Original Article

Potential diagnostic marker gene set for non-alcoholic steatohepatitis associated hepatocellular carcinoma with lymphocyte infiltration

Xueyun Wang[1], Mengzhou Gao[1], Zexi Zhang[1], Xiang Ao[2], An Luo[1], Zhenguo Wen[1], Xingquan Pan[1], Mengge Sun[1], Teng Wang[1], Zhaojun Jia[1]

[1]Beijing Key Laboratory of Enze Biomass Fine Chemicals, College of New Materials and Chemical Engineering, Beijing Institute of Petrochemical Technology, Beijing, China; [2]School of Basic Medicine, Qingdao University, Qingdao, China

Contributions: (I) Conception and design: X Ao, A Luo, Z Wen, X Pan, M Sun, T Wang, Z Jia; (II) Administrative support: None; (III) Provision of study materials or patients: Z Jia; (IV) Collection and assembly of data: X Wang, M Gao, Z Zhang, A Luo; (V) Data analysis and interpretation: X Wang, M Gao, Z Zhang, A Luo; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Teng Wang, Pharmaceutical Analysis, PhD; Zhaojun Jia, Biochemical Engineering, PhD. Beijing Key Laboratory of Enze Biomass Fine Chemicals, College of New Materials and Chemical Engineering, Beijing Institute of Petrochemical Technology, 19 Qingyuan North Road, Daxing District, Beijing 102617, China. Email: wangteng@bipt.edu.cn; jiazj@bipt.edu.cn.

Background: Non-alcoholic steatohepatitis (NASH), a prominent driver of hepatocellular carcinoma (HCC) besides virus and alcohol, induces a series of complex liver structural and immune microenvironment changes, which make the early diagnosis and treatment of NASH-associated HCC (NASH-HCC) more challenging. This study aims to identify signature genes and explore the role of immune cell infiltration in NASH-HCC to improve early detection and prognosis assessment.

Methods: Differential gene and immune cell infiltration are important indicators for predicting the progress of oncology and responsiveness of tumor patients to immunotherapy, usually confirmed through biopsy tests with poor patient compliance. To obtain a highly correlated signature gene set and validate immune cell infiltration status, the GSE164760 and GSE102079 datasets from the Gene Expression Omnibus (GEO) database were analyzed using machine learning algorithms. Feature genes were identified based on differentially expressed genes and key modular genes identified by weighted gene co-expression network analysis (WGCNA). The signature genes were screened using the least absolute shrinkage and selection operator (LASSO), random forest, and support vector machine recursive feature elimination (SVM-RFE) machine learning algorithms. Subsequently, the signature genes were subjected to diagnostic efficacy tests, gene set enrichment analysis, immune cell infiltration assessment and real-time reverse transcription polymerase chain reaction (RT-qPCR) validation.

Results: Six signature genes were identified, including C-C motif chemokine ligand 14 (*CCL14*), C-type lectin domain family 4 member G (*CLEC4G*), ficolin-2 (L-ficolin, *FCN2*), insulin-like growth factor binding protein 3 (*IGFBP3*), C-X-C motif chemokine ligand 14 (*CXCL14*), and vasoactive intestinal polypeptide type I receptor (*VIPR1*). The area under the receiver operating characteristic (ROC) curve for the six signature genes was between 0.927–0.958, and the calibration curves also indicated that they had high prediction accuracy. Six signature genes were positively associated with NASH pathological process pathways including butyric acid metabolism and fatty acid degradation. The infiltration of immune cells such as M2-type macrophages was significantly positively correlated with the signature genes. RT-qPCR revealed a significant decrease in the expression of *CLEC4G* and *IGFBP3* in the NASH-HCC model.

Conclusions: *CLEC4G* and *IGFBP3* hold potential as biomarkers for clinical surveillance, offering new insights for early detection and prognosis evaluation.

Keywords: Non-alcoholic steatohepatitis (NASH); hepatocellular carcinoma (HCC); weighted gene co-expression network analysis (WGCNA); signature genes; immune cell infiltration

Submitted Nov 18, 2024. Accepted for publication Mar 04, 2025. Published online Apr 25, 2025.
doi: 10.21037/tcr-2024-2291
View this article at: https://dx.doi.org/10.21037/tcr-2024-2291

