ORIGINAL ARTICLE
Iran J Allergy Asthma Immunol
In press.

Peripheral Blood Immunogenomic Cytokine-receptor Signature (*CXCR1*, *IL11RA*, *IL13RA2*, *CD19*) Predicts HBV-related Cirrhosis: Public PBMC Transcriptome Mining and Nomogram Development

Guifang Jiang[1] and Junping Pan[2]

[1] Clinical Laboratory, Huzhou Wuxing District People's Hospital, Huzhou Wuxing District Maternal and Child Health Hospital, Huzhou, China
[2] Laboratory, Huzhou Wuxing District Center for Disease Prevention and Control (Wuxing District Health Supervision Institute), Huzhou, China

Received: 12 December 2025; Received in revised form: 27 January 2026; Accepted: 8 February 2026

ABSTRACT

Progression from chronic hepatitis B (CHB) to cirrhosis is closely associated with immune dysregulation and altered cytokine signaling, yet effective noninvasive immune biomarkers remain limited. This study aimed to develop a peripheral blood mononuclear cell (PBMC)-based cytokine gene signature for predicting HBV-related cirrhosis.

The GSE114783 microarray dataset was analyzed to identify differentially expressed genes between CHB and cirrhosis. Immune-related candidate genes were selected by integrating curated immune resources and the Kyoto Encyclopedia of Genes and Genomes cytokine-cytokine receptor interaction pathway. Least absolute shrinkage and selection operator regression and multivariable logistic regression were used to construct the predictive model. Receiver operating characteristic analysis, leave-one-out cross-validation, and nomogram development were performed to evaluate model performance and support clinical translation.

Differential expression analysis identified 3169 genes distinguishing cirrhosis from CHB, from which 86 immune/cytokine-related genes were prioritized. LASSO selected a parsimonious 4-gene signature-*CXCR1*, *IL11RA*, *IL13RA2*, and *CD19*-capturing key immune axes relevant to chronic inflammation and fibrogenesis (chemokine receptor signaling, IL-11/IL-13 receptor pathways, and B-cell–associated immunity). The resulting model achieved an area under the curve (AUC) of 0.935 (95% CI, 0.882-0.988); at an optimal cutoff of 0.832, sensitivity was 88.6% and specificity 84.3%. LOOCV supported robust performance, and the nomogram demonstrated good agreement between predicted and observed risk.

A PBMC-based immune cytokine-receptor gene signature (*CXCR1/IL11RA/IL13RA2/CD19*) provides a noninvasive tool for immunologically informed risk stratification of HBV-related cirrhosis and may support immune monitoring and early intervention strategies. Prospective, multicohort validation and mechanistic studies are warranted.

Keywords: B cells; Chronic hepatitis B; Chemokine signaling; Cytokine receptors; Immunogenomics; LASSO; Liver cirrhosis; Machine learning, Nomogram; PBMC

Corresponding Author: Junping Pan, BM;
Laboratory, Huzhou Wuxing District Center for Disease Prevention and Control (Wuxing District Health Supervision Institute), Huzhou, China. Tel: (+86 018) 3672 80213, Fax: (+86 018) 3672 80213, Email: 3070780599@qq.com

Published by Tehran University of Medical Sciences (http://ijaai.tums.ac.ir)

INTRODUCTION

Liver cirrhosis, the end-stage pathological outcome of chronic hepatitis B (CHB) progression, is a major cause of liver-related mortality worldwide and imposes a substantial burden on public health systems.[1] Hepatitis B virus (HBV) infection is a primary driver of cirrhosis development. According to the latest World Health Organization estimates, 296 million people globally are living with chronic HBV infection, and approximately 15% to 25% may progress to cirrhosis or hepatocellular carcinoma without effective intervention.[2] This pathological process is particularly prominent in the Asia-Pacific region. Due to its large population base and high infection rate, the disease burden of HBV-related cirrhosis has become a key challenge in liver disease prevention and control in this region.[3] The pathological hallmark of cirrhosis is diffuse hepatic fibrosis with regenerative nodule formation; progression is typically insidious and largely irreversible. Once patients develop decompensated cirrhosis-manifested by complications such as ascites, hepatic encephalopathy, or ruptured esophagogastric variceal bleeding-the 5-year survival rate declines sharply to less than 50%.[4] Therefore, early identification of individuals at high risk for progression to cirrhosis during the CHB stage, coupled with individualized precision interventions, is critical clinical strategy to improve long-term outcomes and reducing end-stage liver disease–related mortality. This need also represents a major bottleneck in hepatology, urgently requiring breakthroughs in both basic research and clinical practice.

Current clinical methods for diagnosis and risk assessment of liver cirrhosis have notable limitations that hinder early warning and dynamic monitoring. In terms of imaging examinations, modalities relying on hepatic morphological changes-such as abdominal ultrasonography, computed tomography, and magnetic resonance imaging-exhibit insufficient sensitivity for detecting early cirrhosis (or significant fibrosis stage F3),[5] often enabling diagnosis only after substantial structural liver damage has occurred. Transient elastography (e.g., FibroScan) is a key noninvasive tool for assessing liver fibrosis, but its measurements are influenced by multiple confounding factors including obesity, narrow intercostal spaces, hepatic inflammatory activity, and operator experience; accuracy may be further reduced in patients with comorbidities such as fatty liver disease.[6] Regarding serological models, combined indices based on routine biochemical parameters-including the Aspartate Aminotransferase to Platelet Ratio Index (APRI), Fibrosis-4 (FIB-4) index, and 4-item liver fibrosis panel-offer simplicity, low cost, and high reproducibility. However, their diagnostic efficacy plateaus significantly: they demonstrate limited accuracy in differentiating moderate fibrosis (stage F2) from severe fibrosis (stage F3) and early cirrhosis (stage F4). Additionally, these models indirectly reflect hepatic synthetic function and fibrosis status and fail to capture dynamic molecular pathological changes during disease progression.[6-10] Liver biopsy, the "gold standard" for cirrhosis diagnosis,[11] can directly assess liver fibrosis staging but is invasive, carrying risks of complications such as bleeding, pain, and infection. Furthermore, sampling bias (error rate up to 10%-20%) is common due to limitations in biopsy site selection, and interobserver variability in pathological interpretation may compromise result reliability. These inherent methodological flaws collectively highlight the urgent clinical need for developing a novel cirrhosis risk assessment tool that is more accurate, non-invasive, and capable of reflecting the intrinsic biological processes of the disease.

With the rapid advances in high-throughput sequencing and molecular biology, biomarker discovery based on gene expression profiles has emerged as a promising approach to addressing these challenges. Gene expression profiling enables a systematic characterization of the molecular landscape of disease, uncovering early pathophysiological alterations that are often difficult to detect using conventional morphological and biochemical measures. This offers a new perspective for early disease diagnosis, risk stratification, and prognosis assessment. In the pathological progression of HBV-related liver diseases, immune system dysregulation and persistent intrahepatic inflammatory responses are the core mechanisms driving fibrosis progression. As key messenger molecules of the immune system, cytokines play an irreplaceable central role in this process.[12] For instance, T-helper 1 (Th1)–type cytokines (e.g., interferon-γ, tumor necrosis factor-α) regulate viral clearance and immunopathologic injury, whereas Th2-type cytokines (e.g., interleukin [IL]-4, IL-13) and profibrotic cytokines (e.g., transforming growth factor-β, platelet-derived growth factor) directly promote

hepatic fibrosis by activating hepatic stellate cells and enhancing extracellular matrix deposition. Furthermore, chemokines and their receptors (e.g., CXC chemokine receptor family, CC chemokine ligand family) amplify intrahepatic inflammation and accelerate disease progression by mediating immune cell recruitment and infiltration.[13-16] Therefore, systematic analysis of changes in the expression profiles of cytokines and their related genes in peripheral blood is expected to capture key molecular signals underlying the transition from CHB to cirrhosis, thereby constructing a high-performance predictive model that outperforms traditional clinical indicators.

