Observational Study

Medicine

OPEN

Identification of signature of gene expression in biliary atresia using weighted gene co-expression network analysis

Yongliang Wang, MS[a], Hongtao Yuan, BA[a,*], Maojun Zhao, MS[b], Li Fang, MD[c]

Abstract

Biliary atresia (BA) is the most common cause of obstructive jaundice during the neonatal period. This study aimed to identify gene expression signature in BA. The datasets were obtained from the Gene Expression Omnibus database. Weighted gene co-expression network analysis identified a critical module associated with BA, whereas Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis revealed the functions of the essential modules. The high-connectivity genes in the most relevant module constructed protein–protein interaction networks via the string website and Cytoscape software. Hub genes screened by lasso regression consisted of a disease classification model using the randomforest method. Receiver operating characteristic curves were used to assess models' sensitivity and specificity and the model was verified using the internal and external validation sets. Ten gene modules were constructed by WGCNA, of which the brown module had a strong positive correlation with BA, comprising 443 genes. Functional enrichment analysis revealed that module genes were mainly involved in biological processes, such as extracellular matrix organization, cell adhesion, inflammatory response, and the Notch pathway ($P < .001$), whereas these genes were involved in the metabolic pathways and cell adhesion molecules ($P < .001$). Thirty-nine high-connectivity genes in the brown module constructed protein-protein interaction networks. keratin 7 (*KRT7*) and C-X-C motif chemokine ligand 8 (*CXCL8*) were used to construct a diagnostic model that had an accuracy of 93.6% and the area under the receiver operating curves for the model was 0.93. The study provided insight into the signature of gene expression and possible pathogenesis of BA; furthermore, it identified that the combination of *KRT7* and *CXCL8* could be a potential diagnostic model for BA.

Abbreviations: AUC = area under curves, BA = biliary atresia, *CXCL8* = C-X-C motif chemokine ligand 8, DAVID = annotated, visualized, and integrated Discovery Database, GEO = Gene Expression Omnibus, GO = Gene Ontology, GS = genes traits significance, KEGG = Kyoto Encyclopedia of Genes and Genomes, *KRT7* = = keratin 7, ME = module eigengene, MM = module membership, NC = normal control group, Non-BA = control group for hepatobiliary diseases without the biliary atresia, PPI = protein–protein interaction, ROC = receiver operating characteristic, WGCNA = Weighted Gene Go-expression Network Analysis.

Keywords: biliary atresia, computational biology, statistical model

1. Introduction

Biliary atresia (BA) is one of the most severe diseases of the hepatobiliary system in infancy and is the most common cause of liver transplantation in children.[1] BA also is the most common cause of neonatal cholestasis (25%–55%).[2] The pathological changes include bile duct hyperplasia, cell infiltration, portal fibrosis,[3] and the absence of sinusoidal fibrosis, culminating in cirrhosis. The early operation, including Kassai and its variants, is key to a better prognosis; thus, early diagnosis of BA is crucial.[4] Diagnosis of BA was screened by clinical manifestation, laboratory examination, and imaging examination, confirmed by liver biopsy and intraoperative cholangiography.[5] Nevertheless, as an invasive operation, the appliance of intraoperative cholangiography is limited. Despite the high diagnostic accuracy of liver biopsy for BA,[6] some hepatobiliary diseases have histological features that overlap with BA,[2] including MDR3 (multidrug resistance protein 3) deficiency disease, cystic fibrosis, doublecortin domain containing 2 disease, alpha1-antitrypsin deficiency, and parenteral nutrition-associated cholestasis, with their histology features involving duct proliferation, portal tract fibrosis, inflammation, bile plugs.[7–11] In the early stage of disease, the classic histological changes of BA might be atypical, resulting

The authors have no funding and conflicts of interest to disclose.

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

Supplemental Digital Content is available for this article.

[a] Hepatological Surgery Department, The First People's Hospital of Guiyang City, Guizhou Province, China, [b] Emergency Department, The First People's Hospital of Guiyang City, Guizhou Province, China, [c] Department of Critical Care Medicine, The First People's Hospital of Guiyang City, Guizhou Province, China.

**Correspondence: Hongtao Yuan, Hepatological Surgery Department, The NO.1 People's Hospital of Guiyang City, Guizhou Province, China (e-mail: wylwyflxy@gmail.com.*

How to cite this article: Wang Y, Yuan H, Zhao M, Fang L. Identification of signature of gene expression in biliary atresia using weighted gene co-expression network analysis. Medicine 2022;101:37(e30232).

Received: 3 February 2022 / Received in final form: 9 July 2022 / Accepted: 12 July 2022

http://dx.doi.org/10.1097/MD.0000000000030232

in false-negative diagnoses.[5,12] So a series of liver biopsies are necessary.[13] Clinical features, laboratory parameters, and genetic testing are essential to arrive at a correct diagnosis, as distinguishing early BA from the above disorders by histological features is a challenge.[14] As a complementary method, kinds of molecular markers, such as interleukin-33, matrix metallopeptidase 7, interleukin -8, and microRNAs, have been shown to be diagnostically effective in BA.[15–18] Therefore, we hope to exploit varied analytic tactics to mine more potential molecular biomarkers of BA from existing data, supplement current diagnostic tools, and improve the diagnostic accuracy of BA. Weighted gene co-expression network analysis (WGCNA) is a method to analyze the complicated relationship between gene and phenotype, which has been utilized in various research of biological contexts[19] and can instruct the association of the module with disease results.[20] A unique advantage of the WGCNA is that it retains the continuous nature of the underlying correlation information for construction of a network based on soft thresholding of the correlation coefficient, compared to the unweighted network requiring the choice of a hard threshold.[19,21] The data, divided into multiple groups, requires repeated pairwise comparisons and multiple hypothesis tests when performing the differentially expressed genes (DEG) analysis, whereas, unlike DEG analysis, WGCNA directly modularization relationship between gene expression and phenotype, reducing computational effort. Therefore, we utilized this method to analyze expression profiles of BA in the Gene Expression Omnibus (GEO) database and to discover the genetic signature of BA.

2. Materials and Methods

2.1. Materials

The workflow was shown in (Supplementary Digital Content 1, http://links.lww.com/MD/H91). The mRNA profiles of GSE46960 (85 liver samples from age-matched infants and 10 liver samples from adults) and GSE84954 (11 liver, 13 fat, and 13 muscle samples from children) were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/gds/?term=). All adults and nonliver samples were excluded. According to the known diagnosis, eligible samples were grouped as BA group (biliary atresia group), Non-BA group (control group for hepatobiliary diseases without the BA), and normal control group (NC group). GSE46960 contained 64 BA samples; 14 non-BA samples; 7 NC samples. In GSE84954, 11 liver samples were picked out and divided into BA group (6 samples) and non-BA group (5 samples). Samples' detail were recorded in (Supplementary Digital Content 2, http://links.lww.com/MD/H92).

Data were analyzed and plotted using R 4.12 and R Studio 1.4.1717 software. Additionally, these R-packages,"oligo,"[22] "WGCNA,"[20] "glmnet,"[23] "pROC,"[24] "randomForest,"[25] "caret,"[26] "dplyr,"[27] "kknn"[28] were used for statistical analysis. The following R-packages were used to plot: "ggplot2,"[29] "Pheatmap,"[30] "treemap,"[31] "simplifyEnrichment,"[32] "corplot."[33] String website (https://string-db.org) and Cytoscape software (v3.8.2) to construct the protein–protein interaction network. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis were by the database annotation, visualized, and integrated discovery (annotated, visualized, and integrated Discovery Database [DAVID], https://david.ncifcrf.gov/).