Introduction

The global cancer statistics reveals a consistent trend: primary liver cancer maintains its position as the sixth most frequently diagnosed cancer worldwide. Among these cases, hepatocellular carcinoma (HCC) accounts for nearly 90% (1,2). Nevertheless, the landscape of major risk factors for HCC is undergoing a transformation. Non-alcoholic fatty liver disease (NAFLD) and the following non-alcoholic steatohepatitis (NASH) has emerged as prominent drivers of the rising incidence of cirrhosis and HCC besides virus and alcohol, primarily due to the global surge in obesity, dyslipidemia and type 2 diabetes mellitus (3). NAFLD, which affects approximately 25% of the global population, is expected to see a 56% increase in NASH prevalence from 2016 to 2030 across major countries, including China, France, Germany, Italy, Japan, Spain, the United Kingdom, and the United States (4). Evidence from various regions highlights the rapid rise of NAFLD-associated HCC (NAFLD-HCC). For instance, NAFLD is now the fastest growing cause of HCC in United States liver transplant recipients and waiting list candidates (5). Moreover, a staggering 122% increase in HCC cases attributed to NAFLD in the United States by the year 2030 are being estimated (6). A study from South Korea indicated that NAFLD is strongly associated with the development of HCC, regardless of the presence of fibrosis, after controlling for demographic variables (7). Furthermore, in France, the prevalence of HCC among NAFLD patients increased from 2.6% in 1995 to 19.5% in 2014 (8).

NAFLD-HCC poses unique diagnostic challenges. The elevated level of alpha-fetoprotein (AFP) and des-gamma-carboxyprothrombin (DCP) may suggest an HCC. However, its role in the detection of small tumors and early NAFLD-HCC is unclear. Meanwhile, visual diagnosis of early HCC is limited in cases with complex liver pathological changes, such as NAFLD (9,10). Unfortunately, current clinical evidence does not support the use of abdominal ultrasonography monitoring for early HCC identification in individuals who do not have cirrhosis or severe fibrosis. This difficult issue leads to more NAFLD-HCC patients being diagnosed at later stages of the disease (11). In addition, because the onset of NAFLD-HCC occurs 4–6 years later than HCC caused by other factors (12), these patients may have an older age, more severe frailty, more comorbidities and a worse prognosis (11). It is noteworthy that a meta-study conducted by Younossi *et al.* [2016] (13) revealed a substantial increase in the risk of HCC development only after the transformation of NAFLD into the NASH subtype (annual incidence: 0.44 per 1,000 person-years to 5.29 per 1,000 person-years). This underscores the significance of exploring potential biomarkers in distinguishing between NASH and NASH-associated HCC (NASH-HCC) stages.

Immune cell infiltration is an important indicator for predicting the progress of oncology and responsiveness of tumor patients to immunotherapy, usually confirmed through biopsy tests with poor patient compliance (14,15). Due to the altered liver immune microenvironment caused by NASH, NASH-HCC immunotherapy is promising but challenging. Unfortunately, there is a paucity of studies that have delved into potential immune cell biomarkers for this specific distinction. Bioinformatics analysis can provide cost-effective and high-precision results, which enhances the understanding of the complexities of biology and provides

Highlight box

Key findings

- Six signature genes—C-C motif chemokine ligand 14 (*CCL14*), C-type lectin domain family 4 member G (*CLEC4G*), ficolin-2 (L-ficolin, *FCN2*), insulin-like growth factor binding protein 3 (*IGFBP3*), C-X-C motif chemokine ligand 14 (*CXCL14*) and vasoactive intestinal polypeptide type I receptor (*VIPR1*)—were identified with high diagnostic accuracy (area under the curve: 0.927–0.958). These genes are linked to non-alcoholic steatohepatitis (NASH) related pathways and immune cell infiltration, such as M2 macrophages.

What is known and what is new?

- NASH contributes to hepatocellular carcinoma (HCC), and immune cell infiltration affects tumor progression.
- *CLEC4G* and *IGFBP3* were downregulated in NASH-associated HCC (NASH-HCC), showing potential as biomarkers for early detection.

What is the implication, and what should change now?

- These findings offer a foundation for non-invasive monitoring of NASH-HCC. Validation in clinical settings is required.

valuable insights for future research. However, the sole available study lacked robust indicators for assessing and screening the significance of candidate genes. Furthermore, the results were not subjected to validation using datasets from different sources, which could potentially compromise their accuracy and reliability (16).

To provide precise outcomes, this study used various machine learning algorithms to bioinformatically analyze the transcriptome sequencing data of NASH and NASH-HCC patients to discover NASH-HCC signature genes, which were validated in external datasets and cell model. Finally, the immune cell infiltration of NASH-HCC was evaluated using the cell-type identification by estimating relative subsets of RNA transcripts (CIBERSORT) algorithm, which provides a new direction of thinking for the clinical diagnosis and treatment of NASH-HCC. We present this article in accordance with the TRIPOD reporting checklist (available at https://tcr.amegroups.com/article/view/10.21037/tcr-2024-2291/rc).

