merged modules. Thirty modules were identified by the dynamic tree cutting method. Each module was assigned a color as an identifier. Twenty-eight modules were generated after merging according to the correlation of modules. [Color figure can be viewed at wileyonlinelibrary.com]

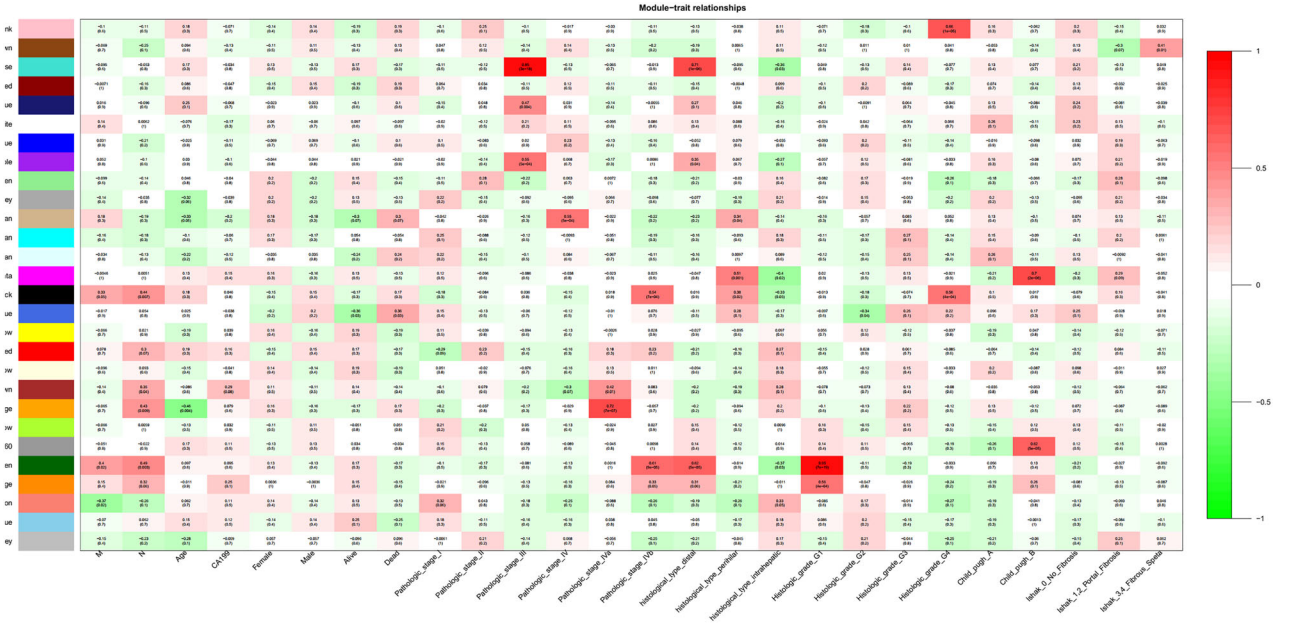

Figure 2 Module–trait relationships and *P*-values for selected traits in cholangiocarcinoma. Each row corresponds to a module eigengene, and each column corresponds to a clinical trait. Each cell contains a corresponding correlation and *P*-value. The table is color-coded by correlation according to the color legend. [Color figure can be viewed at wileyonlinelibrary.com]

tumor type of distal CCA ($r = 0.71$, $P = 1 \times 10^{-6}$). Among them, seven modules (pink, dark green, orange, tan, turquoise, magenta, and saddle brown) were positively correlated with histological grade G1 and G4, pathological stage III–IV, pathological type (distal), Child–Pugh classification, and Ishak fibrosis stage 3–4. The module–module (metamodule) relationship is the groups of correlated eigengenes with correlation of eigengenes >0.5. As shown in Figure 3, the eigengene dendrogram depicts the pink and saddle brown modules as highly related. The heatmap was used to identify groups of correlated eigengenes and the dendrogram indicates that the seven modules were significantly and positively associated with CCA clinic traits. Finally, we plotted a scatter plot of GS versus MM in these selected modules (Fig. S4). There is a highly significant correlation between GS and MM in these modules except for the tan and saddle brown modules.

Functional enrichment analysis of genes in meaningful modules

The biological significance of selected modules was investigated for in-depth understanding by GO term function analysis including biological process, cellular component, and molecular function, and KEGG pathway enrichment analysis. All genes in interesting modules were imported to DAVID software and STRING for online tool analysis. As shown in Table S2, the results illustrated that the dark green module were particularly enriched in biological processes, including tissue development and system development (data not shown), and were enriched in extracellular space and extracellular exosome in molecular function.