2.2. Data processing

Normalization could eliminate the variations in expression (Intensity) caused by experimental techniques, and keep the data at the same level for each sample and parallel experiments. Downloaded microarray data were read and proceeded, including robust multichip average background correction, quantile normalization, base 2 logarithmic conversions, and probe ID transformation. The data without corresponding gene symbol or duplicate gene symbols were removed. The genes with low expression values and missing values were then filtered out. Relative expression plots of datasets are shown in (Supplementary Digital Content 3, http://links.lww.com/MD/H93).

2.3. Construction of co-expression network

The genes with the top 5000 absolute median deviations in GSE46960 have been used to construct a co-expression network via the "WGCNA" R package. First, cluster analysis was performed on the samples using the class average method to remove the outliers. According to sample cluster analysis, there were no outliers in GSE46960. Next, an appropriate soft threshold power was calculated based on the scale-free topology criterion, with the optimal soft threshold power of 12 and the corresponding scale-free R^2 of 0.9. The co-expression network was constructed based on the soft threshold power, and then the cluster dendrogram of gene modules was plotted after hierarchical clustering and branching cuts. The minimum size of the module was set to 30 and CutHeight = 0.25 and the above parameters were used to merge the colser modules into new modules.

2.4. Identification of significant module and high-connectivity genes

The WGCNA package reckoned each sample's module eigengene (ME) matrix, figured the correlation matrix and correlation *P* value between ME and clinical traits, plotted the "module-trait relationship" heatmap, and selected the module with the strongest correlation to the group BA. The verboseScatterplot of the engaging module was depicted based on the Module Membership (MM) value of genes and genes traits significance (GS). Following this, genes with high MM and GS were used to construct protein–protein interaction networks on the String website (https://string-db.org). The connectivity degree and combination score of genes were entered into Cytoscape software to output the protein-protein interaction (PPI) network plot.

2.5. Detection of module function

Enrichment analyses of GO and KEGG pathways for module genes were performed on the DAVID website, and a *P* value of < 0.05 for GO terms or KEGG terms was considered statistically significant.

2.6. Compressing variates

The expression heatmap of genes in the PPI network was presented. The number of variates needs to be compressed, as there were still too many variates that can build diagnostic models. Lasso regression was selected to compress variates. Since the NC group could be easily distinguished from the BA group by laboratory examination and clinical manifests, the NC group was no longer included in the subsequent analysis. Sixty percent of the remaining were set as a training set and 40% as the test. The training set was used to build a lasso regression model. The minimum λ value of the training set was determined by tenfold cross-validation and the minimum λ was brought into the test set to work out variates' coefficients. Those genes retaining coefficients were viewed as nominated genes to construct a diagnostic model by randomforest; meanwhile, the correlation matrix of genes in the PPI network was plotted.

2.7. Construction of the diagnostic model

Various randomforest models were constructed by different collocations of the above-nominated genes, and their diagnostic accuracy was calculated respectively. As for the accuracy of model = 100-OOB (Out of Bag or estimated error rate), the best model was identified by comparing the parameters of models, the receiver operating characteristic (ROC) curve of the models was plotted to assess the diagnostic ability of the model, and the area under the curve of models was ciphered. The principal component analysis (PCA) of original data was compared with PCA of diagnostic model to evaluate the classification effectiveness of the model.

2.8. Validation for the model

In GSE46960, the expression level of model genes in the BA was contrasted with other groups to verify the difference in model genes between the BA and other groups. Tukey honestly significant difference (Honestly Significant Difference) was executed on the outcomes; $P < .05$ was considered statistically significant. For testing the diagnostic ability of the model in external data, the model predicted the grouping results of data in GSE84954 utilizing randomforest and k-nearest neighbors algorithm (KNN). Ultimately, the accuracies were calculated.

3. Results

3.1. Construction of co-expression network & identification of significant module

A systematic clustering tree of gene modules was illustrated in Figure 1A, including 10 gene modules. The label heatmap "Module-trait relationship" showed the correlation of module with clinical traits (Fig. 1B). The correlations contained positive