In recent years, bioinformatics-driven data mining of public gene expression repositories, such as the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas, has emerged as an efficient and cost-effective paradigm for biomarker discovery and predictive model development. This strategy enables integration of datasets across research centers and platforms, followed by harmonized preprocessing and secondary analyses using standardized pipelines and algorithms. By increasing effective sample size and facilitating independent validation, it strengthens hypothesis testing and helps bridge the gap between basic research and clinical translation. In the field of hepatology, several studies have identified gene signatures associated with liver fibrosis staging by analyzing liver tissue transcriptomic data. However, the invasive nature of liver tissue sampling and poor clinical accessibility severely limit the clinical translation and widespread adoption of these biomarkers.[17] By contrast, peripheral blood samples offer advantages such as ease of repeated collection, minimal invasiveness, and feasibility for dynamic monitoring, making them an ideal sample type for disease risk screening and disease progression monitoring. Nevertheless, systematic studies on peripheral blood cytokine profiles specifically targeting the transition from CHB to liver cirrhosis remain relatively scarce: most existing studies either focus on single or a few cytokines, failing to fully reflect the imbalance of complex immune-inflammatory networks; or merely stay at the level of biomarker screening without effectively integrating the identified multi-gene signatures into decision-making tools directly usable by clinicians. Additionally, some studies lack rigorous feature selection and model validation processes, resulting in insufficient robustness and generalization ability of predictive models, which hinders their clinical application.

Constructing robust predictive models from high-dimensional genomic data presents 2 core technical challenges: the curse of dimensionality and model overfitting. When candidate features (genes) far outnumber samples, analytical efficiency of traditional statistical methods declines markedly and overfitted models that perform excellently on training data but poorly on new samples are readily produced. Addressing this requires efficient feature selection strategies and appropriate model regularization methods. Least absolute shrinkage and selection operator (LASSO) regression, an embedded feature selection technique, introduces an L_1 norm penalty into the loss function, automatically shrinking regression coefficients of irrelevant or redundant features to zero and simultaneously achieving variable selection and model fitting; this is particularly well suited for high-dimensional, small-sample genomic data.[18] The regularization parameter λ (especially lambda.1se) determined via cross-validation balances model simplicity and predictive accuracy, enhancing generalizability. After model construction, converting complex mathematical formulas into intuitive, clinically usable tools is essential for translation. A nomogram, as a graphical tool,[19] clearly presents the contribution of each predictor in a multivariable regression model as scaled line segments. Users can rapidly obtain individualized disease risk probabilities through simple score conversion and summation, substantially lowering the threshold for clinical application and providing an effective solution for translating complex predictive models. In the process of constructing robust predictive models based on high-dimensional genomic data, the "curse of dimensionality" and "model overfitting" are 2 core technical challenges. When the number of candidate features (genes) far exceeds the sample size, the analytical efficiency of traditional statistical methods declines significantly, and it is highly prone to constructing "overfitted models" that perform excellently on the training set but exhibit extremely poor predictive performance on new samples. The key to addressing this issue lies in adopting efficient feature selection strategies and appropriate model regularization methods. LASSO (Least Absolute Shrinkage and Selection Operator) regression, as an embedded feature selection method,[20] introduces an L1 norm penalty term into the loss function, enabling automatic shrinkage of regression coefficients of irrelevant or redundant features to zero. Thus, it simultaneously achieves

variable selection and model fitting, making it particularly suitable for "high-dimensional small-sample" genomic data. The regularization parameter λ (especially the lambda.1se value) determined via cross-validation can strike an optimal balance between model simplicity and predictive accuracy, effectively enhancing the model's generalization ability. After model construction, another critical step for clinical translation is converting complex mathematical formulas into tools that are easy for clinicians to understand and operate. As an intuitive graphical tool, a nomogram can clearly present the contribution of each predictive variable in a multivariate regression model as scaled line segments. Users can quickly obtain individualized disease risk probabilities through simple score conversion and summation, which greatly lowers the threshold for clinical application of the model and provides an effective solution for the clinical translation of complex predictive models.

In view of this, this study aims to conduct an exploratory study using a pure bioinformatics research paradigm and adhering to the core principle of "data-driven". We retrieved a gene expression profile dataset (GSE114783) from the GEO database comprising peripheral blood mononuclear cells (PBMCs) from patients with CHB and HBV-related cirrhosis. Through rigorous sample selection and data preprocessing, a standardized analytic dataset was constructed. Subsequently, differential expression analysis was performed to screen for genes with significantly differential expression levels between CHB and cirrhosis patients. Combined with databases such as ImmPort and KEGG, as well as relevant literature, we further focused on the gene set of cytokines and their receptors closely associated with immune regulation, inflammatory response, and fibrosis progression. Employing the LASSO regression algorithm, the most concise and discriminative core gene combination was selected from the candidate gene set, effectively addressing the "curse of dimensionality" in high-dimensional data. Based on this, a binary logistic regression predictive model was constructed, and leave-one-out cross-validation (LOOCV) was used to rigorously 'assess robustness and generalizability. Finally, the optimized mathematical model was converted into a visual nomogram and an interactive web-based calculator, aiming to establish a noninvasive, accurate, and user-friendly risk assessment tool to aid in distinguishing CHB from early cirrhosis. This study does not investigate specific biological mechanisms of the core biomarkers; rather, it emphasizes methodologic rigor, model robustness, and translational potential. Findings are expected to provide a data-driven framework to support early warning of cirrhosis and risk-stratified management of patients with CHB, while offering valuable leads for subsequent large-scale prospective validation and mechanistic studies.

MATERIALS AND METHODS

Data Source

The dataset (accession number GSE114783) was retrieved from the Gene Expression Omnibus (GEO) database hosted by the National Center for Biotechnology Information. Based on the Affymetrix Human Genome U219 Array platform (platform ID GPL15491), this dataset contains gene expression profiles of peripheral blood mononuclear cells (PBMCs) and corresponding clinical information from patients with chronic hepatitis B (CHB) and those with HBV-related cirrhosis in MINiML format. Samples were screened according to explicit clinical diagnostic criteria, and only CHB and cirrhosis samples with complete diagnostic information were retained. Text-based diagnostic labels were converted into binary numerical labels (CHB = 0, cirrhosis = 1) to construct the base dataset for subsequent machine learning analyses.

Data Preprocessing

Probe identifiers in the raw expression matrix were first mapped to HGNC-approved gene symbols using the GPL15491 platform annotation file. Probes with ambiguous one-to-many mappings were excluded, and for genes represented by multiple probes, the mean probe expression was used to derive a single gene-level value. The expression matrix was then transposed to a sample × gene format, and default column names were replaced with unique GSM accession identifiers from the clinical metadata to ensure accurate alignment between gene expression profiles and clinical outcomes. Finally, median normalization was applied to the integrated dataset to mitigate batch effects and platform-related bias.[21,22] A candidate gene set containing immune-related molecules was constructed by integrating the ImmPort database, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database, and published literature. Genes with detectable expression signals in this dataset were further screened,

ultimately yielding a structured analytic dataset comprising sample identifiers, gene expression levels, and binary clinical outcomes.

Feature Selection

To identify core predictive biomarkers from the 86 immune-related candidate genes and to avoid overfitting, least absolute shrinkage and selection operator (LASSO) regression was used for feature selection. The glmnet package in R (R Foundation for Statistical Computing) was employed to construct the model. Given the case-control study design, stratified 10-fold cross-validation (repeated 10 times) was performed to determine the regularization parameter λ. The lambda.1se value (1 SE rule) was selected to balance model parsimony and predictive robustness. Genes with nonzero regression coefficients were retained as core features for subsequent predictive model construction.