The biological processes of the magenta module were enriched in small molecule metabolic processes and single-organism metabolic processes; the molecular functions were enriched in catalytic activity, and cellular components were enriched in the cytoplasm and membrane-bounded organelles. The metabolic pathway plays an important role in KEGG pathway. The pink module was enriched in metabolic processes, including oxidation reduction and regulation of small molecular activation, in cellular components including extracellular vesicles and exosomes, and in molecular functions including endopeptidase activity and oxidoreductase activity. KEGG was enriched in metabolic pathways and complement and coagulation cascades. The biological processes of the tan module mainly regulated the metabolism of mitochondria, and were enriched in the mitochondrial membrane and mitochondrial parts. The turquoise module illustrated that these gene clusters were enriched

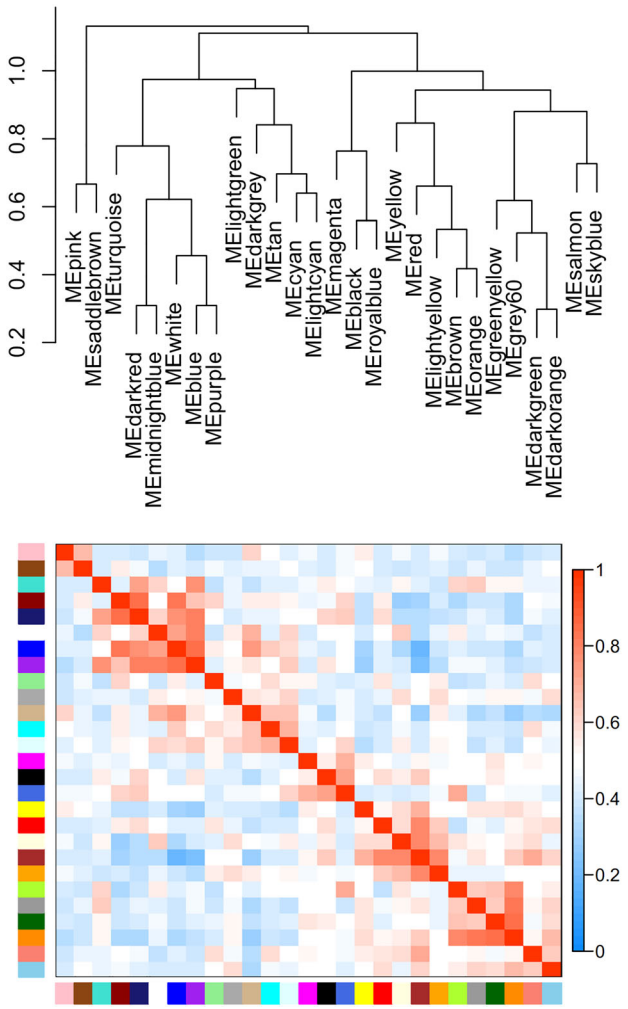

Figure 3 Eigengene network including dendrogram and heatmap shows the correlation among the module and clinical traits in cholangiocarcinoma. (a) Hierarchical clustering of module eigengenes indicates the branches of the dendrogram cluster together eigengenes that are positively correlated. Pink and saddle brown modules are highly related. (b) Heatmap plot of the adjacencies in the hub gene network. Red represents positive correlation with high adjacency; blue represents negative correlation with low adjacency. Squares of red color along the diagonal are the metamodule, positively correlated with clinical traits. [Color figure can be viewed at wileyonlinelibrary.com]

in immune system processes and immune response of biological processes that regulate leukocyte activation. These genes, which are associated with cytoplasm and cytosol components, also played molecular function roles in protein and enzyme binding. The pathway indicated by the turquoise module was associated with human T-lymphotropic virus-1 infection, allograft rejection, and cell adhesion molecules. However, the results of GO enrichment and KEGG analysis were not achievable in the orange module.

Identification of hub genes during tumorigenesis in candidate modules

The significance of hub genes with high MM value in a module was consistent with the significance of the module. These genes are also centers of the network and play essential roles in the network. In order to identify the central nodes that well represent these modules, we analyzed module genes with high intramodular connectivity in further detail. One hundred and fifty-three genes with high connectivity in the dark green, magenta, turquoise, and pink modules were identified as hub genes (Fig. 4, Table S3). In addition, under the threshold value of confidence >0.4 and connectivity degree of ≥ 10, the PPI network analysis showed 442 genes in the dark green, magenta, pink, tan, and turquoise modules individually. Finally, 40 common hub genes were identified in the coexpression network and PPI network, which were validated further for the process of CCA oncogenesis.

