COL3A1, CXCL8, VCAN, THBS2, and COL1A2 are correlated with the onset of biliary atresia

Hui Li, MM[a], Lei Cao, PhD[b], Hong Li, BD[a],* [ID]

Abstract

Background: Biliary atresia (BA) is a devastating progressive fibro inflammatory disorder in infants. The exact etiology of BA is still unclear. This study aimed screen key genes potentially associated with the occurrence of BA.

Methods: All BA data was obtained from GSE46960 dataset. The limma package in R language was used for differentially expressed gene (DEG) analyses. gene ontology and Kyoto encyclopedia of genes and genomes enrichment analysis were performed on the screened DEGs, using "clusterProfiler" package. protein-protein interaction network was built based on STRING Cytoscape software (Bethesda, Rockville, MD). The logistic regression model was constructed based on the selected DEGs.

Results: There were totally 78 DEGs in BA samples compared with normal samples, which were significantly enriched in 200 biological process terms, 37 molecular function terms, 17 cellular component terms, and 18 Kyoto encyclopedia of genes and genomes pathways. Among which, the top 10 genes with the highest importance in protein-protein interaction network were selected. Subsequently, on the basis of the stepwise regression method and 5-fold cross-validation, the logistic regression model constructed based on COL3A1, CXCL8, VCAN, THBS2, and COL1A2 was finally evidenced to predict the BA sample relatively reliably.

Conclusions: In conclusion, COL3A1, CXCL8, VCAN, THBS2, and COL1A2 are potentially crucial genes in BA. The logistic regression model constructed based on them could predict the BA sample relatively reliably.

Abbreviations: BA = biliary atresia, DEG = differentially expressed gene, GO = gene ontology, KEGG = Kyoto encyclopedia of genes and genomes, PPI = protein-protein interaction, WB = Western blot.

Keywords: biliary atresia, bioinformatic analysis, biomarker, diagnostic model

1. Introduction

Biliary atresia (BA) is a devastating progressive fibroinflammatory disorder in infants, which is usually involved in the extrahepatic and intrahepatic biliary tree and finally results in cholestasis, fibrosis, and cirrhosis.[1,2] Commonly, the main clinical manifestations of BA are prolonged jaundice over 14 days of age, dark urine, and pale-pigmented stools.[3] To date, the exact etiology of BA is still unclear. Kasai portoenterostomy and liver transplantation are typical treatments for most BA infants.[4] Unfortunately, along with the increase of surgery age, the success rate of Kasai portoenterostomy decreased significantly.[5] Once BA is untreated timely, the patient would invariably suffer from end-stage liver disease, even death.[6] Clearly, early detection is important for BA. However, on the 1 hand, the phenotypic heterogeneity of different BA patients, for instance, biliary tract obstruction along with cystic components, and extrahepatic biliary tracts completely atretic,[7] on the other hand, overlapping clinical symptoms of cholestasis due to other causes (not BA),[8] both of which bring great challenges for the early diagnosis of BA in clinical cases. Consequently, looking for reliable diagnostic biomarkers is necessary to improve the prognosis of BA patients.

Despite the exact etiology of BA is unknown, increasing studies have evidenced that complex factors including genetic alterations drive the onset of BA.[9,10] Several studies have made attempts in this regard with the hope of applying it to the early diagnosis of BA. A recent study has suggested that hepatic IL8 and LAMC2 expression had a relatively high diagnostic sensitivity and specificity for BA, and they might be potential biomarkers of BA.[11] GPC1 was reported to be a potential BA susceptibility gene.[1] Besides, several researchers have reported that the potential roles of various genes in the occurrence of BA, such as JAG1,[12,13]

Figure 1. (A) Volcano map of differentially expressed genes (DEG) in GSE46960. X-axis: Log2FC; Y-axis: -log10 (FDR). Blue: downregulated genes; red: upregulated genes; green: nonsignificant. (B) Heat map of DEGs in GSE46960. X-axis: sample; Y-axis: different genes. Red: high expression; Blue: low expression.

SERPINA1,[14] PKHD1,[15] and so on. However, there are still few biomarkers that can be applied in the clinical diagnosis of BA. Accordingly, it is still urgent and important to explore biomarkers potentially involved in the pathogenesis of BA.