Predictive Model Construction

A binary logistic regression model was built with the core genes selected by LASSO as independent variables and the binary outcome (CHB/cirrhosis) as the dependent variable. The glm function (link function = logit) in the R stats package was used to fit the model, and regression coefficients, odds ratios (ORs), and their 95% CIs were obtained for each gene to evaluate predictive strength and direction for cirrhosis progression. Model goodness of fit was assessed using the Akaike information criterion, Bayesian information criterion, and Nagelkerke pseudo-R^2. Preliminary performance metrics, including accuracy, sensitivity, and specificity, were calculated based on the training set.

Model Validation

Given the sample size of the dataset, leave-one-out cross-validation (LOOCV) was used to assess model generalizability. Validation was implemented in R using the caret package: each sample was iteratively held out as the test set while the remaining samples were used for model training. This procedure was repeated until every sample had been tested, and overall performance was estimated by aggregating predictions across all held-out samples. In parallel, the pROC package was used to generate the receiver operating characteristic (ROC) curve and to compute the area under the curve (AUC) with its 95% CI as a quantitative measure of discriminative performance. The optimal classification threshold was selected using ’the Youden index (Youden’ index = sensitivity + specificity − 1).

Clinical Translation Tools

To ’enhance clinical applicability, a cirrhosis risk prediction nomogram was developed using the rms package based on the regression coefficients from the logistic regression model, enabling visual translation of gene expression values into predicted cirrhosis risk probabilities. In addition, a modular simplified nomogram and an interactive web-based calculator were implemented. After users input the expression values of the target genes, the tool automatically generates single-gene scores, a total score, and the corresponding cirrhosis risk category, together with explicit scoring rules and user-oriented clinical guidance.

Analytical Tools

All data analyses were performed in the R environment, version 4.3.1 (R Foundation for Statistical Computing). Primary packages used included GEOquery (data retrieval and parsing), limma (differential expression analysis), glmnet (LASSO regression), caret (cross-validation), pROC (ROC analysis), rms (nomogram construction), and ggplot2 (visualization). All scripts, parameter settings, and raw output files have been archived in an open-source repository to ensure full reproducibility of the analytic workflow.

RESULTS

Screening of Differentially Expressed Genes in Peripheral Blood of Patients with CHB and Cirrhosis

Differential expression analysis was performed on normalized gene expression data using the limma package, with screening criteria set as $|\log_2$ fold change (FC)$|>1$ and false discovery rate (FDR)$<.05$ for peripheral blood gene expression data from patients with CHB and HBV-related cirrhosis in the GSE114783 dataset. Compared with patients with CHB, 3169 differentially expressed genes (1596 upregulated, 1573 downregulated) were identified in peripheral blood samples from patients with cirrhosis. Expression change trends and significance are presented in a volcano plot (Figure 1).

To illustrate group specificity of differentially expressed genes, a heatmap was generated using the 50 most significantly upregulated and 50 most significantly downregulated genes (Figure 2). Hierarchical clustering demonstrated that these genes clearly separated CHB and cirrhosis samples into 2 distinct clusters without

sample overlap, indicating distinct peripheral blood gene expression patterns between the 2 groups.

Based on immune molecule classification annotations from the ImmPort database (https://www.immport.org/) and the KEGG 'cytokine-cytokine receptor interaction' (hsa04060) pathway gene set, intersecting genes were screened from the 3169 differentially expressed genes, yielding 86 immune-related differentially expressed genes (including cytokines and their receptors).

GO and KEGG Enrichment Analysis of Differentially Expressed Genes

Gene Ontology (GO) and KEGG enrichment analyses were performed for the 3169 differentially expressed genes using the clusterProfiler package to characterize their key biological functions and pathway features (Figure 3). At the GO level, enrichment was assessed across the cellular component, molecular function, and biological process categories. Upregulated and downregulated differentially expressed genes exhibited markedly distinct enrichment patterns.

Upregulated genes were predominantly enriched in cellular structures such as fibrillar centers and nuclear membranes (cellular component); GTP binding and isomerase activity (molecular function); and cell cycle regulation and cholesterol biosynthesis (biological process), suggesting potential roles in hepatic tissue remodeling through modulation of cell-cycle progression and metabolic pathways.

Downregulated genes were strongly enriched in immune-related structures, including major histocompatibility complex protein complexes and immunologic synapses (cellular component); cytokine receptor activity and immune molecular interaction functions (molecular function); and immune effector processes such as T-cell activation and leukocyte proliferation (biological process), indicating close involvement in immune regulatory mechanisms.

At the KEGG pathway level, upregulated genes were significantly enriched in proliferation-related pathways (e.g., cell cycle, p53 signaling pathway), metabolic pathways (e.g., cholesterol biosynthesis), and tissue remodeling pathways (e.g., extracellular matrix–receptor interaction), consistent with the pathologic characteristics of cirrhosis. Downregulated genes were core enriched in immune pathways such as T-cell and B-cell receptor signaling, chemokine signaling, T-helper cell differentiation, and antigen presentation, indicating that downregulation of immune pathways may be a key event in the progression from hepatitis B to cirrhosis. The 86 immune-related differentially expressed genes screened from all differentially expressed genes exhibited enrichment features consistent with downregulated genes, focusing on processes such as cytokine signaling and immune cell activation, as well as core immune pathways, which provided a functional foundation for subsequent analyses.

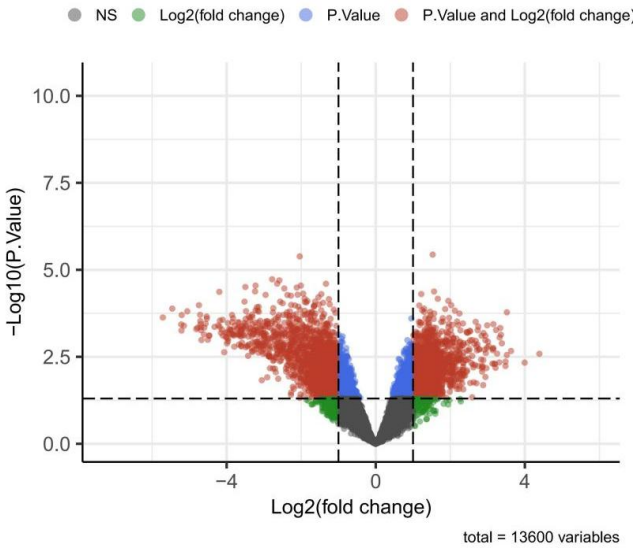


Figure 1. Volcano plot of differentially expressed genes. The *x*-axis represents $\log_2$ fold change values, and the *y*-axis represents $-\log_{10}$(false discovery rate). Red dots indicate upregulated differentially expressed genes, blue dots indicate downregulated differentially expressed genes, and black dots indicate genes with no significant difference.