Survival analysis on hub genes

In order to validate whether these 40 common hub genes in the dark green, magenta, pink and turquoise modules were associated with the prognosis of CCA patients, and were regarded as potential prognostic biomarkers. Survival analysis, including overall survival (OS) time and disease-free survival time, were detected with GEPIA. Only two hub genes (*FERMT3* and *HCST*) with higher expression levels in CCA were screened out from the 40 common hub genes, but the two upregulated genes indicated good survival time of CCA patients. Next, we extensively and sequentially analyzed the other hub genes from PPI networks and coexpression networks. Finally, genes *MRPS18A* and *CST1* with higher expression level and *SCP2* with lower expression level compared to normal tissue reflected poor prognosis and survival time of CCA patients (Fig. 5); in contrast, although *POLE4*, *NDUFA2*, *COX6A1*, *HSP90B1*, *CTSD*, *SAR1A*, *TYROBP*, *RAP1A*, and *HCST* were closely correlated with prognosis, their expression level between CCA and normal tissue showed no significant difference. In addition, we found the differential expression of *FERMT3*, *LYN*, *CASP8*, *CDH1*, *TNFRSF18*, *MBNL1*, and *TNFRSF14* with better OS. To this end, after survival analysis of single genes, we undertook multigene survival analyses in three categories of hub genes, including differentially expression genes (DEG) with better prognosis (seven hub genes),

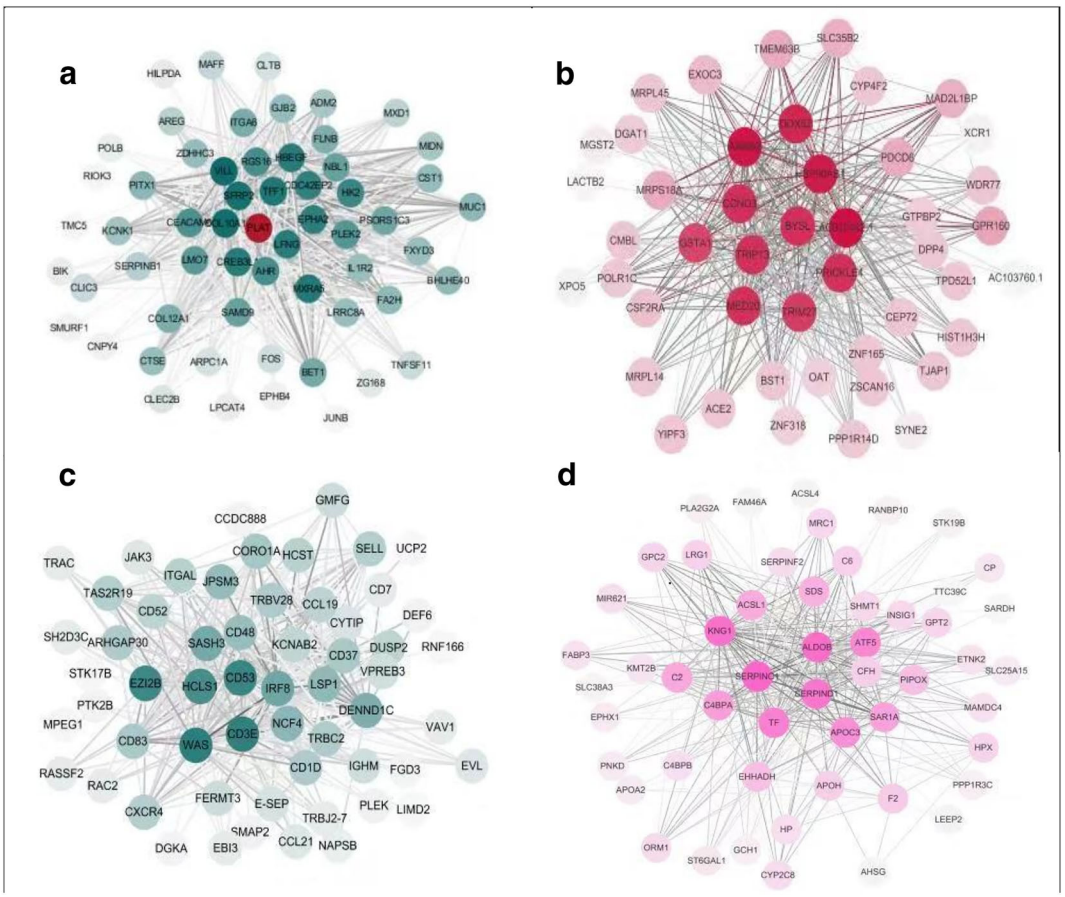

Figure 4 Coexpression network of most connected genes involved in cholangiocarcinoma in the modules (a) dark green, (b) magenta, (c) turquoise, and (d) pink. Nodes and edges represent the correlation of genes. The nodes with saturated circle represent network hub genes and the edge width was proportional to the score of weight by weighted gene coexpression network analysis. [Color figure can be viewed at wileyonlinelibrary.com]

DEG with poorer prognosis (three hub genes), and undifferentiated expression genes with better prognosis (nine hub genes). Specifically, we found that the OS times of the high-risk group of patients were more than twofold shorter than those of patients in the low-risk group (HR 51.43 [95% CI, 6.28–421.4], $P = 2.4e\text{-}04$ for all of 19 hub genes; HR 16.02 [95% CI, 3.43–74.94], $P = 4.2e\text{-}04$ for 16 hub genes except for *SCP2*, *MRPS18A*, and *CST1*) (Fig. 6).