Methods

Data sources

In this study, the data from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/GEO/) were accessed. GSE164760 was utilized for the training set and GSE102079 for validation. The GSE164760 dataset includes transcriptome sequencing data from 74 NASH patients and 53 NASH-HCC patients and is derived from the GPL13667 platform of the Affymetrix Human Genome U219 Array (17). The dataset was provided by Pinyol *et al.*, who reported detailed demographic characteristics and relevant clinical indicators of the study subjects in their research. For the validation set, we utilized the GSE102079 dataset, sourced from the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570 platform). This dataset encompasses transcriptome sequencing data from cancerous tissue samples of 152 HCC patients, alongside 91 adjacent normal liver tissue samples (18,19). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Identification of differentially expressed genes (DEGs)

The dataset was processed using the "limma" package (20) within the R software. This involved background correction and normalization. Subsequently, the probe names in the probe annotation file were converted into gene symbols. DEGs were identified separately for the NASH and NASH-HCC groups. The specific criteria were applied requiring a threshold of $|\log_2$ fold change (FC)$|$ >1 and an adjusted P value <0.05 (21). These DEGs were visualized using volcano plots. Furthermore, the top 30 up-regulated and top 30 down-regulated DEGs were presented in a heat map.

Functional and pathway enrichment analyses

To assess the biological relevance the DEGs, the "clusterProfiler" package in R was employed to analyze Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (22). The Gene Ontology (GO) analysis was also conducted, which investigates the biological processes (BP) of the DEGs across three categories: BP, cellular component (CC) and molecular function (MF). In addition, the KEGG analysis was performed to explore potential signaling pathways associated with the DEGs.

WGCNA

In this study, WGCNA was performed using the R package "WGCNA" to identify the gene modules most closely associated with NASH-HCC (23). The WGCNA analysis was carried out based on scale-free topology criteria. Initially, all outliers were removed from the dataset. After the calculating of the soft-threshold power and transforming of topological overlap matrix, the corresponding dissimilarity was computed for hierarchical cluster analysis. To identify co-expressed gene modules, a dynamic tree-cutting method was applied. Each gene module contained a minimum of 30 genes. Then, the gene significance values and module membership values were calculated for assessing the correlation between gene modules and NASH-HCC. This process allowed us to identify critical gene modules. Furthermore, the candidate hub genes have been identified through the examination of the overlap between DEGs and key module genes.

Identification and assessment of signature genes

Three machine learning algorithms, the LASSO, random forest and support vector machine recursive feature elimination (SVM-RFE), were used to screen for significant features (24-26). The LASSO regression model determines variable importance by selecting optimal regularization parameters through cross-validation and

applying regularization to penalize regression coefficients, with all other parameters kept at their default settings. It is particularly effective in handling high-dimensional data compared to traditional regression methods. In the Random Forest analysis, two key parameters were specified: the number of trees in the forest was set to 500, and the number of variables used for node splitting in each tree was set to 11, while all other parameters remained at their defaults. Genes with importance scores greater than 0.5 were extracted from the results of the new Random Forest model, which was constructed by determining the number of trees corresponding to the model with the lowest error rate. A common supervised machine learning method for classification and regression called the SVM-RFE algorithm was also employed. For SVM-RFE, the control parameter for RFE was adjusted by setting the number of folds for cross-validation to 5, with all other settings left as default. The signature genes for NASH-HCC were identified by finding the intersection of genes selected by these three machine learning algorithms.

The receiver operating characteristic (ROC) curves and calibration curves were employed to assess the diagnostic effectiveness of the signature genes. Model calibration was based on the Brier score, which ranges from 0 to 0.25. The lower the Brier score, the higher the degree of calibration. Additionally, the internal validation using 1,000-fold bootstrap replication and external validation were performed on a separate dataset to confirm the diagnostic efficacy of the signature genes.

Gene set enrichment analysis (GSEA)

To investigate the correlation between these selected key genes and pathways, each gene was classified into two categories based on its median gene expression: high and low expression using GSEA (27) on these subgroups. The P values were used for correction with a significance threshold set at $P<0.05$.

Immune cell infiltration

CIBERSORT, a method that uses linear support vector regression (LSVR) was used to deconvolute expression matrices of human immune cell subtypes, to evaluate immune cell infiltration (28). CIBERSORT is widely recognized for its robust deconvolution analysis, particularly when dealing with unknown mixtures and matrices containing similar immune cell types (29). This makes it a reliable method for estimating immune cell infiltration. Initially, it was calculated that the relative abundance of 22 immune cell subtypes in each NASH and NASH-HCC sample. Subsequently, the student's *t*-test was employed to assess the potential differences in the 22 immune cell subtypes between the two groups. Pearson correlation analysis was conducted to determine whether immune cell subtypes are related to signature genes.