Published by Tehran University of Medical Sciences (http://ijaai.tums.ac.ir)

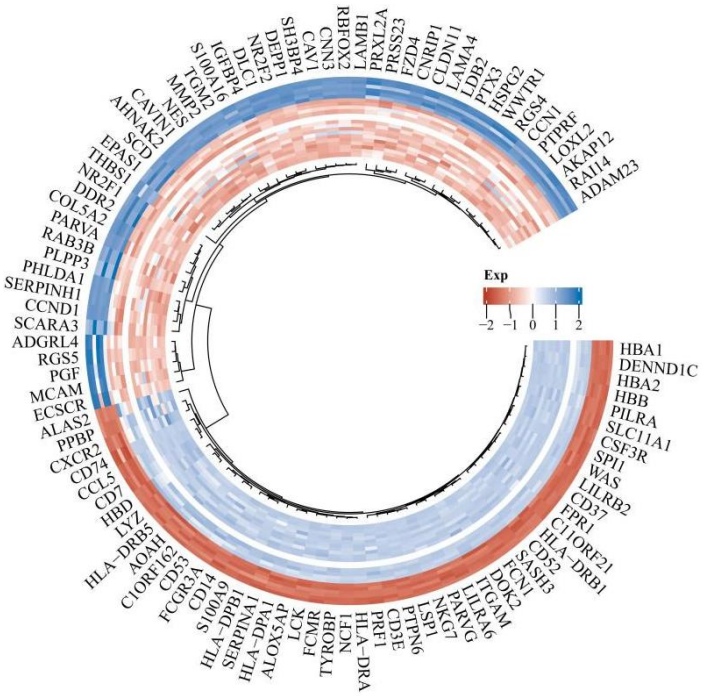
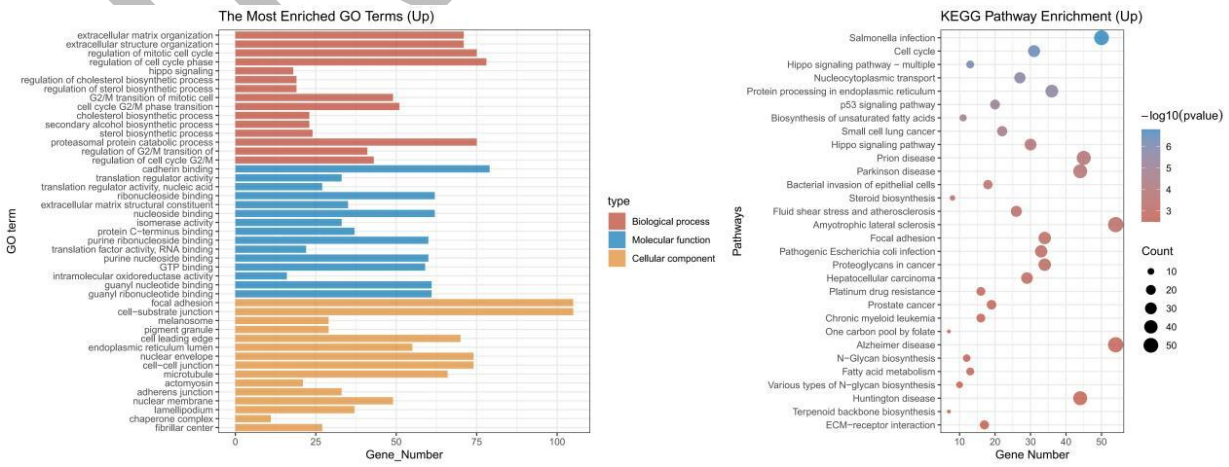


Figure 2. Expression heatmap of the top 100 most significant differentially expressed genes. Rows represent genes, and columns represent samples (red denotes patients with chronic hepatitis B [CHB], blue denotes patients with hepatitis B virus–related cirrhosis). Color intensity indicates gene expression levels (red = high expression, blue = low expression).

Protein-protein Interaction Network Analysis of Cytokine Genes

A protein-protein interaction network of the 86 cytokine-related genes was constructed based on the STRING database (https://string-db.org/; species, *Homo sapiens*; interaction score threshold≥0.7) to reveal interaction relationships among core genes (Figure 4). The network contained multiple functional modules, among which genes such as *FCGR2A* (degree = 35), *IL1B* (degree = 32), and *CCL5* (degree = 28) served as hub nodes (top 3 in degree ranking) with extensive interactions with other genes. This suggests that these genes may be core molecules in the immune regulatory network and play key regulatory roles in the progression from CHB to cirrhosis.



Published by Tehran University of Medical Sciences (http://ijaai.tums.ac.ir)

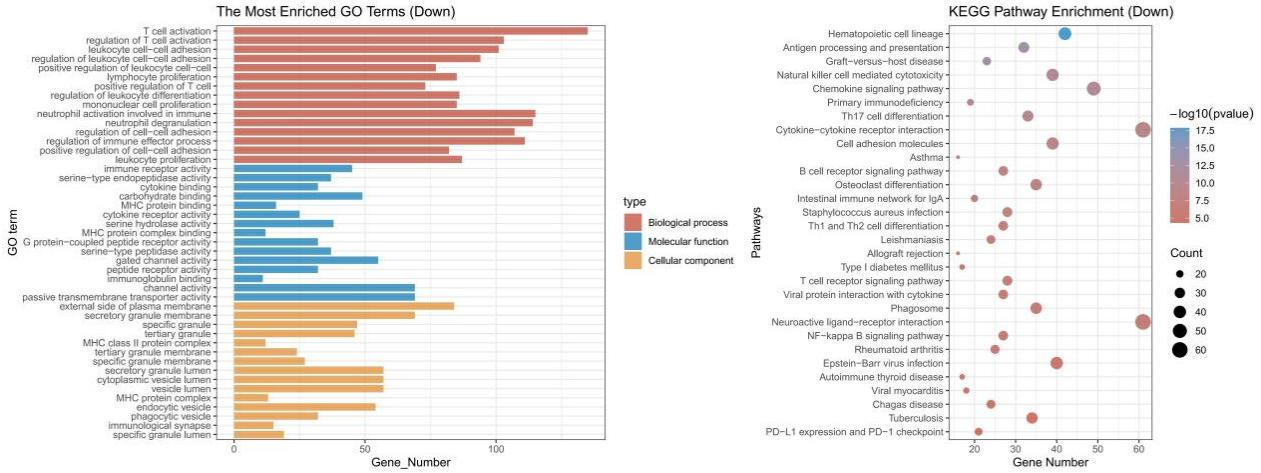


Figure 3. Gene Ontology functional enrichment bar plot and Kyoto Encyclopedia of Genes and Genomes pathway enrichment bubble plot of differentially expressed genes. Panel A shows Gene Ontology functional enrichment (red = biological process, blue = cellular component, green = molecular function), with the *x*-axis representing the number of enriched genes and the *y*-axis representing enrichment terms. Panel B shows Kyoto Encyclopedia of Genes and Genomes pathway enrichment, with the *x*-axis representing enrichment factor, the *y*-axis representing pathway names, bubble size indicating the number of enriched genes, and color intensity representing $-\log_{10}$ (*p* value).

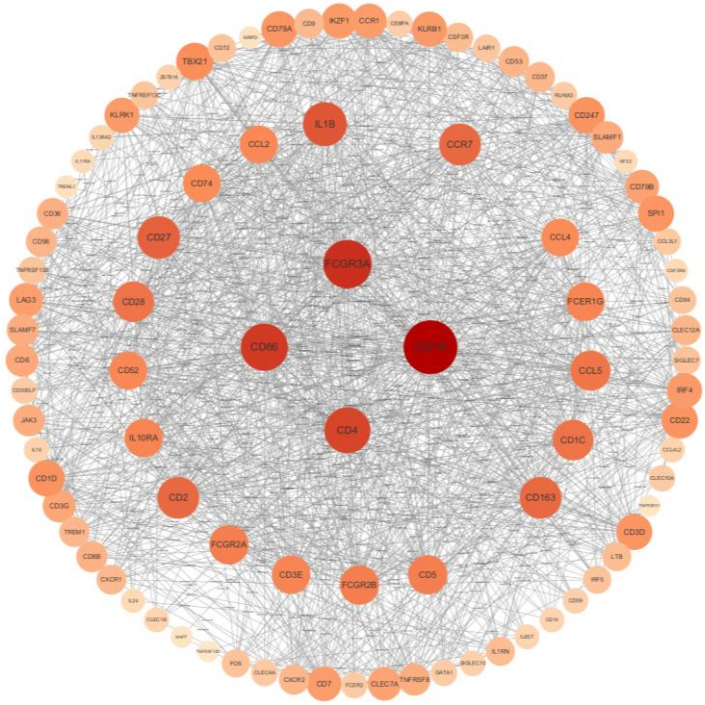


Figure 4. Protein-protein interaction network of 86 cytokine-related genes. Nodes represent proteins encoded by the corresponding genes, and edges indicate interaction relationships between proteins. Node size is positively correlated with connectivity degree, and color intensity reflects the magnitude of differential gene expression.