DISCUSSION

THE PROGRESSION AND prognosis of CCA are quite variable in different patients. Although some molecular signatures involved in the process of CCA tumorigenesis have been discovered, specific and reliable biomarkers for CCA prognosis and progression are rarely reported. Accordingly, better and valuable biomarkers are urgently needed to provide more accurate clinical information that could significantly enhance decision-making for patient management. Here, we used WGCNA to screen progression- and prognosis-related biomarkers.

The WGCNA has many distinct advantages over other methods as the analysis focuses on the association between coexpression modules and clinic traits and the results had much higher reliability and biological significance.[29] Hotta *et al*. reported core gene networks and hub genes associated with progression of non-alcoholic fatty liver disease by RNA sequencing.[30] Through our study, by utilizing WGCNA, 28 modules were generated through 10 000 genes and 36 CCA samples globally, and relationships between module genes and clinical traits were constructed. By means of correlating gene modules with

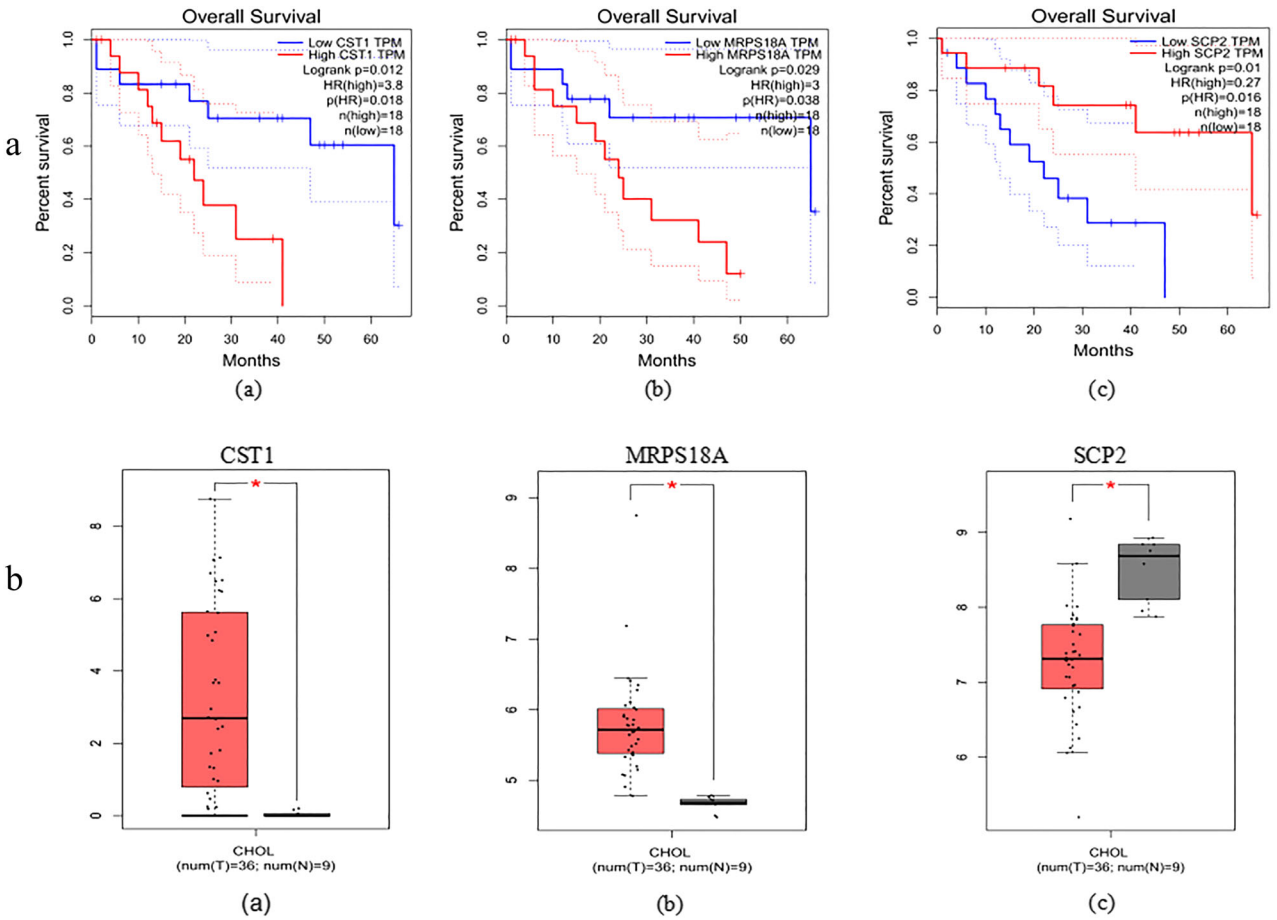

Figure 5 Validation of hub genes involved in cholangiocarcinoma (CCA) progression and survival. (a) Overall survival analysis of three hub genes in dark green, pink, and magenta modules individually (i) *CST1*, (ii) *MRPS18A*, and (iii) *SCP2*. Red line represents the samples with highly expressed genes; blue line represents samples with low gene expression. HR, hazard ratio. (b) Validation of gene expression levels between CCA samples and normal tissue. (i) *CST1*, (ii) *MRPS18A*, (iii) *SCP2*. [Color figure can be viewed at wileyonlinelibrary.com]

clinical features, seven modules positively correlated with CCA clinical traits were picked out. Among these modules, dark green, orange, tan, and turquoise were involved in pathological stage, pink and dark green were associated with histological grade, and magenta and dark green were correlated with liver function and aggressive location. The highest positive correlation was found in the turquoise module corresponding with pathological stage III and the dark green module corresponding with histological grade G1.