Real-time reverse transcription polymerase chain reaction (RT-qPCR) verified the expression of signature genes in HCC cell models

Hepatocellular carcinoma G2 (HepG2, Cobioer, Nanjing, China) cells were treated with 0.3 mmol/L sodium palmitate (Meryer, Shanghai, China) for 24 hours to establish a NAFLD-HCC cell model characterized by excessive fat accumulation (30). Then, after staining the cells using the Oil Red O staining kit (EallBio, Beijing, China), lipid droplets in untreated HepG2 cells and HepG2 cells treated with 0.3 mmol/L sodium palmitate for 24 hours were observed using a fluorescence microscope (KEYENCE, Shanghai, China). Nuclear factor erythroid-2-related factor 1 (Nrf1) was statically knocked out in HepG2 cells to construct NASH-HCC cell model (31). Total RNA of HepG2 and 2 model cells was extracted, and complementary DNA (cDNA) was obtained by reverse transcription using Hifair® III 1st Strand cDNA Synthesis Kit (Yeasen, Shanghai, China). The reaction condition of RT-qPCR experiment: 95 ℃ 3 min; 95 ℃ 10 s, 55 ℃ 30 s, 39 cycles; 65 ℃ 31 s; 65°C 15 s, 0.5 ℃/cycle gradient temperature rise, 60 cycles. Each experiment was set up with 3 replicate wells. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as an internal reference for correction, and the $2^{-\Delta\Delta CT}$ method was used to calculate the messenger RNA (mRNA) expression of the signature genes (32). The sequence of primers used is shown in Table S1.

Statistical analysis

The version R 4.3.0 was used for all statistical analyses. Group differences were assessed using the student's *t*-test and correlations were examined through Pearson correlation analysis. Statistical significance was defined as $P<0.05$, unless otherwise indicated. All P values were two-tailed. *Figure 1* displays the flowchart of the study.