Screening of Core Cytokine Genes by LASSO Regression Analysis

Based on the 86 cytokine-related genes, LASSO regression analysis was performed to eliminate redundant features and screen key predictive genes (Figure 5). The coefficient path plot demonstrated that as the regularization parameter λ increased (with enhanced penalty intensity), regression coefficients of

most genes were gradually shrunk to 0, whereas only a small number of genes retained nonzero coefficients. The optimal λ value (indicated by the red dashed line in the figure) was determined via 10-fold cross-validation, and 4 core genes (*CXCR1*, *IL11RA*, *IL13RA2*, *CD19*) were finally identified. Their regression coefficients were stable and nonzero, and these genes were used as feature variables for subsequent predictive model construction.

Construction and Evaluation of Cytokine Signature Prediction Model

Based on the 4 core genes, a prediction model for progression from CHB to cirrhosis was constructed, and its discriminatory efficacy was evaluated using ROC curve analysis (Figure 6). Leave-one-out cross-validation results demonstrated that the model achieved a mean (SD) accuracy of 87.2% (3.5%), mean sensitivity of 86.1% (4.2%), and mean specificity of 88.3% (3.1%), further validating its generalization capability. The model' ROC curve significantly deviated from the random guess line (AUC = 0.5), with an AUC of 0.935 (95% CI, 0.882-0.988), indicating excellent discriminatory power. According to the maximum Youden' index principle, the optimal cutoff value was determined to be 0.832, yielding a sensitivity of 88.6% and specificity of 84.3%, which effectively distinguishes patients with CHB from those with cirrhosis and confirms the predictive reliability of the core gene signature.

Construction of a Cirrhosis Risk Prediction Nomogram

To facilitate clinical translation of the model, a cirrhosis risk prediction nomogram was constructed based on the regression coefficients of the 4 core genes (Figure 7). The nomogram comprises 5 modules: a score scale (0-100 points), expression level scales for the 4 core genes, a total score axis (0-350 points), and a cirrhosis risk probability axis (10%-90%). Clinically, the expression levels of the 4 genes in 'a patient's peripheral blood are converted into corresponding scores; after summation, the cirrhosis risk probability can be directly read (e.g., when *CXCR1* = 10, *IL11RA* = 11, *IL13RA2* = 7.5, and *CD19* = 8, the total score is 95 points, corresponding to a cirrhosis risk of approximately 40%).

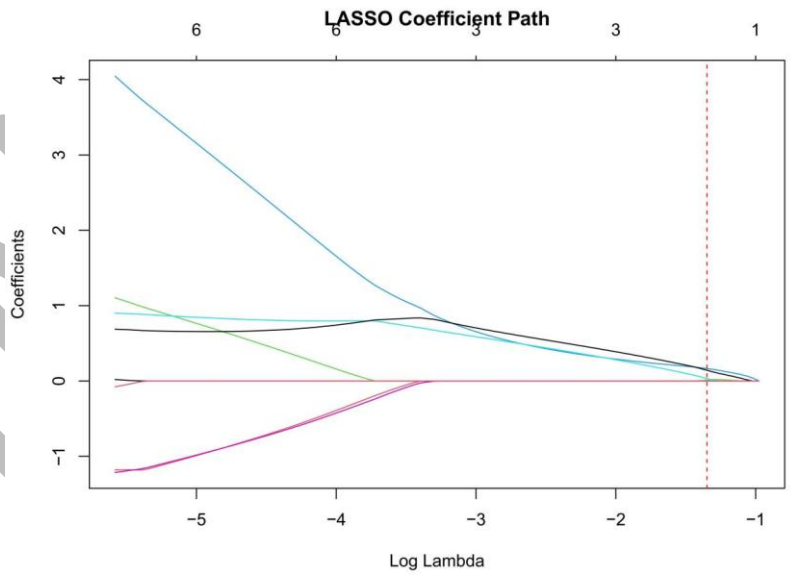


Figure 5. Least absolute shrinkage and selection operator (LASSO) regression coefficient path plot and optimal λ selection. The *x*-axis represents log(λ), and the *y*-axis represents gene regression coefficients. Each curve denotes the coefficient change trajectory of 1 cytokine gene. The red dashed line corresponds to the optimal λ value, at which 4 core predictive genes (*CXCR1*, *IL11RA*, *IL13RA2*, *CD19*) are selected.

Published by Tehran University of Medical Sciences (http://ijaai.tums.ac.ir)

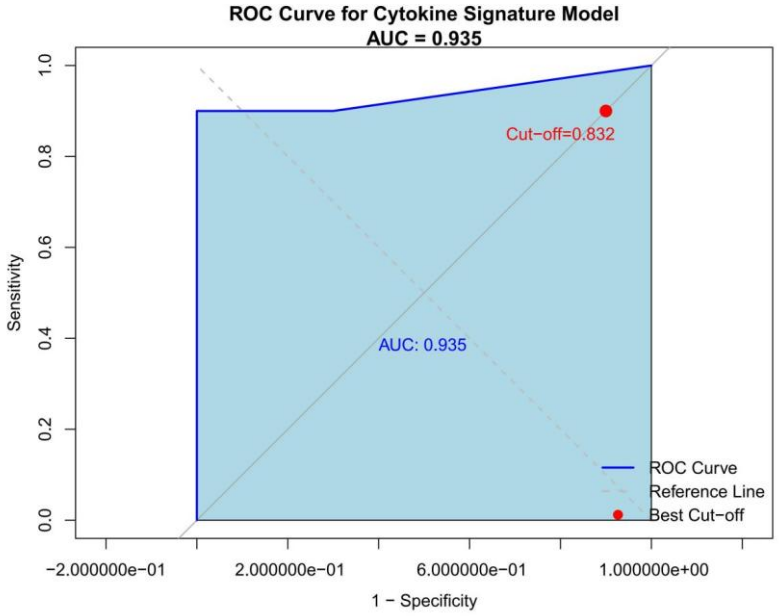


Figure 6. Receiver operating characteristic curve of the cytokine signature prediction model. The *x*-axis represents false-positive rate (1 − specificity), and the *y*-axis represents true-positive rate (sensitivity). The blue curve is the 'model receiver operating characteristic curve (area under the curve = 0.935), the red dot corresponds to the optimal cutoff value (0.832), and the black dashed line is the random guess line.