Herein, via a series of bioinformatic analyses of the BA data downloaded from gene expression omnibus database and further validation experiments, we expected to screen key genes potentially associated with BA occurrence, to provide new diagnostic ideas and methods for BA in the near future.

2. Materials and Methods

2.1. Data sources

Our study was mainly based on the gene chip data in gene expression omnibus database (https://www.ncbi.nlm.nih.gov/geo/), and the BA sample data was obtained from GSE46960 dataset. There were totally 95 samples in GSE46960 dataset, which included 64 liver tissue samples from patients with BA, 21 normal liver tissue samples (no BA), and 10 internal quality control samples. The mRNA expression level of all samples in GSE46960 dataset was detected on Affymetrix Human Gene 1.0 ST Array chip platform.

2.2. Clinical sample collection

All BA samples were collected in Tianjin First Central Hospital. Totally 15 BA tissue specimens and 15 healthy controls were collected. All experiments were approved by the ethics committee of Tianjin First Central Hospital [Review number: 2021N158KY], in accordance with The Helsinki Declaration.

The informed consents were obtained from all participants. The patients detailed clinical information were summarized in Table S1, Supplemental Digital Content, http://links.lww.com/MD/I674.

2.3. Differential expression analyses

Firstly, all raw data were normalized according to the robust multi-array average method, which was then standardized by taking the log2 logarithm value. The differential expression analyses were performed on the data in GSE46960 dataset after standardization. The limma function package[16] in R language (version 3.5.2, the same below) was used for differentially expressed gene (DEG) analyses. Finally, the DEGs with |Log2FC| >1 and FDR ≤ 0.05 were screened.

2.4. Functional enrichment analyses

Subsequently, gene ontology (GO) enrichment analysis (including biological process, molecular function, and cellular component) and Kyoto encyclopedia of genes and genomes (KEGG) enrichment analysis were performed on the screened DEGs, using "clusterProfiler" function package[17] in R language. The GO and KEGG terms with P value. adjust < .05 were considered as significantly enriched terms.

2.5. Protein-protein interaction (PPI) analyses

The STRING database is usually used for protein functional connections and PPI prediction. Thus, protein functional connections and PPI in our study were analyzed using STRING
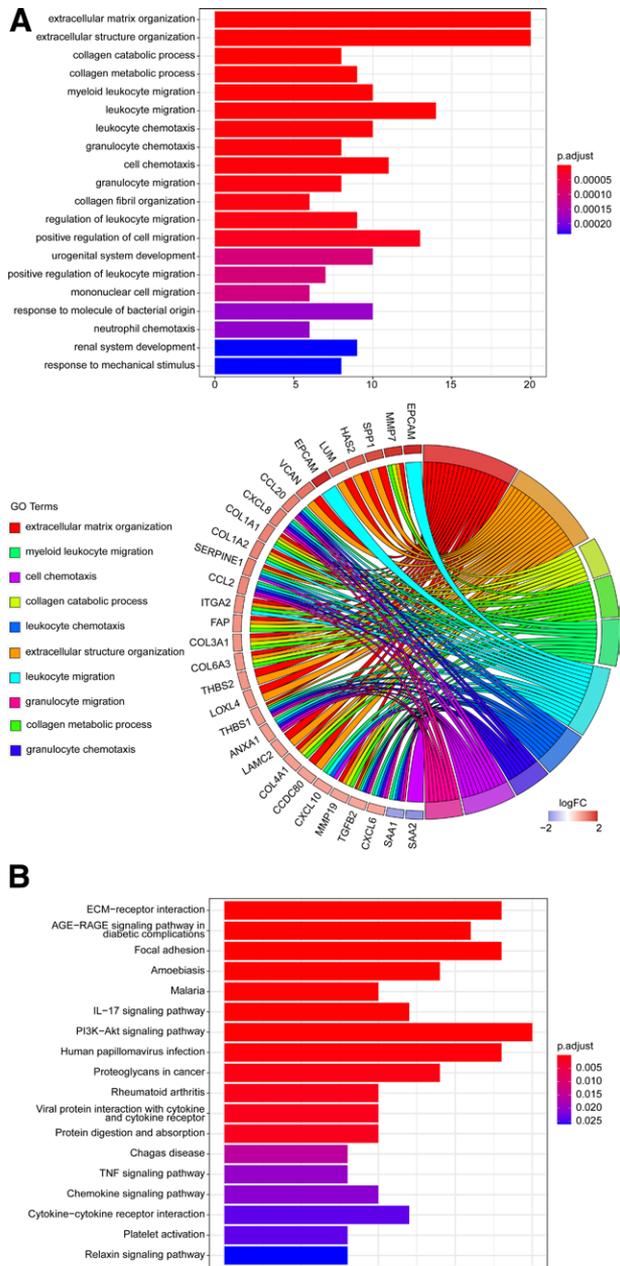
Figure 2. Functional enrichment analysis results. (A) The top 20 significantly enriched GO terms (X-axis: gene number; Y-axis: the title of GO terms) and genes contained in the top 10 GO terms (Right semicircle: 10 GO terms; left semicircle: genes enriched in these 10 GO terms). (B) The 18 significantly enriched KEGG pathways. X-axis: gene number; Y-axis: the title of KEGG pathways. GO = gene ontology, KEGG = Kyoto encyclopedia of genes and genomes.