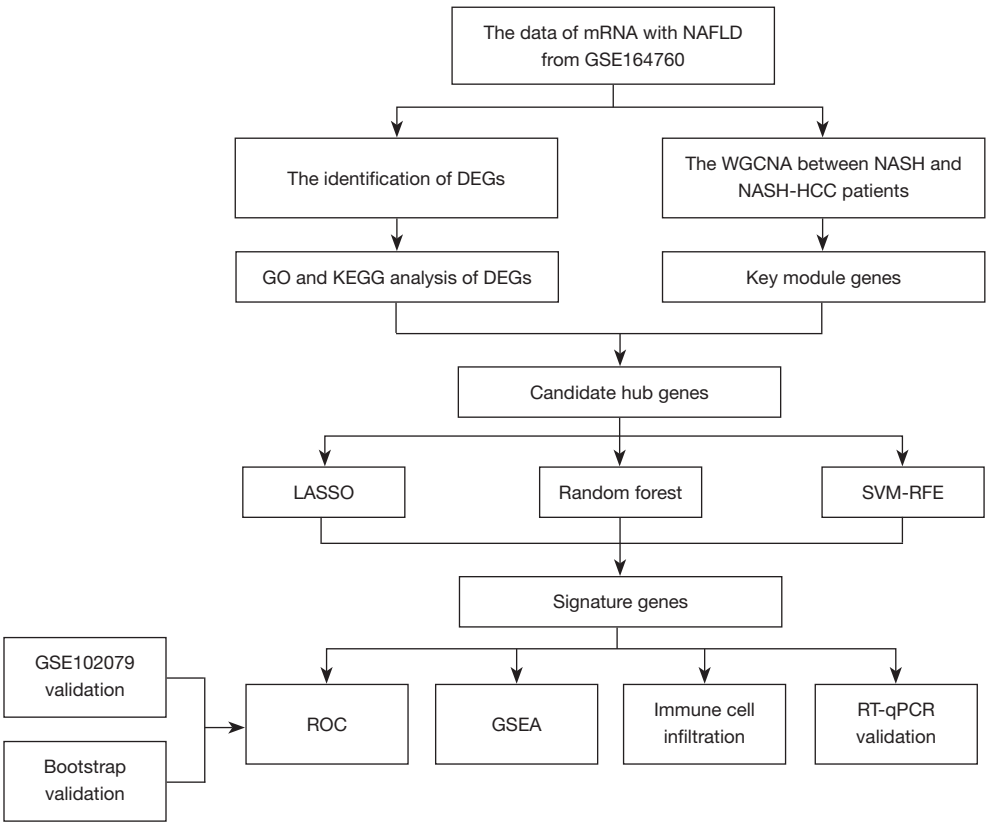

Figure 1 Flow chart of the analysis process. DEGs, differentially expressed genes; GO, Gene Ontology; GSEA, gene set enrichment analysis; HCC, hepatocellular carcinoma; KEGG, Kyoto Encyclopedia of Genes and Genomes; LASSO, least absolute shrinkage and selection operator; mRNA, messenger RNA; NAFLD, non-alcoholic fatty liver disease; NASH, non-alcoholic steatohepatitis; NASH-HCC, NASH-associated HCC; ROC, receiver operating characteristic; RT-qPCR, real-time reverse transcription polymerase chain reaction; SVM-RFE, support vector machine recursive feature elimination; WGCNA, weighted gene co-expression network analysis.

Results

Identification of DEGs

To study oncogenes associated with hepatocarcinogenesis in the context of NASH, the DEGs between NASH and NASH-HCC samples were analyzed using the "limma" package. In total, 164 DEGs were identified, comprising 118 downregulated genes and 46 upregulated genes (*Figure 2A*). *Figure 2B* illustrated a heat map displaying the top 30 upregulated and top 30 downregulated DEGs between NASH and NASH-HCC.

Function enrichment analysis

To study the function of the DEGs, the GO analysis encompassed three key categories: BP, CC and MF (*Figure 3A*). In the BP analysis, the top three enriched processes were steroid metabolic process, regulation of hormone levels and response to xenobiotic stimulus. For CC analysis, collagen-containing extracellular matrix and blood microparticle were significantly enriched. In MF analysis, receptor ligand activity, signaling receptor activator activity and carbohydrate binding were found to play crucial roles. Additionally, the KEGG analysis revealed the top 3 enriched pathways: drug metabolism-cytochrome P450, cytokine-cytokine receptor interaction and chemical carcinogenesis-DNA adducts (*Figure 3B*).

Weighted gene co-expression network analysis (WGCNA)

To select the gene modules most closely associated with NASH-HCC, the scale-free co-expression networks were