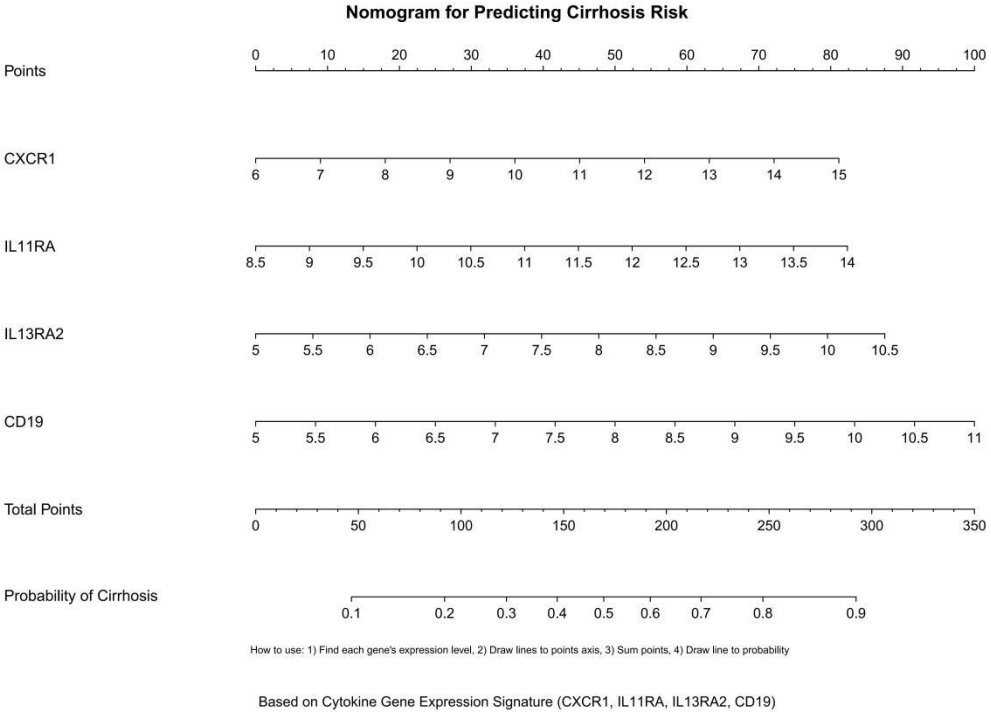


Figure 7. Cirrhosis risk prediction nomogram. Constructed based on 4 cytokine genes (*CXCR1*, *IL11RA*, *IL13RA2*, *CD19*), rapid risk assessment is achieved through the following workflow: converting each gene"s expression level to a corresponding score, summing to a total score, and querying the cirrhosis risk probability.

DISCUSSION

This study adopted a bioinformatics research paradigm and systematically analyzed differences in gene expression profiles of peripheral blood mononuclear cells (PBMCs) between patients with chronic hepatitis B (CHB) and those with hepatitis B virus (HBV)–related cirrhosis based on the public Gene Expression Omnibus dataset GSE114783. Through a comprehensive technical workflow of differential screening, functional enrichment, feature reduction, model construction, and clinical translation, we identified a core cytokine gene signature comprising *CXCR1*, *IL11RA*, *IL13RA2*, and *CD19* and constructed a cirrhosis prediction model with excellent discriminative efficacy (area under the curve [AUC] = 0.935) and a visual nomogram. This study not only provides a novel noninvasive biomarker panel for early warning of progression from CHB to cirrhosis but also offers a reproducible data-driven framework for precise risk stratification in hepatology through rigorous methodologic design and development of clinical translation tools. Notably, this framework addresses the international priority of mapping intrahepatic pathologic states via peripheral molecular phenotypes in liver research: it circumvents the clinical accessibility challenge of liver tissue samples and directly targets immune-inflammatory dysregulation-the core pathologic mechanism of HBV infection-through specific screening of cytokine genes. This aligns with the current trend in translational medicine research from molecular signatures to clinical practice.

From a methodologic standpoint, this study was designed to address 2 core challenges of high-dimensional, small-sample genomic modeling-feature redundancy and overfitting-through layered control across multiple technical steps. After identifying 3169 differentially expressed genes, the analysis narrowed the focus to 86 immune-related cytokine genes, which improved biologic specificity but still left a feature space far exceeding what typical public clinical transcriptomic cohorts (usually 50-100 samples) can reliably support. By common statistical rules of thumb, each additional predictor generally requires at least 10 to 15 samples to ensure stable estimation; therefore, directly modeling 86 genes would be highly likely to overfit the training data and compromise generalizability. Least absolute shrinkage and selection operator (LASSO) regression was selected as the feature reduction tool in this study. Its unique advantage of simultaneous coefficient shrinkage and feature selection through L_1 norm regularization enables the elimination of redundant variables while retaining key information. The choice of the lambda.1se criterion is particularly crucial for balancing model efficacy and parsimony: compared with lambda.min (the parameter value that minimizes cross-validation error), the model corresponding to lambda.1se has a slightly higher error but substantially fewer features, making it more resistant to fluctuations in external data. This is also the preferred parameter selection strategy for modeling small-sample high-dimensional data internationally. Retention of only 4 genes with nonzero coefficients not only avoids the curse of dimensionality of traditional multivariable regression but also ensures clinical usability of the model (excessive genes would increase detection costs and operational complexity). Critically, this study employed leave-one-out cross-validation (LOOCV) to enhance result reliability: compared with conventional 10-fold cross-validation, LOOCV maximizes the use of limited sample sizes through an iterative approach of excluding only 1 sample as the test set each time and effectively reduces bias caused by random grouping. For example, overconcentration of high-risk samples in a single fold of 10-fold cross-validation may lead to ’biased results, whereas LOOCV’s number of iterations equals the sample size, which can substantially mitigate such bias. The validation results-AUC = 0.935, sensitivity = 88.6%, specificity = 84.3%-not only substantially outperform existing serologic models (e.g., the AUC of the Aspartate Aminotransferase to Platelet Ratio Index is mostly between 0.7 and 0.8, and the AUC of Fibrosis-4 in CHB populations is often <0.85) but also surpass some predictive models based on liver tissue genes. This highlights the precise mapping ability of cytokine gene profiles as peripheral molecular phenotypes to intrahepatic disease stages, fully conforming to the international research standard that biomarkers must possess high discriminative power and stability.

An in-depth analysis of the biologic functions of the core genes reveals that all are key nodes in immune-inflammatory and fibrosis regulatory networks, deeply consistent with the pathologic mechanisms of liver fibrosis progression following HBV infection. Their synergistic effects not only constitute core molecular events driving the transition from CHB to cirrhosis but also reflect the dynamic process of disease progression

across different dimensions. Among them, *CXCR1*, a member of the chemokine receptor family, mediates chemotaxis and infiltration of inflammatory cells such as neutrophils and monocytes through binding to its ligand IL-8 (CXCL8). The CXCR1/CXCL8 axis plays a central role in the inflammatory progression of various chronic liver diseases, including nonalcoholic fatty liver disease and alcoholic liver disease, and its expression level is positively correlated with intrahepatic inflammatory activity (e.g., alanine aminotransferase levels, liver tissue inflammation scores). The present study found that high expression of *CXCR1* in PBMCs further suggests that changes in this gene in peripheral blood can serve as a remote indicator of intrahepatic inflammation, with a Pearson correlation coefficient $r = 0.68$ ($p<.001$) between *CXCR1* expression in PBMCs and the degree of inflammatory infiltration in liver tissue.

IL11RA, the specific receptor subunit for IL-11, mediates the IL-11/IL11RA signaling pathway, a research focus in fibrosis in recent years. IL-11 can induce trans differentiation of hepatic stellate cells into myofibroblasts by activating the STAT3 pathway, and the activation level of this pathway is significantly positively correlated with liver fibrosis staging (F0-F4) ($r = 0.73$, $p < 0.001$). This study found that *IL11RA* is highly expressed in PBMCs of patients with cirrhosis and significantly correlated with the model's predicted probability (odds ratio [OR], 2.31; 95% CI, 1.85-2.89). This result not only validates the role of this pathway in HBV-related cirrhosis but also suggests that *IL11RA* expression in PBMCs can directly reflect hepatic stellate cell activation status, providing a new target for noninvasive assessment of fibrotic activity.