Function enrichment analyses of the dark green module illustrated that these genes played important roles in biological processes in tissue and system development, which are closely associated with tumor evolution, including aggressive depth, location, and malignancy. Module turquoise, corresponding to pathological grade III, was found to be mainly enriched in immune system processes, and enriched pathways of cell adhesion molecules. With the abnormal regulation of this module's genes, the dysfunction of immune cells facilitates tumor immune evasion and tissue infiltration of tumor cells; moreover, weakened cell adhesion could promote metastasis or movement of tumor cells. The biological processes of the tan module were mainly regulated by the metabolism of mitochondrion. The biological processes of the pink module were enriched in metabolic processes including oxidation reduction, and the magenta module was enriched in small molecule metabolic processes. Changes in cell metabolism are central to cancer development.[31] The mitochondria, which plays a central role in regulating

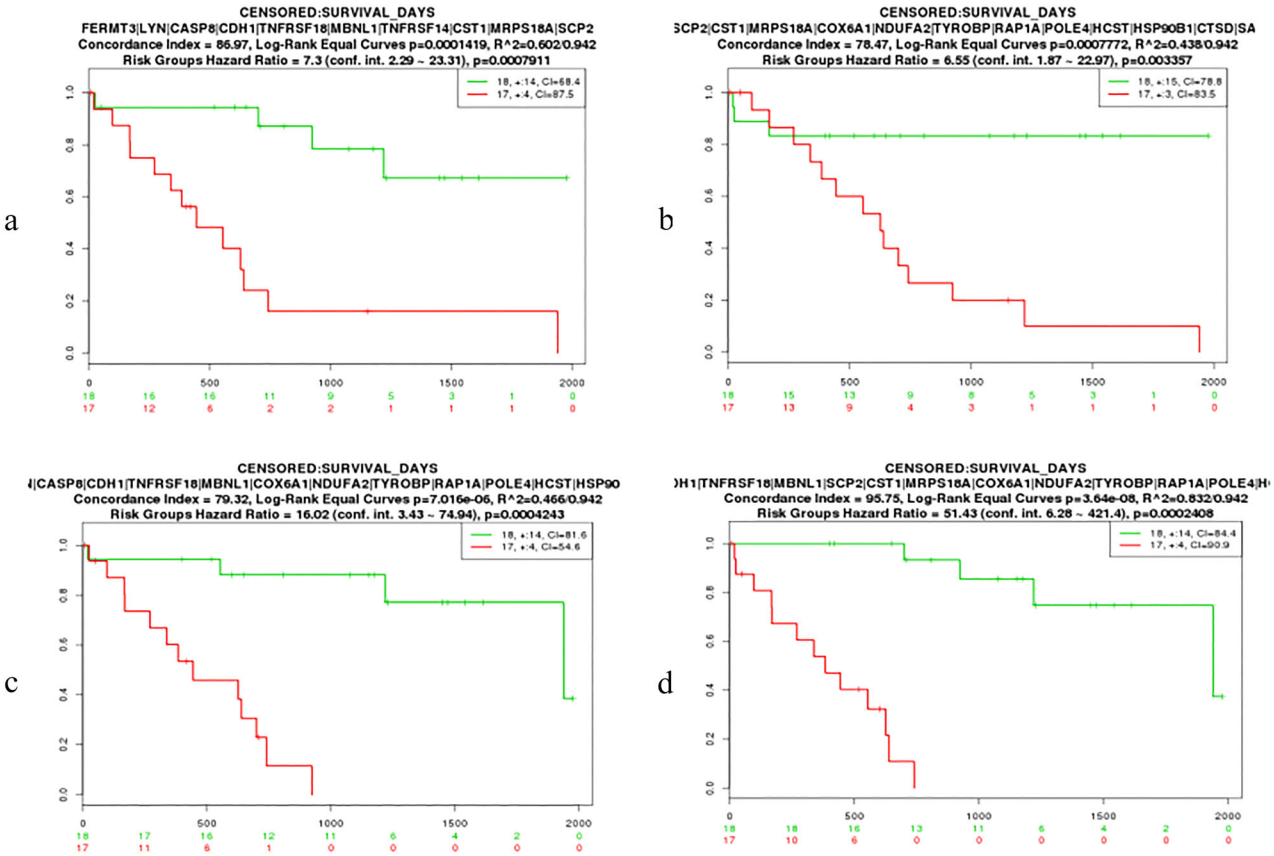

Figure 6 Kaplan–Meier survival plots for overall survival in patients with cholangiocarcinoma related to multiple hub genes. (a) Seven hub genes with differential upregulation and *SCP2*, *MRPS18A*, and *CST1*. (b) Nine hub genes with undifferentiated upregulation and *SCP2*, *MRPS18A*, and *CST1*. (c) Sixteen hub genes for differential expression and undifferentiated expression. (d) All 19 hub genes. *X* and *Y* axes represent survival time (months) and percent of survival, respectively. Lower curves represent high-risk groups and upper curves represent low-risk groups. [Color figure can be viewed at wileyonlinelibrary.com]