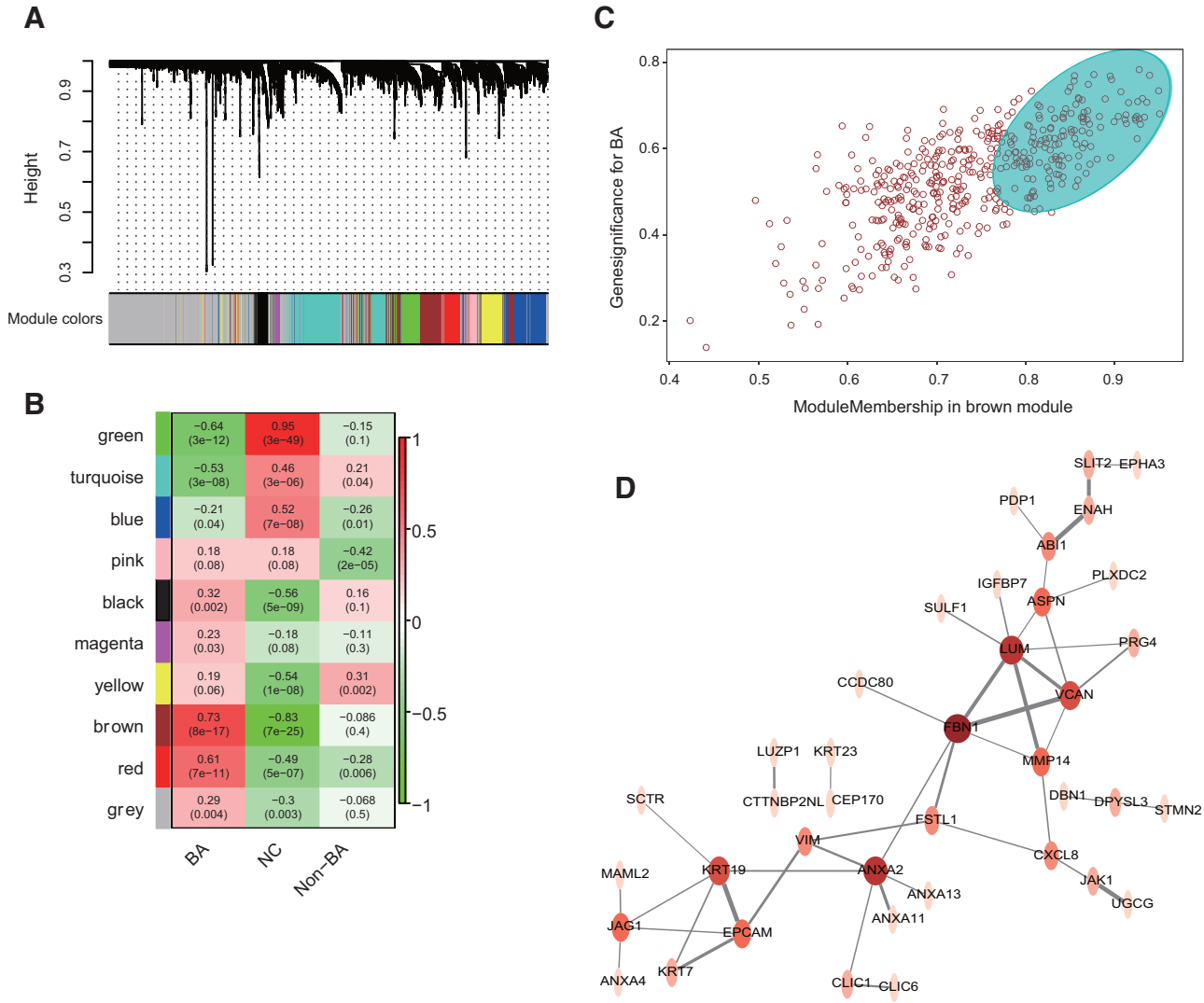

Figure 1. (A) Cluster dendrogram. The upper interface was a dendrogram of genes, and the lower was the corresponding cluster of genes. Varied colors typified various modules; a total of 10 modules were identified. The gray module was a collection of genes with no distinctive features. (B) Module-Trait relationship. Each column conformed to a clinical phenotype, and each row fitted to a module. The numbers in each grid represented the correlation coefficients between the module and the clinical trait. The numbers in brackets were parallel *P* value of correlation coefficients. The depth of color with grid symbolized the strength of the association between a module and a clinical phenotype; red denoted a positive association, and green represented a negative association. (C) VerboseScatterplot of the brown module. The X-axis was the MM, Y-axis was the gene significance of BA. The points in the range of the blue ellipse were genes with high connectivity. (D) Protein–protein interaction network of hub genes (PPI network). The color depth and width of the ellipse indicated the connectivity degree of a gene; the width of edge represented the combination score that reflects degree of link during genes. BA = biliary atresia, GS = genes traits significance, MM = module membership, NC = normal control group, Non-BA = control group for hepatobiliary diseases without the biliary atresia, PPI = protein–protein interaction.

and negative correlations, with color depth representing the strength of the correlation. The brown module (cor = 0.73, $P = 8 \times 10^{-17}$) had the highest positive correlation with BA among these modules. VerboseScatterplot (Fig. 1C) corresponded to the MM and GS values of genes in the brown module. In addition, the 84 genes (Supplementary Digital Content 4, http://links.lww.com/MD/H94) in the upper right quadrant of the verbosescatterplot (MM > 0.8, GS > 0.6) formed the PPI network, which is shown in Figure 1D based on connectivity degree and combination score.

3.2. Function enrichment analysis of significant module

Cellular Component enrichment analysis showed that genes mainly existed in the cytoplasm, extracellular exosome, and cytosol (Fig. 2A). The main molecular functions referred to protein binding, transcription factor activity, DNA binding, and calcium ions binding (Fig. 2B). The biological processes involved include the Notch signaling pathway, signal transduction, positive regulation of transcription from RNA polymerase II promoter, and inflammatory response. (Fig. 2C). KEGG pathway enrichment analysis showed that genes in the module were mainly involved in the metabolic pathways, axon guidance, and cell adhesion molecules (CAMs; Fig. 2D).

3.3. Compressing variates

Figure 3A exhibited the expression heatmap of 39 genes in the PPI networks. Thirty-nine genes were clearly expressed differently between the BA and NC groups, in contrast, a few genes were significantly differentially expressed between the BA and the non-BA groups. According to the result of lasso regression, 5 variates still owned coefficients, when minimal $\lambda = 0.0497$ (Table 1). The correlation matrix of genes in the PPI network (Fig. 3B) demonstrated the robust correlation between the 5 genes and other genes.

3.4. Construction of the diagnostic model

Model constituted of keratin 7 (*KRT7*), *KRT19*, Versican (*VCAN*), Annexin A2 (*ANXA2*), and C-X-C motif chemokine ligand 8 (*CXCL8*) possessed the highest classification accuracy of 94.5% (Table 2, Supplementary Digital Content 5, http://links.lww.com/MD/H95); moreover, the model only containing *CXCL8* and *KRT7* owned a similar accuracy of 93.6%, which was as high as the above-stated model. *CXCL8* and *KRT7* for the BA group had a classification error rate of <2%, while the non-BA group had a classification error rate of 29%. Figure 4 showed the ROC and area under curve (AUC) for the 2 models; The AUC equated to 0.93 and 0.92. Finally, the model consisting