Figure 3. PPI network construction results. (A) PPI network diagram. Each dot in network represented a node. The more line segments connected to the node represented, the greater degree of this node, which suggested that the gene on this node was more important in PPI network. (B) The top 10 genes with the highest degree in the PPI network were screened via the MNC algorithm. The darker the red color, the higher the degree of gene. MNC = maximum neighborhood component, PPI = protein-protein interaction.

(version 11.0) (https://string-db.org/).[18] The PPI network was then visualized using Cytoscape software (version 3.7.2).[19] Moreover, based on maximum neighborhood component algorithm, the core genes in PPI network were further screened using the cyto Hubba plug-in in Cytoscape software.

2.6. Logistic regression model construction

Logistic regression is a common method in classification, which is predicting a classification result based on a set of variables. In the present study, based on logistic regression, the mRNA expression was used to predict the sample type (BA or normal).

All samples were divided into 2 categories, BA samples and normal samples. Then mRNA expression as a continuous predictive variable was included in glm function in R language to construct the multivariable logistic regression model, and the sample type (BA or normal) was taken as the classification response value. Subsequently, the variables were further screened according to the stepwise regression method and the screened variables were used to reconstruct the model. Based on the P value of each variable calculated by the model, the variables with $P < .05$ were taken as candidate mRNA to reconstruct the final model. According to 5-fold Cross-Validation method, the sample data in the GSE46960 dataset was divided into 2 parts, training set and validation set, to construct and validate the logistic model.
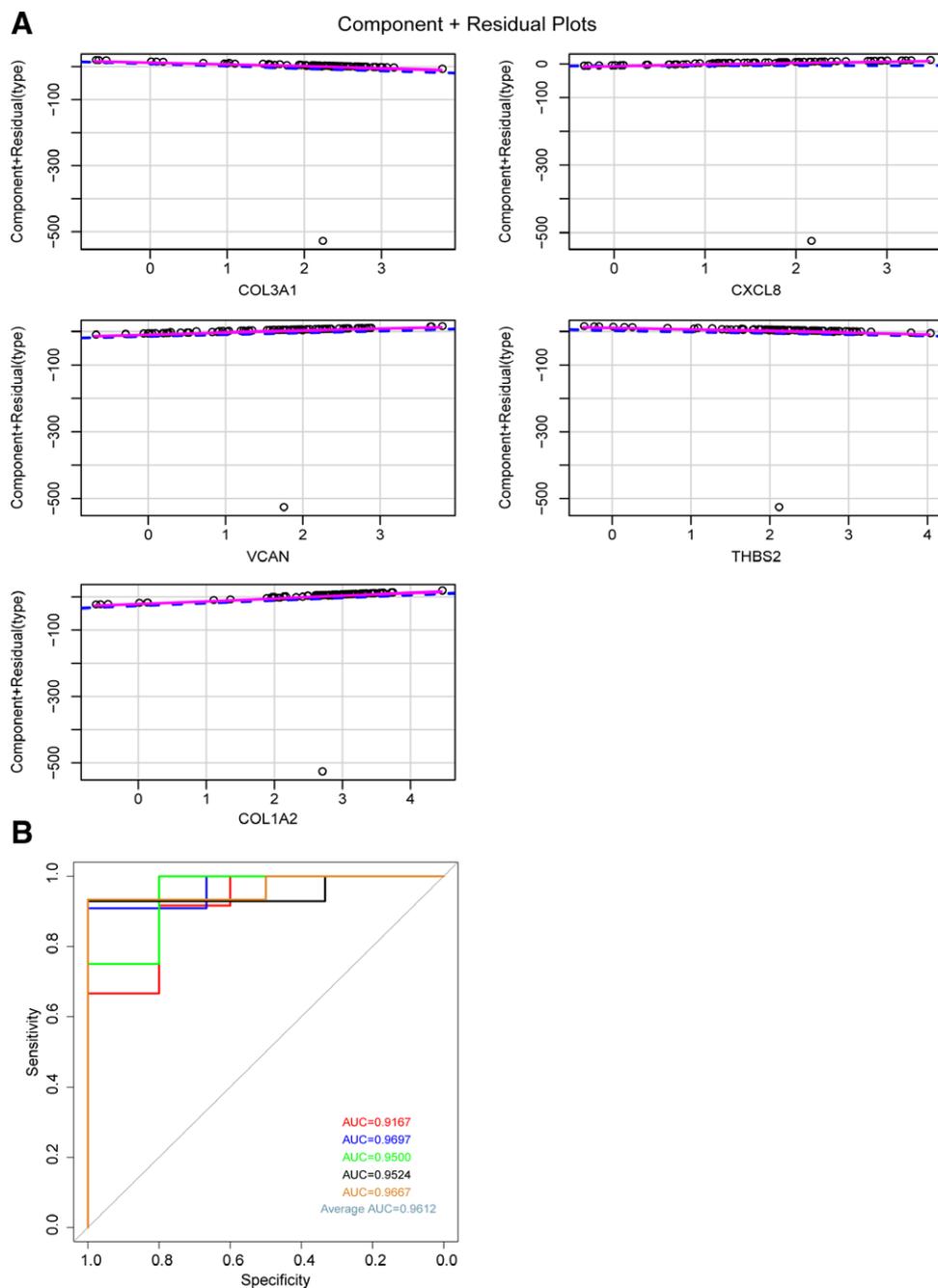
Figure 4. Logistic regression model construction and prediction results. (A) The component plus residual plot of the 5 final mRNAs included in the model. The obvious linear relationship between X-axis and Y-axis indicated that the independent variables were suitable for the model. (B) ROC curve. X-axis: false positive rate (FPR) specificity; Y-axis: true positive rate (TPR) sensitivity. The AUC value could intuitively evaluate the quality of the model. AUC = area under the curve, ROC = receiver operating characteristic.

2.7. Quantitative real-time polymerase chain reaction and western blot (WB) assays

Total RNA extraction was conducted using RNA prep pure Tissue Kit (DP431, TIANGEN Biotech, Beijing, China). During reverse transcription and PCR amplification, Takara Prime Script RT Reagent Kit with gDNA Eraser (RR047A, Takara), Takara TB Green Premix Ex Taq II (RR820A, Takara), and Roche LightCycler96 quantitative PCR instrument were used. The primer sequences were shown in Table 1. The qPCR procedure was as below: pre-denaturation at 95°C for 30 seconds;

40 cycles of 95°C for 3 seconds, 60 °C for 30 seconds (3 repeats per sample). The internal reference was GAPDH. The relative mRNA expression was calculated using formula $2\text{-}\Delta\Delta CT$.