parameters of the metabolism such as energy production, production of biosynthetic precursors, redox status, reactive oxygen species (ROS) generation, cytosolic calcium levels, and the initiation of apoptosis, has been a fascinating focus of oncologic investigation, and somatic mitochondrial DNA (mtDNA) mutations have been identified in some solid tumors, and have been suggested as playing a critical role in carcinogenesis.[32] Mutations in mtDNA are enhanced by ROS generated by the oxidative phosphorylation pathway. In fact, Zhou *et al.*, using the MitoChip Affymetrix (commercially available GeneChip Human Mitochondrial Resequencing Array 2.0; Santa Clara, CA) with an oligonucleotide sequencing array, found a high incidence of mtDNA mutations in squamous cell carcinomas of the head and neck that might contribute to the development of a malignant phenotype by direct genotoxic effects from increased ROS production.[33]

Consequently, these modules were associated with higher pathological stage and histological grade.

Among these candidate modules, 40 common hub genes were derived from coexpression networks and PPI networks were extracted. Even though none of these common hub genes were identified as ideal prognosis biomarkers, we searched hub genes in two networks individually and thoroughly. Interestingly, we found numbers of upregulated hub genes with differential expression levels (*FERMT*, *LYN*, *CASP8*, *CDH1*, *TNFRSF18*, *MBNL1*, *TNFRSF14*, and *MXRA5*) and without differential expression (*COX6A1*, *NDUFA2*, *TYROBP*, *RAP1A*, *POLE4*, *HCST*, *HSP90B1*, *CTSD*, and *SAR1A*) showed good prognosis for CCA patients. Apparently, these hub genes could not be regarded as warning signatures for CCA prognosis. However, when these hub genes were taken together for survival analyses, the results were reversed. Multiple hub genes

within the high expression group were more than twofold shorter than the low expression group in terms of OS, which illustrated these hub genes could serve together as prognostic biomarkers.

The factors accounting for different prognostic results might include insufficient numbers of CCA samples, and the expression levels of hub genes tended to be unstable. Thus, we need to increase the CCA samples to validate the expression levels in further studies. Most importantly, we excluded 16 hub genes. Three hub genes (*MRPS18A*, *CST1*, and *SCP2*) attracted our attention as meaningful prognostic biomarkers.

The mitochondrial ribosomal proteins (MRPS18) coded by nuclear DNA were reported in early proteomic analyses, of which the s18 protein family were localized on the surface of the large subunit of the mitochondrial ribosome.[31] However, the function of these proteins is basically unknown. Recent studies reported that the expression of MRPS18A is upregulated in breast cancer cells compared to normal cells, but it cannot be regarded as a unique biomarker because it was found both in normal and cancer cells. However, the increased expression level in cancer cells is explained by the increased energy metabolism of cancer cells, and it could pave the way for new diagnostic and therapeutic routes to be explored.[34]

As a secretory protein encoded by the *CST1* gene, CST1 which belongs to the type 2 cystatin (CST) superfamily. Studies have implied that cystatins play pivotal roles in tumor invasion and metastasis. The salivary activity cystatins were associated with both local invasion at early stage and remote metastasis in colorectal cancer.[35,36] It is reported that CST1 was upregulated in cancerous lesions of gastric cancer tissue, which suggests its important role in the regulation of the proteolysis system and its effect on gastric tumorigenesis through T-cell factor-mediated proliferative signaling.[37] Previous studies implied that CST1 could contribute to the processes of carcinogenesis and tumor progression.

Sterol carrier protein 2 (SCP2), famous as non-specific lipid transfer protein, is a 13.2-kDa base protein expressed in peroxisome, mitochondria, endoplasmic reticulum, and cytoplasm.[38] The function of SCP2 is involved in the biosynthesis of cholesterol[39,40] and the transformation of cholesterol to bile acid.[41] Hence, SCP2 plays an essential role in cholesterol metabolism as a moderating factor. At present, there has been no direct research on the relationship between SCP2 and tumors. However, our study indicates that SCP2 might promote the development of CCA. Evidence given above illustrates that *SCP2* could be a novel oncogene and worth further study.