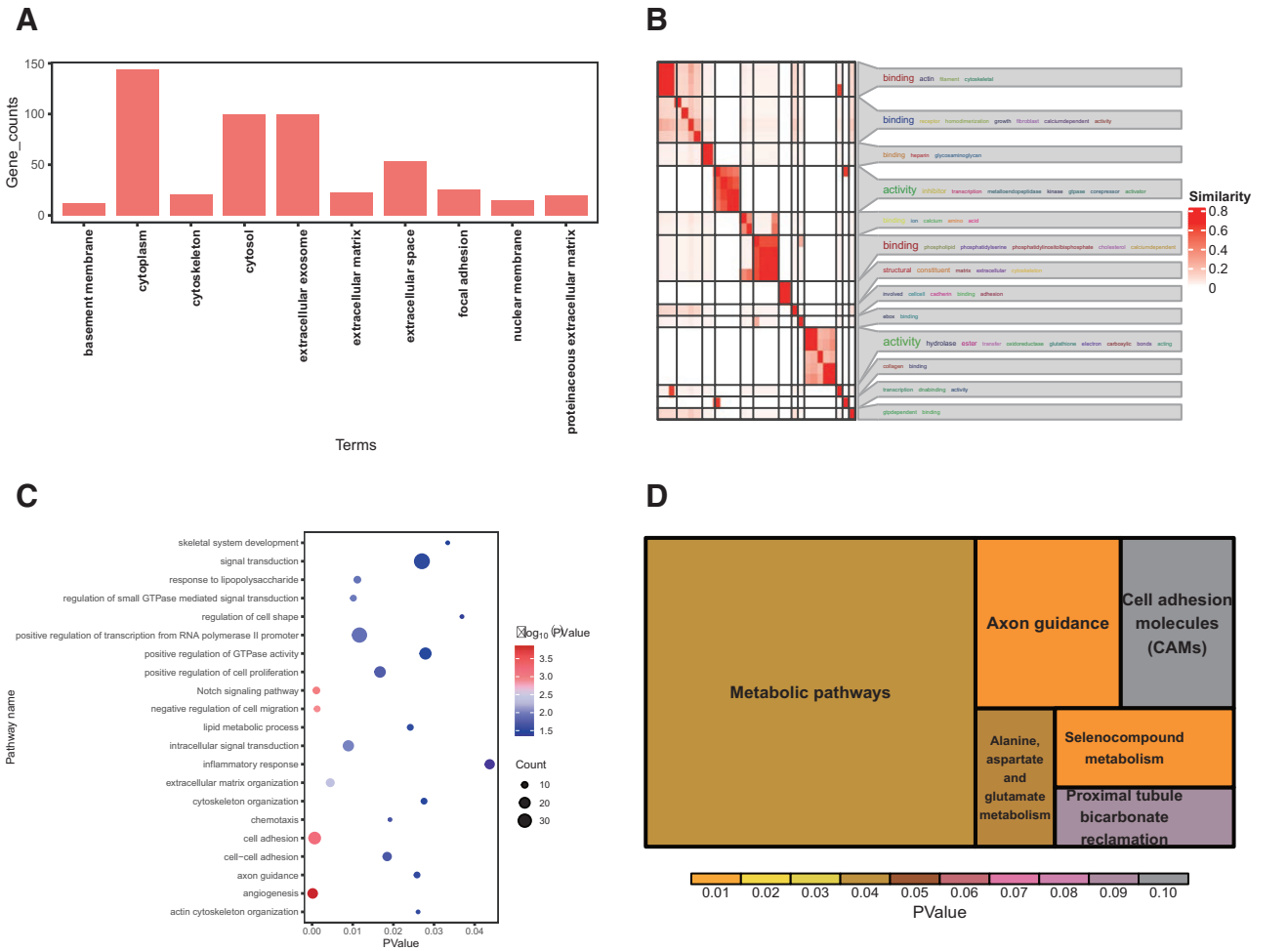

Figure 2. (A) GO for CC. The X-axis denoted GO TERM, and the Y-axis represented gene counts of GO TERM. (B) GO for MF enrichment analysis. Each column or row was a BP term, and the color depth symbolized the similarity between terms, while terms with high similarity formed a cluster, and the font size of the word cloud next to the heatmap was positively proportional to the number of terms. (C) GO for BP enrichment analysis. The bubble size represented the gene counts of GO TERM, and the color deepness coincided with the *P* value. (D) Treemap of KEGG pathway analysis. Box size conformed to gene counts in a pathway and the color depth matched up *P* value. BP = biological processes, CAM = cell adhesion molecule, CC = cellular component, GO = gene ontology, MF = molecule function.

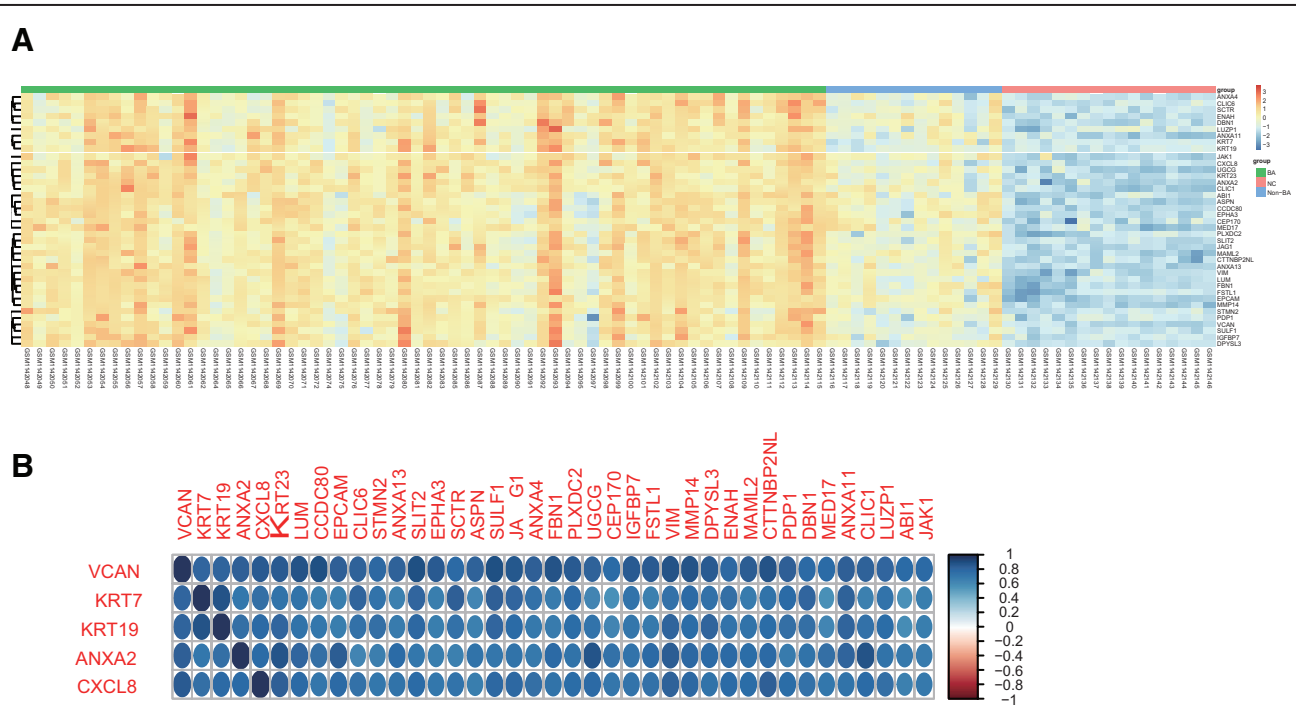

Figure 3. (A) Heatmap of the hub genes in the PPI network. (B) Correlation matrix of the hub genes in the PPI network. The deeper color indicated a stronger correlation. BA = biliary atresia, NC = normal control group, Non-BA = control group for hepatobiliary diseases without the biliary atresia, PPI = protein–protein interaction.

of *CXCL8* and *KRT7* was selected as best model, because it had a high accuracy and minor number of genes. Figure 5A and B exhibited that the PCA of the model could better separate BA and Non-BA, compared to the PCA of primitive data.

3.5. Validation for the model

In GSE46960, the expression level of *KRT7* in the BA group was significantly higher than in other groups (F = 42.65, BA vs NC $P < .001$; BA vs non-BA $P < .001$), and the expression level of *CXCL8* in the BA group also was significantly higher than other groups (F = 74.74, BA vs NC $P < .001$; BA vs non-BA $P < .001$) (Fig. 6A). In the data set GSE84954, the classification accuracy of the model utilizing randomforest was 73% (95% confidence interval: 0.39–0.94). The result of classification is predicted as shown in Table 3. Figure 6B revealed the probability of each sample being classified as BA calculated by the randomforest and the expression level of *KRT7* and *CXCL8* in the corresponding sample. A sample with consistently high expression of *KRT7* and *CXCL8* has a high probability of being identified as the BA. GSM2254637 and GSM2254649 were misclassified as the BA, whereas GSM2254643 was incorrectly assigned to non-BA. KNN algorithm (3 folds crossing validation) had better results, with the mean accuracy rising slightly around 0.83 (Table 4). Comparing randomforest, similarly, GSM2254637 and GSM2254649 were also misclassified as BA.

4. Discussion

The etiology and mechanism of BA are still obscure, and there are many hypotheses for its pathogenesis, including abnormal bile duct development, inflammation, genetic variants, and immune disorder caused by a viral infection and even the occurrence of BA may be an end-stage phenotype induced by multiple mechanisms.[34]