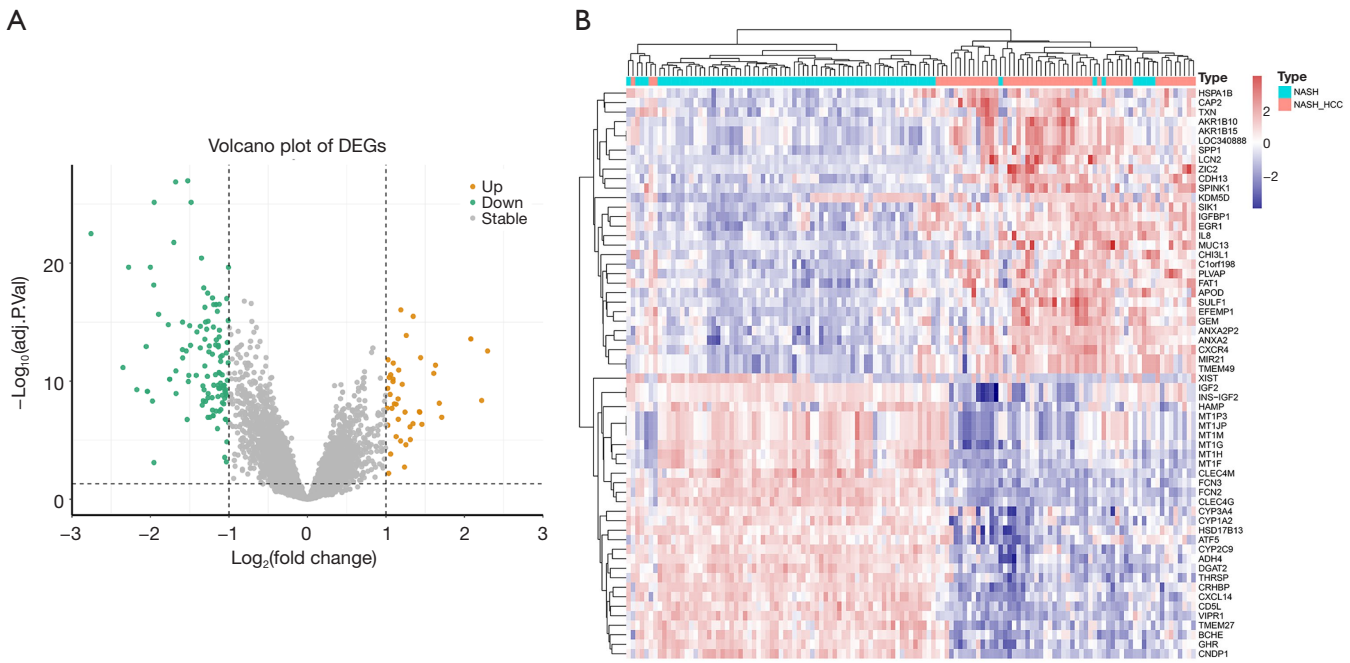

Figure 2 Screening for differential genes. (A) Volcano plots of 118 downregulated genes and 46 upregulated genes; (B) the top 30 upregulated and top 30 downregulated DEGs between NASH and NASH-HCC. DEGs, differentially expressed genes; HCC, hepatocellular carcinoma; NASH, non-alcoholic steatohepatitis; NASH-HCC, NASH-associated HCC.

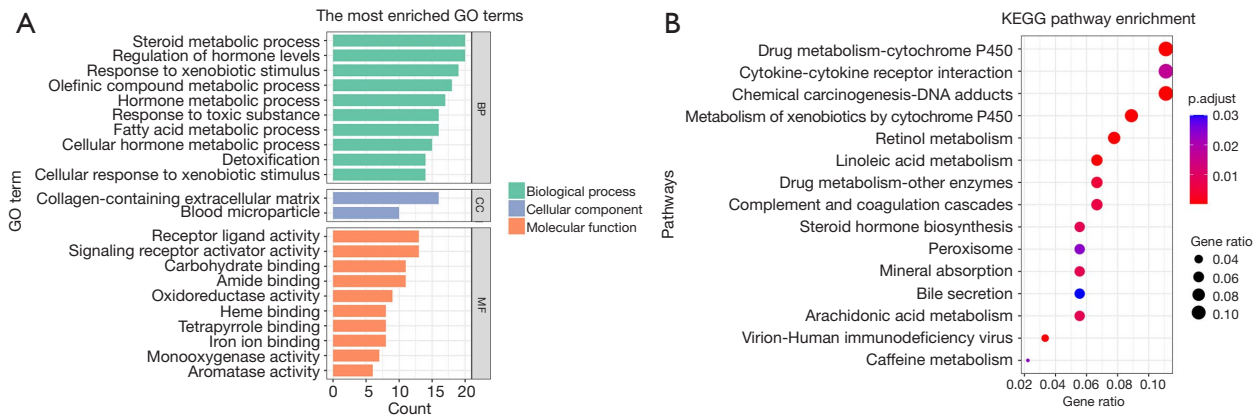

Figure 3 Functional enrichment analysis of DEGs. (A) Bar plot of the GO enrichment analysis; (B) bubble plot of the KEGG enrichment analysis. BP, biological process; CC, cellular component; DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; MF, molecular function.

established using the "WGCNA" package in R software. The selection of the soft threshold was based on achieving an initial value of the scale-free topology fitting index exceeding 0.85, which was found to be 7 (*Figure 4A*). The network diagram (*Figure 4B*) visually demonstrates that the network is well-connected when the soft threshold is set to 7. To represent gene clustering, a tree diagram was employed, including subgroups of genes (*Figure 4C*). The data were categorized into 13 modules, with the MEgreen module displaying a notable correlation with the disease