Taken together, *MRPS18A*, *CST1*, and *SCP2*, as candidate genes in CCA, were first found to be associated with CCA with tremendous excitement. The expression levels of these genes are involved in pathological stage, histological grade, Child–Pugh grade, and OS of patients, which illustrates that these hub genes participate in the development and progression of CCA. Thus, the candidate genes we identified can be taken as novel prognostic biomarkers or therapeutic targets of CCA and deserve further study. A large-scale study needs to be carried out to validate these findings. The outcomes of this study surely provide new insight into the tumorigenesis and progression of CCA.

ACKNOWLEDGMENTS

WE THANK THE Department of Infectious Diseases, The First Hospital of Lanzhou University. The project was supported by Award Number 1606RJZA127 from the Gansu National Science Fund.

REFERENCES

1 Nakeeb A, Pitt HA, Sohn TA *et al.* Cholangiocarcinoma. A spectrum of intrahepatic, perihilar, and distal tumors. *Ann Surg* 1996; **224**: 463–73.
2 Nakanuma Y, Sato Y, Harada K, Sasaki M, Xu J, Ikeda H. Pathological classification of intrahepatic cholangiocarcinoma based on a new concept. *World J Hepatol* 2010; **2**: 419–27.
3 Razumilava N, Gores GJ. Cholangiocarcinoma. *The Lancet* 2014; **383**: 2168–79.
4 Macias RIR, Banales JM, Sangro B *et al.* The search for novel diagnostic and prognostic biomarkers in cholangiocarcinoma. *Biochim Biophys Acta* 2018; **1864**: 1468–77.
5 Patel AH, Harnois DM, Klee GG, LaRusso NF, Gores GJ. The utility of CA 19-9 in the diagnoses of cholangiocarcinoma in patients without primary sclerosing cholangitis. *Am J Gastroenterol* 2000; **95**: 204–7.
6 Tshering G, Dorji PW, Chaijaroenkul W, Na-Bangchang K. Biomarkers for the diagnosis of cholangiocarcinoma: a systematic review. *Am J Trop Med Hyg* 2018; **98**: 1788–97.
7 Jiao Y, Pawlik TM, Anders RA *et al.* Exome sequencing identifies frequent inactivating mutations in *BAP1*, *ARID1A* and *PBRM1* in intrahepatic cholangiocarcinomas. *Nat Genet* 2013; **45**: 1470–3.
8 Zou S, Li J, Zhou H *et al.* Mutational landscape of intrahepatic cholangiocarcinoma. *Nat Commun* 2014; **5**: 5696.
9 Ikenoue T, Terakado Y, Zhu C *et al.* Establishment and analysis of a novel mouse line carrying a conditional knockin allele of a cancer-specific *FBXW7* mutation. *Sci Rep* 2018; **8**: 2021.
10 Zender S, Nickeleit I, Wuestefeld T *et al.* A critical role for Notch signaling in the formation of cholangiocellular carcinomas. *Cancer Cell* 2013; **23**: 784–95.