The study found that the brown module genes most closely associated with BA enriched in *Cell adhesion molecule* (*CAM*), extracellular matrix (ECM) organization, inflammatory response, and notch pathway. One concept is that biliary epithelium cells actively participate in the pathogenesis of cholangiopathies via their transformation into reactive cholangiocytes; they have a critical role in biliary fibrosis via crosstalk with ECM-producing cells, inflammatory cells, and ECM, and also promote fibrosis via the secreting of proinflammatory or chemotactic cytokines and the expression of adhesion molecules.[35] Laminin Subunit Gamma 1 (*LAMC1*) is an ECM glycoprotein that participated in many processes, such as cell adhesion, and its location on the base membrane and the positive association between *LAMC1* and liver fibrosis was discovered long ago.[36] Integrins (*ITG*), one of cellular receptors of the LAMC family, bind to *LAMC* and transmit base membrane signals to cells.[37] *LAMC* binds to *ITGA/B* and formats the focal adhesion pathway in which Focal adhesion kinase regulates downstream hedgehog pathway in response to injury of liver or biliary, fibrosis.[38] Yu et al[39] of Nanjing Medical University proved expression of *LAMC1* was modulated via *rs3768617* of *LAMC1*, thereby affecting the binding of *miRNA-548b-3p* to *LAMC1*, which could be a potential therapeutic target for BA. CAMs like Intercellular Adhesion Molecule 1 (*ICAM1*), Vascular Cell Adhesion Molecule 1 (*VCAM1*), and E-selectin are overexpressed in patients' liver tissue and serum with BA.[40,41] When the injury occurred, CAM mediated cell-cell contact and recruited leukocytes to the site of injury, followed by sustained inflammation.[42] A process that also existed in the biliary epithelium was infected by the virus and led to BA.[5] The activity of CAM might reflect inflammation of the biliary ducts and the development of cirrhosis.[41] During follow-up to the BA patients after operation, Professor Shan Zheng's team at Fudan University found that patients with their jaundice not eliminated possessed high expression levels of *VCAM1*, suggesting that *VCAM1* may play an important role in the pathogenesis of liver fibrosis in infants with BA.[43] The same phenomenon was detected in patients with primary biliary cirrhosis, primary sclerosing cholangitis, and alcoholic liver disease.[44] In addition, the excessively accumulated extracellular matrix represents the process of liver fibrogenesis; extracellular affords complex functions, such as cell adhesion, cell migration, and proliferation.[45,46]

Table 1

Lasso regression coefficients of 39 hub genes.

Symbol	Coefficient
VCAN	−0.84
KRT23	
LUM	
CCDC80	
EPCAM	
CLIC6	
STMN2	
ANXA13	
CXCL8	−0.10
SLIT2	
KRT7	−1.04
KRT19	−0.10
EPHA3	
SCTR	
ASPN	
SULF1	
JAG1	
ANXA4	
FBN1	
PLXDC2	
UGCG	
CEP170	
IGFBP7	
FSTL1	
VIM	
MMP14	
DPYSL3	
ENAH	
MAML2	
CTTNBP2NL	
PDP1	
DBN1	
ANXA2	−2.50
MED17	
ANXA11	
CLIC1	
LUZP1	
ABI1	
JAK1	

These pathological changes match the progress of BA. The notch pathway is an evolutionarily conserved intercellular signaling pathway to maintain progenitor cells.[47] It is a regulator in the shape of the intrahepatic bile duct.[48] The abnormal alterations of the notch pathway and its ligand Jagged Canonical Notch Ligand 1 (*JAG1*) are the etiology of Alagille syndrome characterized by a paucity of intrahepatic bile ducts.[49] The notch pathway plays a key role in BA, promoting the differentiation of hepatic progenitor cells into cholangiocytes for repairing bile duct epithelium.[50–52] *KRT7* and *KRT19* were reported as markers of biliary epithelium, which appeared on biliary progenitor cells during development and regeneration,[53,54] a process, however, led to the obstruction of the bile ducts.

The PPI network exhibited that hub genes of brown module principally converged to 3 centers: Epithelial Cell Adhesion Molecule (*EPCAM*) and *JAG1*, *ANXA2*, Fibrillin 1(*FBN1*), which were associated with notch pathway, annexins family, and elastin respectively. Notch pathway had been described in the above statement, likewise, annexins family and elastin were also in charge of the BA process. Annexins were involved in tissue regeneration as these genes are involved in cell-to-cell communication and extracellular matrix growth.[55] A hallmark of liver cirrhosis was the accumulation of large amounts of elastic fiber.[56] *FBN1*, usually together with elastin, was a component of the extracellular matrix[57] and was regularly co-expressed with elastin in children with cholestatic diseases.[58] *FBN1* in cholestatic disease was characterized as an apparent high expression with fibrosis of the bile duct, such as BA, sclerosing cholangitis, *FBN1* cross-linked with tropoelastin to format microfibrils, further formed elastic fiber with elastin.[59] *TGF-beta* is one of the most crucial cytokines regulating elastin expression levels and is perceived as the most potent fibrogenic cytokine in the liver; *FBN1* could bind *TGF-beta* to modulate levels of *TGF-beta*.[60,61]

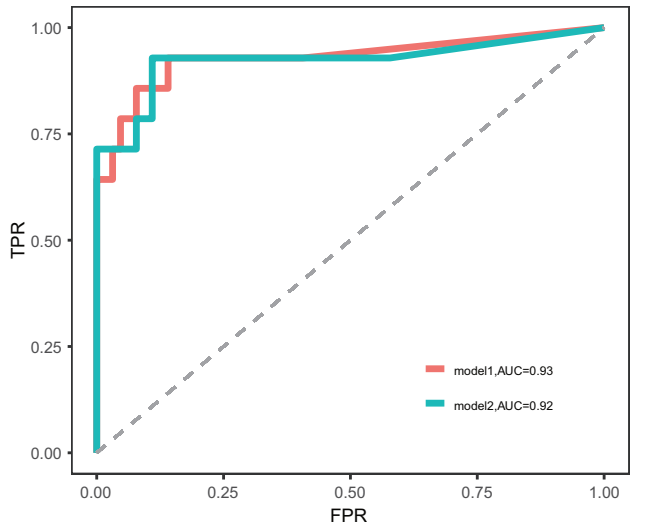

Figure 4. ROC curves of 2 models. Model 1: KRT7 and CXCL8, Model 2: VCAN, ANXA2, KRT7, KRT19, and CXCL8. ANXA2 = annexin A2, AUC = area under curves, CXCL8 = C-X-C motif chemokine ligand 8, FPR = false positive rate, KRT7 = keratin 7, KRT19 = keratin 19, ROC = receiver operating characteristic, TPR = true positive rate, VCAN = versican.

Table 2

Parameters of different model.

Model	Accuracy	95% CI	Sensitivity	Specificity
ANXA2+KRT19+VCAN+CXCL8+KRT7	0.95	(0.87–0.99)	1.0	0.71
ANXA2+KRT19+VCAN+CXCL8	0.91	(0.82–0.96)	0.97	0.64
ANXA2+VCAN+CXCL8	0.91	(0.82–0.96)	0.95	0.71
ANXA2+VCAN+CXCL8+KRT7	0.94	(0.86–0.98)	0.98	0.71
ANXA2+VCAN+KRT7	0.90	(0.81–0.95)	0.95	0.64
ANXA2+VCAN	0.85	(0.75–0.92)	0.92	0.50
KRT19+CXCL8+KRT7	0.92	(0.84–0.97)	0.98	0.64
CXCL8+KRT7	0.94	(0.86–0.98)	0.98	0.71
KRT7	0.78	(0.67–0.87)	0.89	0.28
KRT19+KRT7	0.83	(0.73–0.91)	0.92	0.43
CXCL8	0.90	(0.81–0.95)	0.94	0.71

ANXA2 = annexin A2, CI = confidence interval, CXCL8 = C-X-C motif chemokine ligand 8, KRT7 = keratin 7, KRT19 = keratin19, VCAN = versican.