Subsequently, the WB assay was done utilizing the general previous methods.[20] The reagents we used included BCA protein quantification kit (ZJ101, Shanghai Epizyme Biomedical Technology, China), internal reference β-actin (SC-47778, Santa Cruz, 1:1000), primary antibody COL3A1 antibody (sc-271249, 1:1000, Santa Cruz), and secondary antibody Goat Anti-Mouse IgG H&L (HRP) (ab205719, 1: 10,000, abcam). Optical density analysis was done in ImageJ software.
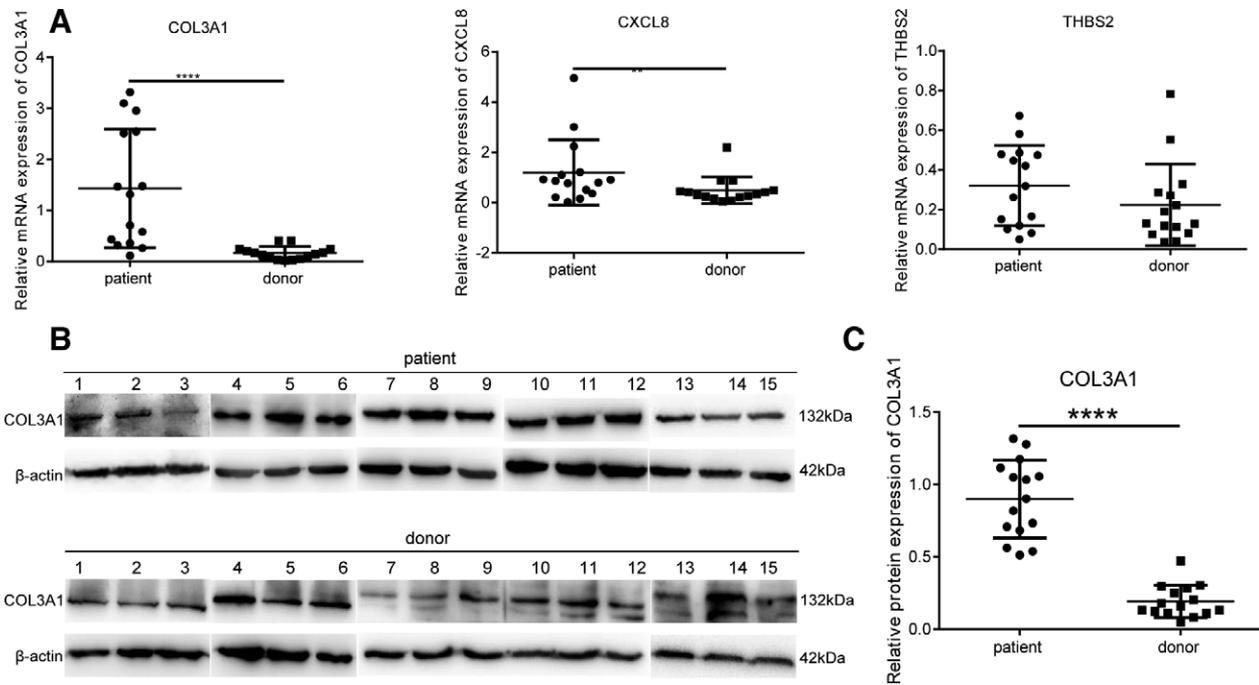
Figure 5. The expression of COL3A1, CXCL8, and THBS2 in clinical BA samples. (A) The mRNA expression levels of COL3A1, CXCL8, and THBS2 in BA samples (n = 15, *: BA patients compared with donors (healthy controls), ** $P < .01$, **** $P < .0001$). (B–C) The protein expression of COL3A1 in BA samples compared with donors (healthy controls) (n = 15, **** $P < .0001$). BA = biliary atresia.

3. Results

3.1. Differentially expressed genes

Since the data in GSE46960 dataset has been standardized, the standardized data was used for further DEG analysis. In GSE46960 dataset, compared with normal samples, totally 78 DEGs were identified in BA samples, including 68 upregulated genes and 10 downregulated genes (Fig. 1A). The expression levels of DEGs were significantly different between BA samples and normal samples (Fig. 1B).

3.2. Results of functional enrichment analyses

To better know the functional information of these 78 DEGs, we have performed GO and KEGG enrichment analyses on these 78 DEGs. They were significantly enriched (P value. adjust < .05) in 200 biological process terms, 37 molecular function terms, 17 cellular component terms, and 18 KEGG pathways (detailed enrichment analysis results see Table S2, Supplemental Digital Content, http://links.lww.com/MD/I675). The top 20 significantly enriched GO terms and the genes in the top 10 significantly enriched GO terms were shown in Figure 2A. All 18 significantly enriched KEGG pathways were displayed in Figure 2B.