11 Zender S, Nickeleit I, Wuestefeld T *et al*. A critical role for Notch signaling in the formation of cholangiocellular carcinomas. *Cancer Cell* 2016; **30**: 353–6.
12 Sia D, Losic B, Moeini A *et al*. Massive parallel sequencing uncovers actionable *FGFR2-PPHLN1* fusion and *ARAF* mutations in intrahepatic cholangiocarcinoma. *Nat Commun* 2015; **6**: 6087.
13 Borger DR, Tanabe KK, Fan KC *et al*. Frequent mutation of isocitrate dehydrogenase (IDH)1 and IDH2 in cholangiocarcinoma identified through broad-based tumor genotyping. *Oncologist* 2012; **17**: 72–9.
14 Liu S, Xie F, Xiang X *et al*. Identification of differentially expressed genes, lncRNAs and miRNAs which are associated with tumor malignant phenotypes in hepatoblastoma patients. *Oncotarget* 2017; **8**: 97554–64.
15 Liang Z, Liu X, Zhang Q, Wang C, Zhao Y. Diagnostic value of microRNAs as biomarkers for cholangiocarcinoma. *Dig Liver Dis* 2016; **48**: 1227–32.
16 Zheng B, Jeong S, Zhu Y, Chen L, Xia Q. miRNA and lncRNA as biomarkers in cholangiocarcinoma (CCA). *Oncotarget* 2017; **8**: 100819–30.
17 Csosz E, Kallo G, Markus B, Deak E, Csutak A, Tozser J. Quantitative body fluid proteomics in medicine – a focus on minimal invasiveness. *J Proteomics* 2017; **153**: 30–43.
18 Shen J, Wang W, Wu J *et al*. Comparative proteomic profiling of human bile reveals SSP411 as a novel biomarker of cholangiocarcinoma. *PLoS One* 2012; **7**: e47476.
19 Lankisch TO, Metzger J, Negm AA *et al*. Bile proteomic profiles differentiate cholangiocarcinoma from primary sclerosing cholangitis and choledocholithiasis. *Hepatology* 2011; **53**: 875–84.
20 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf* 2008; **9**: 559.
21 Li J, Zhou D, Qiu W *et al*. Application of weighted gene co-expression network analysis for data from paired design. *Sci Rep* 2018; **8**: 622.
22 Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol* 2010; **17**: 1471–4.
23 Horvath S. *Weighted Network Analysis. Applications in Genomics and Systems Biology*. Los Angeles, US. Springer, 2011.
24 Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005; **4**: Article17.
25 Huang d W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.
26 Szklarczyk D, Franceschini A, Wyder S *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015; **43**: D447–D452.
27 Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res* 2017; **45**: W98–W102.
28 Aguirre-Gamboa R, Gomez-Rueda H, Martinez-Ledesma E *et al*. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One* 2013; **8**: e74250.
29 Chou WC, Cheng AL, Brotto M, Chuang CY. Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer. *BMC Genomics* 2014; **15**: 300.
30 †Hotta K, Kikuchi M, Kitamoto T *et al*. Identification of core gene networks and hub genes associated with progression of non-alcoholic fatty liver disease by RNA sequencing. *Hepatol Res* 2017; **47**: 1445–58.
31 Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011; **144**: 646–74.
32 Chatterjee A, Mambo E, Sidransky D. Mitochondrial DNA mutations in human cancer. *Oncogene* 2006; **25**: 4663–74.
33 Zhou S, Kachhap S, Sun W *et al*. Frequency and phenotypic implications of mitochondrial DNA mutations in human squamous cell cancers of the head and neck. *Proc Natl Acad Sci USA* 2007; **104**: 7540–5.
34 Sorensen KM, Meldgaard T, Melchjorsen CJ *et al*. Upregulation of Mrps18a in breast cancer identified by selecting phage antibody libraries on breast tissue sections. *BMC Cancer*. 2017; **17**: 19.
35 Hirai K, Yokoyama M, Asano G, Tanaka S. Expression of cathepsin B and cystatin C in human colorectal cancer. *Hum Pathol*. 1999; **30**: 680–6.
36 Saleh Y, Sebzda T, Warwas M, Kopec W, Ziolkowska J, Siewinski M. Expression of cystatin C in clinical human colorectal cancer tissues. *J Exp Ther Oncol* 2005; **5**: 49–53.
37 Choi EH, Kim JT, Kim JH *et al*. Upregulation of the cysteine protease inhibitor, cystatin SN, contributes to cell proliferation and cathepsin inhibition in gastric cancer. *Clin Chim Acta* 2009; **406**: 45–51.
38 Wirtz KW. Phospholipid transfer proteins revisited. *Biochem J* 1997; **324**(Pt 2): 353–60.
39 Amigo L, Zanlungo S, Miquel JF *et al*. Hepatic overexpression of sterol carrier protein-2 inhibits VLDL production and reciprocally enhances biliary lipid secretion. *J Lipid Res* 2003; **44**: 399–407.
40 Mukherji M, Kershaw NJ, Schofield CJ, Wierzbicki AS, Lloyd MD. Utilization of sterol carrier protein-2 by phytanoyl-CoA 2-hydroxylase in the peroxisomal alpha oxidation of phytanic acid. *Chem Biol* 2002; **9**: 597–605.
41 Lidstrom-Olsson B, Wikvall K. The role of sterol carrier protein2 and other hepatic lipid-binding proteins in bile-acid biosynthesis. *Biochem J* 1986; **238**: 879–84.

SUPPORTING INFORMATION

ADDITIONAL SUPPORTING INFORMATION may be found online in the Supporting Information section at the end of the article.

Figure S1 Clustering dendrogram of tumor samples and clinical traits. The color intensity is proportional to the clinical feature.

Figure S2 Analysis of network topology for various softthresholding powers. (a) Analysis of the scale-free fit index for various soft-thresholding powers (β). (b) Analysis of the mean connectivity for various soft-thresholding powers.

Figure S3 Network heatmap plot. The heatmap shows the topological overlap matrix (TOM) among selected genes in the analysis. Branches in the hierarchical clustering dendrograms correspond to each module. Two bars of color coded module membership are located under and right of the dendrograms. Light color shows low overlap and progressively saturated yellow and red color represents higher overlap. Genes of high intramodular connectivity are located at the tip of the module branches because they show the highest interconnectedness with the rest of the genes in the module.

Figure S4 (a) Scatterplots of Gene Significance (GS) for pathological stage III–IV versus Module Membership (MM) in the turquoise, orange, and dark green modules. (b) Scatterplots of GS for histological distal type versus MM in the dark green module. (c) Scatterplots of GS for histological grade G1 and G4 versus MM in the dark green and pink modules. (d) Scatterplots of GS for liver fuction classification Child–Pugh B versus MM in the magenta. There is a highly significant correlation between GS and MM in these modules.

Table S1 Number of genes in modules

Table S2 Gene Ontology enrichment analysis and the Kyoto Encyclopedia of Genes and Genomes pathway of coexpression module genes

Table S3 Hub genes identified in interesting modules