Next, *KRT7* and *CXCL8* were found to play a core role among all genes. *CXCL8* chiefly responded for specificity, distinguishing non-BA liver diseases from BA. Therefore, *CXCL8* plus *KRT7* elevated the correct rate of identifying BA. The accuracy of combining *KRT7* and *CXCL8* was very close to the model consisting of *VCAN*, *ANXA2*, *CXCL8*, *KRT19*, *KRT7*, which only misidentified a BA as a non-BA. In contrast, the model's accuracy descended to 74.74% in external validation set GSE84954. Through in-depth analysis, the classification error of GSE46960 mainly occurred when the Non-BA group was classified as BA group, and the samples misclassified were principally idiopathic cholestasis. In GSE84954, the propensity for error was similar to GSE46960; non-BA diseases miscategorized separately were biliary cirrhosis and alpha-1-antitrypsin deficiency.

Furthermore, samples with the higher expression level of *KRT7* and *CXCL8* tended to be sorted as BA to a great extent, a phenomenon that was striking as shown by Figure 6B. The errors would occur when one of *KRT7* and *CXCL8* was expressed at an extremely high or low level, and another gene was expressed at a normal level. Given the expression of *KRT7* and *CXCL8* observed in GSE46960, thereby we took into account that the reason that induced the low accuracy of the model in GSE84954 mostly was the deviation of the results arising from the smaller sample size. The overall accuracy of the model, which was disturbed for errors caused by aberrant values, may improve with increasing sample size. Another reason we speculated on was that not all cholestatic diseases suited the model to differentiate.

In our study, the result of WGCNA analysis and function enrichment disclosed the changes of mRNA profiles with BA were in line with previously identified pathogenesis of cholestatic diseases containing BA and implied the core role of biliary progenitor, ECM and Notch pathway in the progress of BA. A novel diagnostic model was developed by lasso regression and randomforest algorithm, which was comparable in accuracy to the model comprised of *CXCL8* and *LAMC2* originating from the same dataset. This conclusion inferred that the model of *KRT7* and *CXCL8* might be worthy of further study and made us notice the unique value of *CXCL8* to discriminate BA with several cholestatic diseases. As the overlap in pathogenesis characteristic between BA and partial cholestasis, no single preoperative examination that enables the diagnosis of BA to be made certainty and thus a complex model that incorporates clinical manifest, laboratory examinations, image examinations, histological cachet, and biomarkers might be promising. Patients who opt for diagnosis by an exploratory laparotomy will benefit from it.

Given the limitations of our study in the cohort size (14 disease control samples) and the small size of the external validation set (11 samples), the outcome should be investigated in a broader context. Another aspect, since these data were from European and American countries, the practical value of the model in the mainland China needs to be inspected with native biospecimen.

5. Conclusion

WGCNA constructed a co-expression net encompassing ten modules and identified that the brown module is mainly related to BA. GO and KEGG enrichment analysis uncovered the genes of the brown module are predominantly enriched in biological

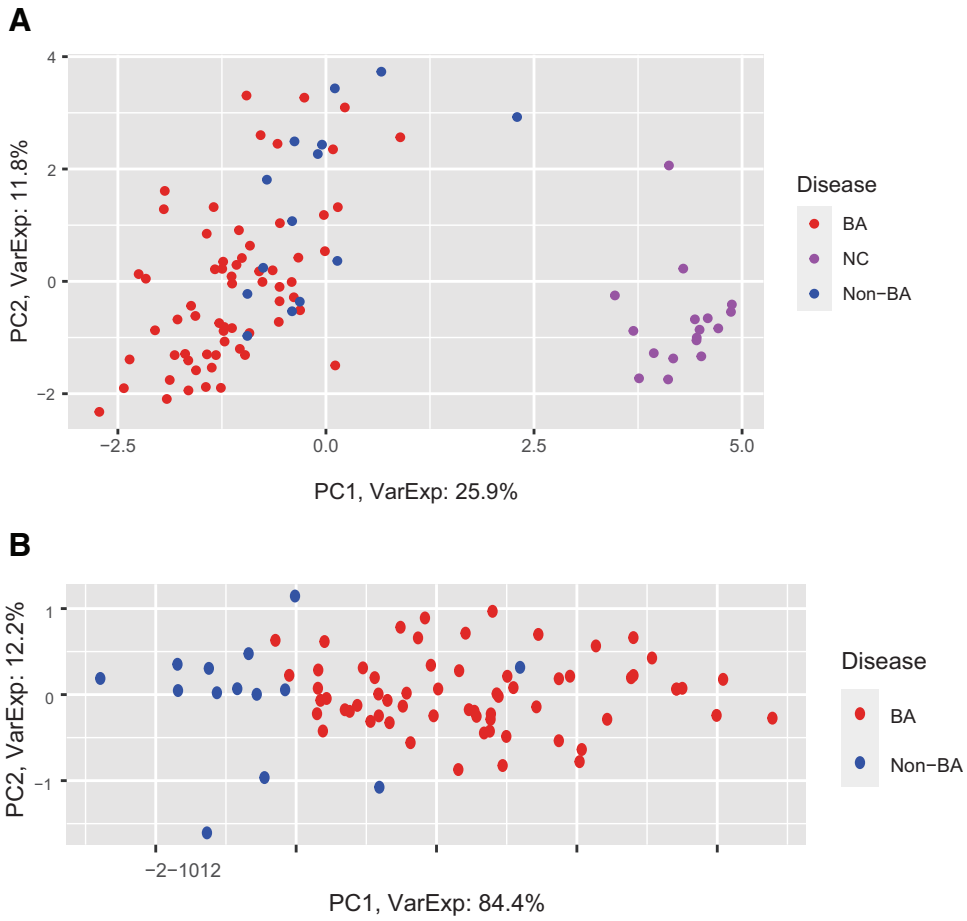

Figure 5. (A) PCA of original data. (B) PCA of the model of KRT7 and CXCL8. BA = biliary atresia, CXCL8 = C-X-C motif chemokine ligand 8, KRT7 = keratin 7, NC = normal control group, Non-BA = control group for hepatobiliary diseases without the biliary atresia, PCA = principal component analysis, PC1 = principal component 1, PC2 = principal component 2, VarExp = variance explained.