3.3. PPI network construction and core gene screening

Furthermore, the STRING database was used to construct a PPI network based on 78 DEGs, and interaction pairs with minimum required interaction score > 0.4 were screened. The PPI network was visualized using Cytoscape software, see Figure 3A. There were 74 nodes and 174 edges in Figure 3A, in which, each node represented a gene and the edge represented the interaction between them. Then the topological structure of the whole PPI network was analyzed using Cytoscape software. The importance of each node in PPI network was scored according to maximum neighborhood component

algorithm, the top 10 genes with the highest importance were selected; the darker the color, the higher the importance of the node (Fig. 3B). Detailed results of the 10 genes see Table S3, Supplemental Digital Content, http://links.lww.com/MD/I676.

3.4. Logistic diagnostic model and validation

Next, the 10 genes screened by PPI network were used to construct a logistic regression model. As a stronger model was expected to be constructed based on fewer variables, 5 genes were further screened out of the 10 genes according to the stepwise regression method, including COL3A1, CXCL8, VCAN, THBS2, and COL1A2.

Taking all these 5 genes into the model as variables, we found that the p values of these 5 genes were <0.05, which indicated that they contributed a lot to the model. Then the model was evaluated, and the results suggested that the model conforms to the normal distribution (Figure S1A, Supplemental Digital Content, http://links.lww.com/MD/I677). Not only that, there was a good linear relationship between the independent variables and response variables included in the model (Fig. 4A), and there was no extreme value that significantly affected the accuracy of the model (Figure S1B, Supplemental Digital Content, http://links.lww.com/MD/I677).

Moreover, the reliability of the model was evaluated according to the 5-fold cross-validation method. Firstly, all samples were randomly divided into 5 groups, 4 groups of data were taken as the training set to construct a logistic model every time, and the other group of data was taken as the validation set, and above process was repeated 5 times. It had been ensured through the cross-validation process that each subsample has participated in training and test, which could reduce errors and reflect the detected ability of model more realistically. We found that AUC of 5 logistic models constructed according to 5-fold cross-validation method in 5 validation sets were 0.9167, 0.9697, 0.9500, 0.9524, and 0.9667 (Fig. 4B), respectively; the average AUC was 0.9612. The results showed that the model

constructed according to the 5 genes could determine the sample type relatively reliably, which indicated indirectly that the 5 genes (COL3A1, CXCL8, VCAN, THBS2, and COL1A2) were correlated with the onset of BA.

3.5. COL3A1, CXCL8, and THBS2 were highly expressed in clinical BA samples

Our bioinformatic analysis results basing on public BA data indicated that COL3A1, CXCL8, and THBS2 showed higher expression levels in BA samples compared with normal samples. In our local clinical BA samples (n = 15), COL3A1, CXCL8, and THBS2 mRNA expression levels were all higher-than-normal controls (n = 15) (Fig. 5A). Moreover, the protein expression of the most significant COL3A1 was determined by WB. The results indicated that COL3A1 protein expression was consistent with the mRNA expression in BA samples (Fig. 5B and C). Collectively, our data suggested that the expression of COL3A1, CXCL8, and THBS2 in local clinical samples were in line with public data mining results.

4. Discussion

Currently, the etiology and pathogenesis of BA is still unclear. In our present study, based on bioinformatic analyses and further experimental validation, we have identified 5 genes, including COL3A1, CXCL8, VCAN, THBS2, and COL1A2, which were potentially associated with the occurrence of BA. Besides, the logistic regression model constructed based on these genes could predict the BA sample relatively reliably.