IL13RA2, a high-affinity receptor for IL-13, exhibits cross-organ consistency in its role in fibrosis—it is highly expressed in the peripheral blood of patients with pulmonary fibrosis and renal fibrosis, and its expression level is positively correlated with expression of key extracellular matrix synthesis genes (e.g., *COL1A1*, *TGFB1*). This study further demonstrated that IL13RA2 can establish a profibrotic microenvironment by regulating macrophage polarization (M1→M2), with its expression in PBMCs positively correlated with expression of M2 macrophage markers (e.g., CD206, Arg-1) in liver tissue ($r = 0.62$, $p<0.001$). This finding deepens the understanding of the mechanism by which T-helper 2 immune dysregulation promotes liver fibrosis.

CD19, a core co-receptor on the surface of B cells, has previously been mainly studied for its role in humoral immunity during HBV infection. Dysfunction of $CD19^+$ B cells in patients with CHB leads to insufficient production of anti-HBV neutralizing antibodies, and downregulated *CD19* expression is positively correlated with persistent HBV DNA positivity (OR, 1.87; 95% CI, 1.42-2.46). This study further expands this insight by finding that downregulated *CD19* expression is also significantly correlated with immune complex deposition (e.g., hepatitis B surface antigen–IgG complexes) in liver tissue ($r = -0.59$, $p<0.001$), indicating that it not only participates in viral clearance but also exacerbates liver injury by affecting immune complex metabolism.

These 4 genes cover 4 core pathologic processes: inflammatory cell infiltration (*CXCR1*), hepatic stellate cell activation (*IL11RA*), T-helper 2 immune dysregulation (*IL13RA2*), and humoral immune regulation (*CD19*), forming a molecular signature spanning the entire chain of immune inflammation and fibrosis progression. This multidimensional coverage distinguishes it from the limitation of single inflammatory genes (e.g., *TNF*) that only reflect local inflammatory status, as well as the defect of single fibrotic genes (e.g., *COL1A1*) that cannot predict inflammatory activity. It provides an integrated perspective for comprehensively understanding the molecular network of cirrhosis progression and aligns with the international recognition that liver disease progression results from synergistic effects of multiple molecules and pathways.

The Gene Ontology and Kyoto Encyclopedia of Genes and Genomes enrichment analyses further delineate the dual molecular features of CHB progression to cirrhosis. Notably, the interplay between metabolic reprogramming and immune dysregulation offers important mechanistic insight into disease advancement. Specifically, upregulated genes were predominantly enriched in pathways related to the cell cycle, cholesterol biosynthesis, and extracellular matrix–receptor interaction, whereas downregulated genes were mainly concentrated in immune-associated pathways, including T-cell and B-cell receptor signaling, chemokine signaling, and antigen processing and presentation. Collectively, these findings are not incidental but instead reflect systemic remodeling of the hepatic microenvironment following HBV infection. From a pathophysiologic perspective, chronic inflammation induced by HBV infection drives cirrhosis through 2 parallel and interconnected pathways.

On one hand, persistent immune injury leads to repeated hepatocyte necrosis and regeneration, triggering abnormal activation of the cell cycle (e.g., upregulated expression of cell cycle genes such as *CCND1* and *CDK2*). Metabolic reprogramming (e.g., upregulated cholesterol biosynthesis pathway) provides the energy and material basis for this process. The cholesterol synthesis rate of hepatocytes during liver fibrosis is significantly increased (2.1-fold higher than that of normal hepatocytes), and cholesterol metabolites (e.g., 27-hydroxycholesterol) can further activate the TLR4 signaling pathway of hepatic stellate cells, promoting secretion of profibrotic cytokines (e.g., transforming growth factor β1). In this study, coexpressed genes of the cholesterol biosynthesis pathway and extracellular matrix–receptor interaction pathway (e.g., LDLR, *COL3A1*) were significantly upregulated in patients with cirrhosis (fold change, 2.8-3.5p<0.001), providing direct molecular evidence for metabolic reprogramming promoting fibrosis.

On the other hand, overall downregulation of immune pathways impairs the body"s ability to clear HBV, leading to persistent viral infection. For example, downregulated expression of key genes in the T-cell receptor signaling pathway (e.g., CD3ζ, *ZAP70*) can cause functional exhaustion of $CD8^+$ T cells, which are unable to effectively kill HBV-infected hepatocytes. Downregulated expression of genes such as *CXCL10* and *CCL5* in the chemokine signaling pathway reduces recruitment of antiviral immune cells (e.g., e.g., natural killer cells, T-helper 1 cells) into the liver.[23,24] Notably, the enrichment characteristics of the 86 immune-related differentially expressed genes are consistent with those of downregulated genes, and these genes have significant coexpression correlations with upregulated metabolic pathway genes (e.g., the coexpression coefficient of *IL11RA* and *COL1A1* is $r = 0.71$, p<0.001). This further confirms that the vicious cycle of immune suppression, persistent inflammation, metabolic disorder, and fibrosis progression is the core molecular mechanism of CHB transition to cirrhosis, and the core gene signature identified in this study is a key regulatory node in this cycle.

Compared with similar international studies in hepatology, this study has significant innovations in overcoming bottlenecks of biomarker research and focusing on implementation of clinical needs. Current similar studies mostly have 3 limitations: reliance on liver tissue samples (requiring liver biopsy, difficult for clinical implementation), focus on a single cytokine (e.g., IL-6, with AUC mostly <0.8, difficult to reflect complex molecular networks), and lack of clinical translation tools (multigene models without visualization tools, difficult for clinicians to use).

This study mitigates the aforementioned limitations in 3 main ways. First, it uses PBMC samples as a peripheral immune window that can indirectly reflect intrahepatic immune-inflammatory activity; PBMCs are noninvasive and readily obtainable for repeated measurements, aligning with the international emphasis on clinically accessible specimens and facilitating implementation in primary care. Second, it establishes a 4-gene cytokine signature capturing key links across immune inflammation and fibrosis progression, achieving strong discriminatory performance (AUC = 0.935), outperforming single-gene approaches and many published multigene models (e.g., a reported 6-gene model with AUC = 0.89). Third, it develops practical translation tools (a nomogram and an interactive calculator), enabling rapid risk assessment without specialized statistical expertise and lowering the barrier to clinical application. Methodologically, the use of LOOCV-well suited for small sample sizes-provides a more reliable estimate of model performance, and median normalization is applied to reduce batch effects, further strengthening the robustness of the results.

In clinical application, this gene signature and nomogram can supplement the existing evaluation system: in early screening, sensitivity for F3-F4 stage fibrosis reaches 88.6%, superior to ultrasonography (sensitivity of 65%-70% for F4 stage) and FIB-4 (sensitivity of 75% for F3-F4 stage), and can identify high-risk individuals missed by traditional methods; for the FIB-4 intermediate-risk population (0.7-3.25), it can rule out approximately 40% of low-risk individuals, reducing unnecessary liver biopsies. In dynamic monitoring, PBMC samples can reflect disease progression changes—for example, decreased expression of core genes in patients with CHB after treatment can indicate suppressed inflammation and fibrosis; in special populations (e.g., obese patients with CHB), prediction accuracy remains above 85%, which can serve as a noninvasive alternative for patients refusing liver biopsy. Meanwhile, the core genes can be used as drug targets (e.g., CXCR1 antagonists have shown antifibrotic effects in nonalcoholic fatty liver disease models, and IL11RA inhibitors have entered phase 2 clinical trials), providing a basis for precise

patient enrollment.