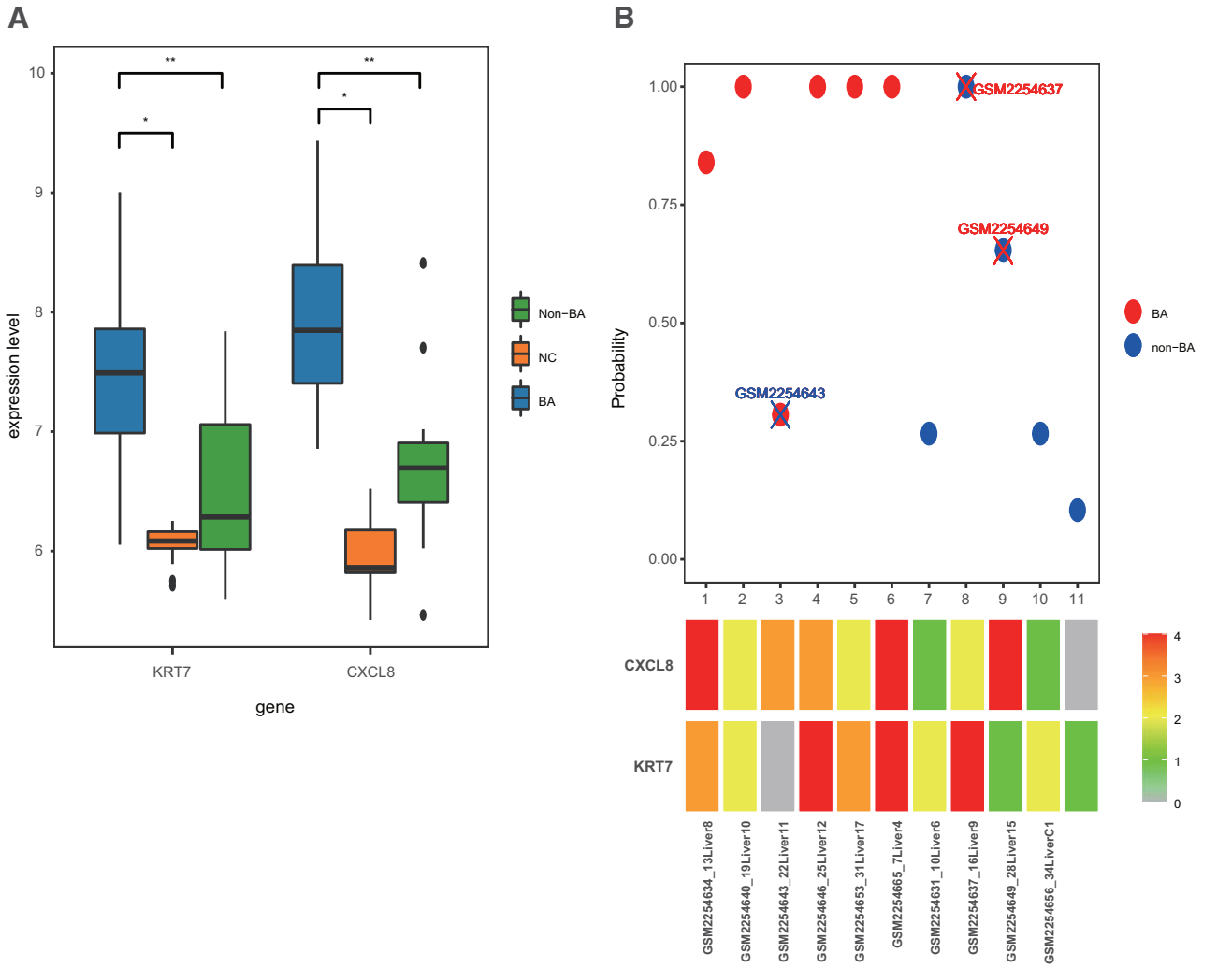

Figure 6. (A) Expression level of KRT7 and CXCL8 in GSE46960.*BA VS NC $P < .001$,**BA VS Non-BA $P < .001$. (B) The probability of each sample being classified as BA, calculated by the randomforest and the expression level of KRT7 and CXCL8 in the corresponding sample. Each point in upper graphic corresponded to the probability of a sample being classified as BA. Red typeface in annotation referred to the non-BA sample being misclassified as BA. Blue typeface in annotation referred to BA sample being misclassified as non-BA. The lower graphic exhibited the expression level of KRT7 and CXCL8 corresponding to every sample. BA = biliary atresia, CXCL8 = C-X-C motif chemokine ligand 8, KRT7 = keratin 7, NC = normal control group, Non-BA =control group for hepatobiliary diseases without the biliary atresia.

Table 3

Confusion matrix of predicting classification in GSE84954 (randomforest).

			Reference	
			BA	Non-BA
Liver tissue	Prediction	BA	5	2
		Non-BA	1	3

BA = biliary atresia, Non-BA =control group for hepatobiliary diseases without the biliary atresia.

Table 4

Confusion matrix of predicting classification in GSE84954 (KNN).

			Reference	
			BA	Non-BA
Liver tissue	Prediction	BA	6	2
		Non-BA	0	3

BA = biliary atresia, KNN = k-nearest neighbors algorithm, Non-BA = control group for hepatobiliary diseases without the biliary atresia.

processes of CAM, extracellular matrix organization, inflammatory response, and notch pathway. Thirty-nine hub genes from the brown module consisted of a PPI network. *KRT7* and *CXCL8* were screened out from the 39 genes by lasso regression and a diagnostic model was formed using a randomforest algorithm to distinguish between BA and non-BA with an approximate accuracy of 93.6% and an AUC of 0.93.

Acknowledgments

The authors thank Heya Wang Prof, from the 6th people's hospital of Guiyang city, for reviewing this paper and her helpful comments.

Author contributions

Yongliang Wang had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data interpretation.
Concept and design: Fang Li.
Acquisition, analysis, or interpretation of data: Yongliang Wang.
Drafting of the manuscript: Yongliang Wang.

Critical revision of the manuscript for important intellectual content: All authors.
Supervision: Hongtao Yuan.