Despite it has been revealed that genetic susceptibility, viral infection, or autoimmune disease might be related to BA,[9,21,22] the details remain to be elucidated. Accordingly, we expected to screen key genes potentially associated with the occurrence of BA. Firstly, based on the BA data in GSE46960 dataset, there were totally 78 DEGs in BA samples compared with normal samples. Subsequently, GO and KEGG enrichment analyses were performed on the 78 DEGs to better know their functional information. These 78 DEGs were significantly enriched in 254 GO terms and 18 KEGG pathways. Many of the KEGG pathways were related to the immune response, which was consistent with some previous studies.[22,23] A recent study has demonstrated that IL-17A signaling pathway was probably involved in chronic inflammation of the bile duct in primary biliary cirrhosis,[24] in which, chronic inflammation was associated with biliary innate immune responses.[25] It seems to be linked to our study. PI3K/Akt signaling pathway is widely involved in many biological processes like cell migration, apoptosis inhibition, and so on.[26] It has been documented that miR-200b-activated PI3K/Akt signaling pathway in BA patients has accelerated the proliferation of hepatic stallate cells.[27] Our findings complemented the influence of DEGs on PI3K/Akt pathway in BA.

Furthermore, based on 78 DEGs, a PPI network was constructed to further screen the top 10 most important DEGs. Subsequently, on the basis of the stepwise regression method and 5-fold cross-validation, the logistic regression model constructed based on COL3A1, CXCL8, VCAN, THBS2, and COL1A2 was finally evidenced to predict the BA sample relatively reliably. Moreover, we have found some previous reports that could support our findings to some extent. For instance, COL3A1 was suggested to be involved in the process of liver fibrosis,[28] and the significant higher expression of COL3A1 was observed in fibrotic tissue compared to healthy tissue,[29] both of which might be related to the liver fibrosis during the development of BA. As for CXCL8, Zhang et al[21] reported that B cells could secrete CXCL8, which was correlated with the pathogenesis (proinflammatory effect) of BA. Regarding THBS2, there

was no direct research about THBS2 in BA, but we found that it had been suggested to be associated with the poor prognosis of hepatocellular carcinoma.[30] Not only that, Byrling et al[31] indicated that stromal THBS2 might be a possible prognostic marker in distal cholangiocarcinoma. However, the relevant reports and evidence we found was relatively fragmented. Thus, how these genes play roles in BA still needs large amounts of further exploration in the future.

5. Conclusions

In summary, based on the BA data publicly obtained from GSE46960 dataset, via bioinformatic analyses and further experimental validation, COL3A1, CXCL8, VCAN, THBS2, and COL1A2 have been evidenced to be correlated with the onset of BA. Our diagnostic model basing on these 5 genes show good predictive performance to distinguish BA samples. Moreover, the functional pathways indicate that the BA related DEGs have exerted crucial roles affecting immune responses. Our findings provide more reference information for better understanding the potential pathogenic genes and expanding candidate therapeutic targets in BA patients. Furthermore, this diagnostic model for BA exhibits a great potential in clinical application after a deepening investigation in a larger sample size.

Author contributions

Conceptualization: Hui Li.
Data curation: Hui Li, Lei Cao, Hong Li.
Formal analysis: Hui Li, Hong Li.
Funding acquisition: Hui Li, Lei Cao, Hong Li.
Investigation: Hui Li, Lei Cao.
Methodology: Hui Li, Hong Li.
Project administration: Hui Li.
Resources: Hui Li, Lei Cao.
Software: Hui Li, Hong Li.
Supervision: Hui Li, Lei Cao.
Validation: Hui Li, Lei Cao.
Visualization: Hui Li.
Writing – original draft: Hui Li, Lei Cao, Hong Li.
Writing – review & editing: Hui Li, Lei Cao, Hong Li.

References

[1] Cui S, Leyva-Vega M, Tsai EA, et al. Evidence from human and zebrafish that GPC1 is a biliary atresia susceptibility gene. Gastroenterology. 2013;144:1107–1115.e3.

[2] Miethke AG, Huppert SS. Fishing for biliary atresia susceptibility genes. Gastroenterology. 2013;144:878–81.

[3] Matsui A. Screening for biliary atresia. Pediatr Surg Int. 2017;33:1305–13.

[4] Ohi R. Surgery for biliary atresia. Liver. 2001;21:175–82.

[5] Mieli-Vergani G, Vergani D. Biliary atresia. Semin Immunopathol. 2009;31:371–81.

[6] Hartley JL, Davenport M, Kelly DA. Biliary atresia. Lancet. 2009;374:1704–13.

[7] Thomson J. On congenital obliteration of the bile-ducts. Trans Edinb Obstet Soc. 1892;17:17–49.