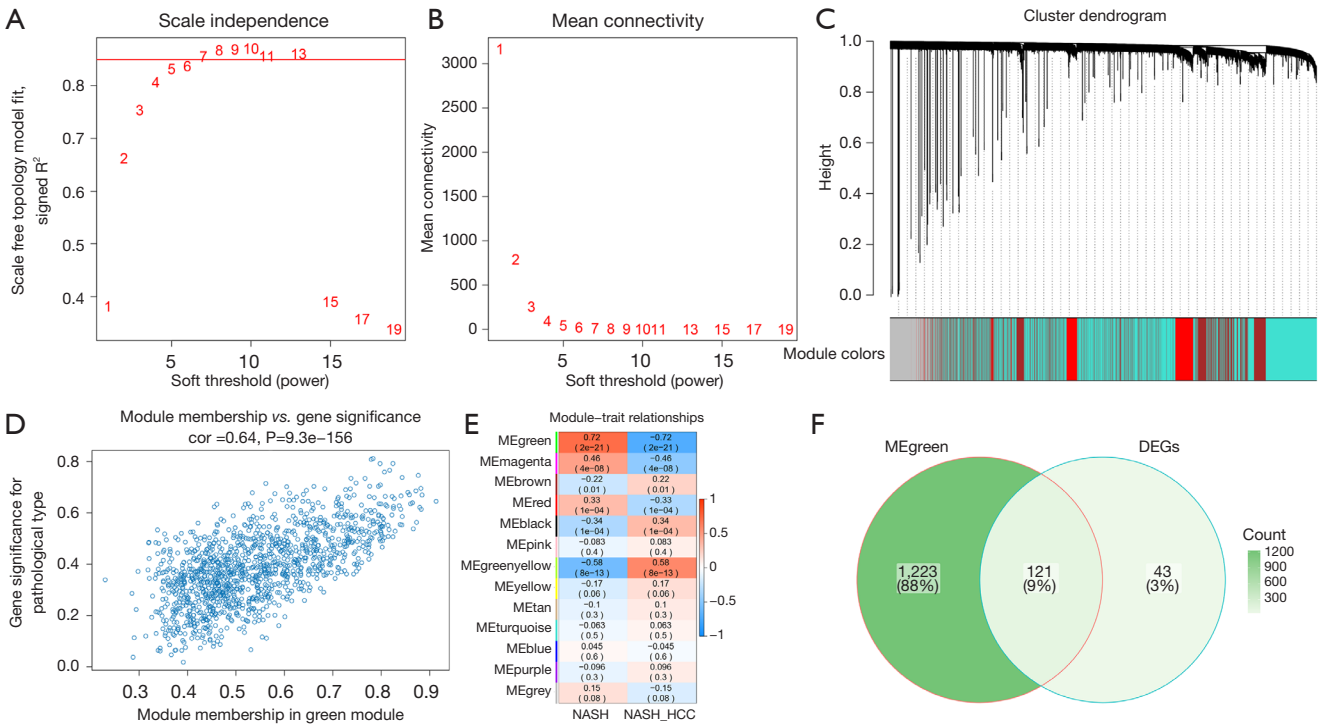

Figure 4 The WGCNA analysis of GSE164760 and identification of candidate hub genes. (A,B) Identification of soft threshold power and mean connectivity of WGCNA. (C) The cluster dendrogram of WGCNA. Different colors represent different modules, each consisting of genes with similar expression patterns. (D) The scatter plots of the correlation for module membership (X-axis) and gene significance (Y-axis) in HCC. (E) The correlation between clustered modules. The numbers in each cell of the graph indicate the correlation coefficient and P value, respectively. (F) The Venn plot of the genes selected by DEGs and MEgreen module genes. DEGs, differentially expressed genes; HCC, hepatocellular carcinoma; NASH, non-alcoholic steatohepatitis; WGCNA, weighted gene co-expression network analysis.

(*Figure 4D*). Correlation analysis further revealed that the MEgreen module exhibited the highest correlation with the disease (r=−0.72, P<0.001, *Figure 4E*). Through the intersection of DEGs and key module genes, we identified a total of 121 candidate hub genes (*Figure 4F*).

Screening of signature genes via machine learning algorithms

Three machine learning algorithms were used to screen signature genes from the hub genes. Thirteen feature genes were screened by LASSO regression analysis (Figure S1A,S1B), and there were 29 feature genes with relative importance greater than 0.5 in random forest analysis (Figure S1C), and support vector machine analysis screened 13 feature genes (Figure S1D). Table S2 listed the filtered feature genes. By taking the intersection of the results from these three algorithms, we identified six signature genes, including C-C motif chemokine ligand 14 (*CCL14*), C-type lectin domain family 4 member G (*CLEC4G*), ficolin-2 (L-ficolin, *FCN2*), insulin-like growth factor binding protein 3 (*IGFBP3*), C-X-C motif chemokine ligand 14 (*CXCL14*) and vasoactive intestinal polypeptide type I receptor (*VIPR1*) (Figure S1E).

Diagnostic performance of signature genes

The expression levels of six genes—*CCL14*, *CLEC4G*, *FCN2*, *IGFBP3*, *CXCL14* and *VIPR1*—were notably lower in NASH-HCC patients compared to NASH patients (Figure S2A-S2F). Furthermore, the area under the curve (AUC) values for these genes were as follows: 0.927, 0.952, 0.945, 0.933, 0.931 and 0.958, respectively (*Figure 5*, A1-A6). Calibration curves showed that the prediction and actual probabilities of the six genes are generally in agreement, with excellence noted for *CCL14*,