References

[1] Sanchez-Valle A, Kassira N, Varela VC, et al. Biliary atresia: epidemiology, genetics, clinical update, and public health perspective. Adv Pediatr. 2017;64:285–305.
[2] Feldman AG, Sokol RJ. Recent developments in diagnostics and treatment of neonatal cholestasis. Semin Pediatr Surg. 2020;29:150945.
[3] Russo P, Magee JC Boitnott J, et al. Design and validation of the biliary atresia research consortium histologic assessment system for cholestasis in infancy. Clin Gastroenterol Hepatol. 2011;9:357–362. e2.
[4] Dubuisson L, Lepreux S, Bioulac-Sage P, et al. Expression and cellular localization of ®brillin-1 in normal and pathological human liver. J Hepatol. 2001;34:514–22.
[5] Jane L, Hartley MD, Deirdre AK. Biliary atresia. Lancet. 2009;374:1704–13.
[6] Rastogi A, Krishnani N, Yachha SK, et al. Histopathological features and accuracy for diagnosing biliary atresia by prelaparotomy liver biopsy in developing countries. J Gastroenterol Hepatol. 2009;24:97–102 .
[7] Teitelbaum DH. Parenteral nutrition-associated cholestasis. Curr Opin Pediatr. 1997;9:270–5.
[8] Morotti RA, Suchy FJ, Magid MS. Progressive familial intrahepatic cholestasis (PFIC) type 1, 2, and 3: a review of the liver pathology findings. Semin Liver Dis. 2011;31:3–10.
[9] Nelson DR, Teckman J, Di Bisceglie AM, et al. Diagnosis and management of patients with alpha1-antitrypsin (A1AT) deficiency. Clin Gastroenterol Hepatol. 2012;10:575–80.
[10] Vogel GF, Maurer E, Entenmann A, et al. Co-existence of ABCB11 and DCDC2 disease: infantile cholestasis requires both next-generation sequencing and clinical-histopathologic correlation. Eur J Hum Genet. 2020;28:840–4.
[11] Assis DN, Debray D. Gallbladder and bile duct disease in cystic fibrosis. J Cyst Fibros. 2017;16 (Suppl 2):S62–9.
[12] Fawaz R, Baumann U, Ekong U, et al. Guideline for the evaluation of cholestatic jaundice in infants: joint recommendations of the North American society for pediatric gastroenterology, hepatology, and nutrition and the European society for pediatric gastroenterology, hepatology, and nutrition. J Pediatr Gastroenterol Nutr. 2017;64:154–68.
[13] Department of Hepatobiliary Surgery PSB, Chinese Medical Association, Division of Pediatric Organ Transplantation BoOTP, Chinese Medical Association. Guidelines for diagnosing & treating biliary atresia (2018 Edition). J Clin Hepato. 2019;35:2435–40.
[14] Morotti RA, Jain D. Pediatric cholestatic disorders: approach to pathologic diagnosis. Surg Pathol Clin. 2013;6:205–25.
[15] Dong R, Dong K, Wang X, et al. Interleukin-33 overexpression is associated with gamma-glutamyl transferase in biliary atresia. Cytokine. 2013;61:433–7.
[16] Jiang J, Wang J, Shen Z, et al. Serum MMP-7 in the diagnosis of biliary atresia. Pediatrics. 2019;144.
[17] Bessho K, Mourya R, Shivakumar P, et al. Gene expression signature for biliary atresia and a role for interleukin-8 in pathogenesis of experimental disease. Hepatology. 2014;60:211–23.
[18] Peng X, Yang L, Liu H, et al. Identification of circulating MicroRNAs in biliary atresia by next-generation sequencing. J Pediatr Gastroenterol Nutr. 2016;63:518–23.
[19] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article17.
[20] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. 2008;9:559.
[21] Carey VJ, Gentry J, Whalen E, et al. Network structures and algorithms in bioconductor. Bioinformatics. 2005;21:135–6.
[22] Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010;26:2363–7.
[23] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22.
[24] Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform. 2011;12:1–8.
[25] Liaw A, Wiener M. Classification and regression by randomforest. R News. 2002;2:18–22.
[26] Kuhn M. caret: classification and regression training. 2021.
[27] Wickham H, Franois R, Henry L, et al. dplyr: a grammar of data manipulation. 2021.
[28] Klaus S, Hechenbichler K. kknn: weighted k-nearest neighbors. R package version 1.3.1. 2016.
[29] Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag; 2016.
[30] Kolde R. pheatmap: pretty heatmaps. 2018.
[31] Tennekes M. treemap: treemap visualization. 2021
[32] Gu Z, Hübschmann D. Simplify enrichment: a bioconductor package for clustering and visualizing functional enrichment results, Genom Proteom Bioinfor. 2022. doi: https://doi.org/10.1016/j.gpb.2022.04.008
[33] Taiyun Wei, Simko V. R package "corrplot": visualization of a correlation matrix. 2021.
[34] Lakshminarayanan B, Davenport M. Biliary atresia: a comprehensive review. J Autoimmun. 2016;73:1–9.
[35] Park SM. The crucial role of cholangiocytes in cholangiopathies. Gut Liver. 2012;6:295–304.
[36] Levavasseur F, Liétard J, Ogawa K, et al. Expression of laminin γ 1 cultured hepatocytes involves repeated CTC and GC elements in the LAMC1 promoter. Biochem J. 1996;313:745–52.
[37] Kiyozumi D, Taniguchi Y, Nakano I, et al. Laminin gamma1 C-terminal glu to gln mutation induces early postimplantation lethality. Life Sci Alliance. 2018;1:e201800064.
[38] Weng Y, Lieberthal TJ, Zhou VX, et al. Liver epithelial focal adhesion kinase modulates fibrogenesis and hedgehog signaling. JCI Insight. 2020;5.
[39] Yu Y, Mao L, Cheng Z, et al. A novel regQTL-SNP and the risk of lung cancer: a multi-dimensional study. Arch Toxicol. 2021;95:3815–27.
[40] Dillon P, Belchis D, Tracy T, et al. Increased expression of intercellular adhesion molecules in biliary atresia. Am J Pathol. 1994;145:263–7.
[41] Kobayashi H, Horikoshi K, Long L, et al. Serum concentration of adhesion molecules in postoperative biliary atresia patients: relationship to disease activity and cirrhosis. J Pediatr Surg. 2001;36:1297–301.
[42] Wahl S, Feldman G, McCarthy J. Regulation of leukocyte adhesion and signaling in inflammation and disease. J Leukoc Biol. 1996;59:789–96.
[43] Wang J, Wang W, Dong R, et al. Gene expression profiling of extrahepatic ducts in children with biliary atresia. Int J Clin Exp Med. 2015;8:5186.
[44] Lim AG, Jazrawi RP, Levy JH, et al. Soluble E-selectin and vascular cell adhesion molecule-1 (VCAM-1) in primary biliary cirrhosis. J Hepatol. 1995;22:416–22.
[45] Parola M, Pinzani M. Liver fibrosis: pathophysiology, pathogenetic targets and clinical issues. Mol Asp Med. 2019;65:37–55.
[46] Guo J, Friedman SL. Hepatic fibrogenesis. Semin Liver Dis. 27:413–26.
[47] Chiba S. Notch signaling in stem cell systems. Stem Cells. 2006;24:2437–47.
[48] Sparks EE, Huppert KA, Brown MA, et al. Notch signaling regulates formation of the three-dimensional architecture of intrahepatic bile ducts in mice. Hepatology. 2010;51:1391–400.
[49] Li L, Krantz ID, Deng Y, et al. Alagille syndrome is caused by mutations in human Jagged1, which encodes a ligand for Notch1. Nat Genet. 1997;16:243–51.
[50] Mao Y, Tang S, Yang L, et al. Inhibition of the notch signaling pathway reduces the differentiation of hepatic progenitor cells into cholangiocytes in biliary atresia. Cell Physiol Biochem. 2018;49:11151074–1123.
[51] Zhang X, Du G, Xu Y, et al. Inhibition of notch signaling pathway prevents cholestatic liver fibrosis by decreasing the differentiation of hepatic progenitor cells into cholangiocytes. Lab Invest. 2016;96:350–60.
[52] Zhang X, Xu Y, Chen JM, et al. Huang qi decoction prevents bdl-induced liver fibrosis through inhibition of notch signaling activation. Am J Chin Med. 2017;45:85–104.
[53] Verhulst S, Roskams T, Sancho-Bru P, et al. Meta-analysis of human and mouse biliary epithelial cell gene profiles. Cells. 2019;8:1117–39.
[54] Asai A, Malladi S, Misch J, et al. Elaboration of tubules with active hedgehog drives parenchymal fibrogenesis in gestational alloimmune liver disease. Hum Pathol. 2015;46:84–93.
[55] Quoseena M, Vuppaladadium S, Hussain S, et al. Functional role of annexins in zebrafish caudal fin regeneration—a gene knockdown approach in regenerating tissue. Biochimie. 2020;175:125–31.
[56] Kanta J. Elastin in the liver. Front Physiol. 2016;7:491.

[57] Lorena D, Darby IA, Reinhardt DP, et al. Fibrillin-1 expression in normal and fibrotic rat liver and in cultured hepatic fibroblastic cells: modulation by mechanical stress and role in cell adhesion. Lab Investig. 2004;84:203–12.
[58] Lamireau T, Dubuisson L, Lepreux S, et al. Abnormal hepatic expression of fibrillin-1 in children with cholestasis. Am J Surg Pathol. 2002;26:637–46.
[59] Rock MJ, Cain SA, Freeman LJ, et al. Molecular basis of elastic fiber formation: critical interactions and a tropoelastin-fibrillin-1 cross-link. J Biol Chem. 2004;279:23748–58.
[60] Wang MC, Lu Y, Baldock C. Fibrillin microfibrils: a key role for the interbead region in elasticity. J Mol Biol. 2009;388:168–79.
[61] Fabris L, Strazzabosco M. Epithelial-mesenchymal interactions in biliary diseases. Semin Liver Dis. 2011;31:11–32.