[8] Yang L, Zhou Y, Xu PP, et al. Diagnostic accuracy of serum matrix metalloproteinase-7 for biliary atresia. Hepatology. 2018;68:2069–77.

[9] Mezina A, Karpen SJ. Genetic contributors and modifiers of biliary atresia. Dig Dis. 2015;33:408–14.

[10] Girard M, Panasyuk G. Genetics in biliary atresia. Curr Opin Gastroenterol. 2019;35:73–81.

[11] 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.

[12] Dedic T, Jirsa M, Keil R, et al. Alagille syndrome mimicking biliary atresia in early infancy. PLoS One. 2015;10:e0143939.

[13] Sangkhathat S, Laochareonsuk W, Maneechay W, et al. Variants associated with infantile cholestatic syndromes detected in extrahepatic biliary atresia by whole exome studies: a 20-case series from Thailand. J Pediatr Genet. 2018;7:67–73.

[14] Campbell KM, Arya G, Ryckman FC, et al. High prevalence of alpha-1-antitrypsin heterozygosity in children with chronic liver disease. J Pediatr Gastroenterol Nutr. 2007;44:99–103.

[15] Hartley JL, O'Callaghan C, Rossetti S, et al. Investigation of primary cilia in the pathogenesis of biliary atresia. J Pediatr Gastroenterol Nutr. 2011;52:485–8.

[16] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

[17] Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.

[18] Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47:D607–13.

[19] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

[20] Chen J, Harding SM, Natesan R, et al. Cell cycle checkpoints cooperate to suppress DNA- and RNA-associated molecular pattern recognition and anti-tumor immune responses. Cell Rep. 2020;32:108080.

[21] Zhang Y, Zhou L, Gu G, et al. CXCL8[high] inflammatory B cells in the peripheral blood of patients with biliary atresia are involved in disease progression. Immunol Cell Biol. 2020;98:682–92.

[22] Ortiz-Perez A, Donnelly B, Temple H, et al. Innate immunity and pathogenesis of biliary atresia. Front Immunol. 2020;11:329.

[23] Kim S, Moore J, Alonso E, et al. Correlation of immune markers with outcomes in biliary atresia following intravenous immunoglobulin therapy. Hepatol Commun. 2019;3:685–96.

[24] Zhan K, Xu Y, Han M, et al. Daifan San intervenes in fork head box P3 and the interleukin (IL)-23/IL-17A signaling pathway to help prevent and treat primary biliary cirrhosis. J Tradit Chin Med. 2020;40:571–83.

[25] Harada K, Shimoda S, Sato Y, et al. Periductal interleukin-17 production in association with biliary innate immunity contributes to the pathogenesis of cholangiopathy in primary biliary cirrhosis. Clin Exp Immunol. 2009;157:261–70.

[26] Xie Y, Shi X, Sheng K, et al. PI3K/Akt signaling transduction pathway, erythropoiesis and glycolysis in hypoxia (Review). Mol Med Rep. 2019;19:783–91.

[27] Xiao Y, Wang J, Chen Y, et al. Up-regulation of miR-200b in biliary atresia patients accelerates proliferation and migration of hepatic stallate cells by activating PI3K/Akt signaling. Cell Signal. 2014;26:925–32.

[28] Tao R, Fan XX, Yu HJ, et al. MicroRNA-29b-3p prevents Schistosoma japonicum-induced liver fibrosis by targeting COL1A1 and COL3A1. J Cell Biochem. 2018;119:3199–209.

[29] Mirzavand S, Rafiei A, Teimoori A, et al. Gene expression in human liver fibrosis associated with Echinococcus granulosus sensu lato. Parasitol Res. 2020;119:2177–87.

[30] Zhang J, Hao N, Liu W, et al. In-depth proteomic analysis of tissue interstitial fluid for hepatocellular carcinoma serum biomarker discovery. Br J Cancer. 2017;117:1676–84.

[31] Byrling J, Kristl T, Hu D, et al. Mass spectrometry-based analysis of formalin-fixed, paraffin-embedded distal cholangiocarcinoma identifies stromal thrombospondin-2 as a potential prognostic marker. J Transl Med. 2020;18:343.