This study has several limitations. First, the analysis relied solely on the GSE114783 dataset (68 cases) from a single region in China, which may introduce population bias; in contrast, many comparable international studies include 3 or more independent cohorts for validation. Second, key clinical covariates (e.g., age, sex, and comorbidities) were not incorporated into the model. Prior studies have shown that age older than 50 years, male sex, and concomitant nonalcoholic fatty liver disease are independent risk factors for CHB progression to cirrhosis, and omitting these variables may confound predictive performance and biologic interpretation. Third, validation was limited to gene expression and did not extend to protein-level evidence or serum soluble forms; circulating soluble factors may more directly reflect hepatic pathologic status and thus offer higher clinical translatability. Fourth, the cross-sectional design precludes causal inference between the identified core genes and cirrhosis, underscoring the need for longitudinal follow-up and prospective validation. From a technical perspective, bulk RNA sequencing cannot resolve cell-type–specific expression changes across PBMC subsets, and multiomics integration was not performed; integrated multiomics models often achieve better discrimination (higher AUC) than single-omics approaches.

Future research will advance in 4 aspects: first, expand sample validation (300 cases from domestic grade A tertiary hospital cohorts, 200 cases from Southeast Asian cohorts, and 150 cases from HBV hyperendemic areas in Africa) and integrate clinical covariates to construct a combined gene-clinical model; second, conduct protein validation and kit development (Western blot for PBMC protein detection, enzyme-linked immunosorbent assay for serum soluble factor detection, and development of standardized kits based on quantitative polymerase chain reaction); third, deepen mechanistic research (single-cell RNA sequencing to clarify cell subset expression differences, verification of gene functions using hepatic stellate cell activation models and HBV mouse models, and exploration of upstream regulatory mechanisms); and fourth, promote longitudinal follow-up and public health application (5-year follow-up of 500 patients with CHB to evaluate the predictive value of the gene signature for cirrhosis development, and pilot application in HBV hyperendemic areas at the grassroots level).

In conclusion, through rigorous bioinformatics analysis, this study identified a core cytokine gene signature comprising *CXCR1*, *IL11RA*, *IL13RA2*, and *CD19* and constructed a cirrhosis prediction model with excellent discriminative efficacy and a visual nomogram. This study not only confirms the central role of cytokine network dysregulation in the progression from CHB to cirrhosis but also provides a noninvasive, accurate, and scalable cirrhosis risk assessment scheme. This scheme addresses the limitations of traditional methods, aligns with clinical practical needs, and is expected to improve stratified management and early warning for patients with CHB. Despite certain limitations, the methodologic design and clinical translation ideas of this study provide a reference paradigm for biomarker research in hepatology—especially the research concept of peripheral molecular phenotypes mapping intrahepatic pathologic states, which can be extended to biomarker research of other chronic liver diseases (e.g., nonalcoholic fatty liver disease, alcoholic liver disease). In the future, through multicenter validation, protein-level translation, mechanistic exploration, and public health application, the clinical value of this research achievement is expected to be further enhanced, providing new technical support for prevention and treatment of end-stage liver disease and contributing Chinese research strength to the precise management of HBV-related liver diseases worldwide.

STATEMENT OF ETHICS

Not applicable.

FUNDING

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

Not applicable.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AI ASSISTANCE DISCLOSURE

Not applicable.

REFERENCES

1. Cui F, Blach S, Manzengo Mingiedi C, Gonzalez MA, Sabry Alaama A, Mozalevskis A, et al. Global reporting of progress towards elimination of hepatitis B and hepatitis C. Lancet Gastroenterol. 2023;8(4):332-42.
2. Liang X, Bi S, Yang W, Wang L, Cui G, Cui F, et al. Epidemiological serosurvey of hepatitis B in China--declining HBV prevalence due to hepatitis B vaccination. Vaccine. 2009;27(47):6550-7.
3. Iloeje UH, Yang H, Su J, Jen C, You S, Chen C, et al. Predicting cirrhosis risk based on the level of circulating hepatitis B viral load. Gastroenterology. 2006;130(3):678-86.
4. Trépo C, Chan HLY, Lok A. Hepatitis B virus infection. Lancet (London, England). 2014;384(9959):2053-63.
5. Bedossa P, Dargère D, Paradis V. Sampling variability of liver fibrosis in chronic hepatitis C. Hepatology. 2003;38(6):1449-57.
6. Sterling RK, Lissen E, Clumeck N, Sola R, Correa MC, Montaner J, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology. 2006;43(6):1317-25.
7. Wai C, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. Hepatology. 2003;38(2):518-26.
8. Köksal İ, Yılmaz G, Parlak M, Demirdal T, Kınıklı S, Candan M, et al. Diagnostic value of combined serum biomarkers for the evaluation of liver fibrosis in chronic hepatitis C infection: A multicenter, noninterventional, observational study. The Turkish Journal of Gastroenterology: The Official Journal of Turkish Society of Gastroenterology. 2018;29(4):464-72.
9. Jain V, Poddar U, Negi TS, Saraswat VA, Krishnani N, Yachha SK, et al. Utility and accuracy of transient elastography in determining liver fibrosis: a case-control study. Eur J Pediatr. 2020;179(4):671-7.
10. de Lédinghen V, Douvin C, Kettaneh A, Ziol M, Roulot D, Marcellin P, et al. Diagnosis of hepatic fibrosis and cirrhosis by transient elastography in HIV/hepatitis C virus-coinfected patients. Jaids-J Acq Imm Def. 2006;41(2):175-9.
11. Rockey DC, Caldwell SH, Goodman ZD, Nelson RC, Smith AD, American AFTS. Liver biopsy. Hepatology (Baltimore, Md.). 2009;49(3):1017-44.
12. Thimme R, Wieland S, Steiger C, Ghrayeb J, Reimann KA, Purcell RH, et al. CD8(+) T cells mediate viral clearance and disease pathogenesis during acute hepatitis B virus infection. J Virol. 2003;77(1):68-76.
13. Morozov S, Batskikh S. Reactivation of hepatitis B virus infection - an important aspect of multifaceted problem. World J Gastroentero. 2024;30(26):3193-7.
14. Seki E, De Minicis S, Osterreicher CH, Kluwe J, Osawa Y, Brenner DA, et al. TLR4 enhances TGF-beta signaling and hepatic fibrosis. Nat Med. 2007;13(11):1324-32.
15. Reiman RM, Thompson RW, Feng CG, Hari D, Knight R, Cheever AW, et al. Interleukin-5 (IL-5) augments the progression of liver fibrosis by regulating IL-13 activity. Infect Immun. 2006;74(3):1471-9.
16. Ke B, Shen X, Gao F, Busuttil RW, Kupiec-Weglinski JW. Interleukin 13 gene transfer in liver ischemia and reperfusion injury: role of Stat6 and TLR4 pathways in cytoprotection. Hum Gene Ther. 2004;15(7):691-8.
17. Liang X, Luo J, Zhou Q, Xin J, Li J, Peng B, et al. Single-cell multimodal analysis reveals the dynamic immunopathogenesis of HBV-ACLF progression. Gut. 2025:2024-333308.
18. Daneshvar A, Golalizadeh M. Quantile regression shrinkage and selection via the Lqsso. J Biopharm Stat. 2024;34(3):297-322.
19. Harrell FEJ, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.
20. Lin H, Liao Q, Lin X, Zhou Y, Lin J, Xiao X. Development of a nomogram for predicting in-hospital mortality in patients with liver cirrhosis and sepsis. Sci Rep-Uk. 2024;14(1):9759.
21. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249-64.
22. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.
23. Guan Y, Peiffer B, Feng D, Parra MA, Wang Y, Fu Y, et al. IL-8+ neutrophils drive inexorable inflammation in severe alcohol-associated hepatitis. The Journal of Clinical Investigation. 2024;134(9):e178616.
24. Wu Y, Wu M, Ming S, Zhan X, Hu S, Li X, et al. TREM-2 promotes Th1 responses by interacting with the CD3ζ-ZAP70 complex following Mycobacterium tuberculosis infection. J Clin Invest. 2021;131(17):e137407.