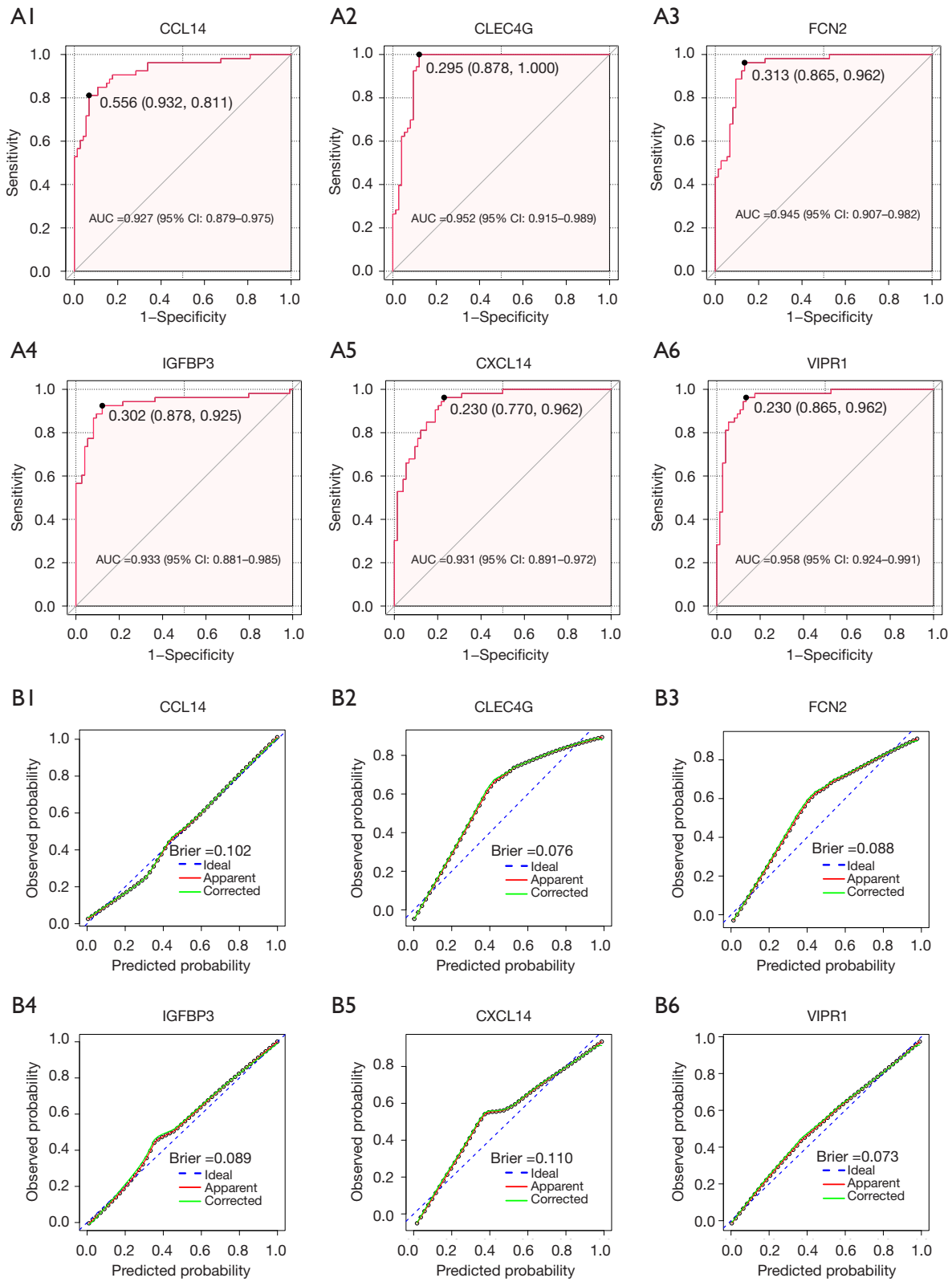

Figure 5 The diagnostic performance of the signature genes in GSE164760. (A1-A6) ROC curves. (B1-B6) Calibration curves. AUC, area under the curve; CI, confidence interval; ROC, receiver operating characteristic.

IGFBP3 and *VIPR1* (*Figure 5*, B1-B6). Upon internal validation using 1000 bootstrap resamples, it was reaffirmed that the average AUC values of the six genes ranged from 0.927 to 0.958, and the average Brier scores ranged from 0.073 to 0.110. Notably, the six identified genes exhibited exceptional diagnostic effectiveness and predictive probability in the external dataset (*Figure 6*, Figure S3).

GSEA

The GSEA was utilized to investigate the signaling pathways associated with the signature genes. Based on the findings presented in *Figure 7*, it was identified the top 10 signaling pathways for each of the six genes. The results revealed noteworthy patterns: "Butanoate metabolism" and "Fatty acid degradation pathway" exhibited positive associations with all six signature genes. In contrast, the "extracellular matrix (ECM)-receptor interaction pathway" displayed negative associations.

Assessment of immune cell infiltration

In comparison to NASH patients, NASH-HCC patients exhibited increased infiltration of specific immune cells, including naive B cells, M0 macrophages, plasma cells and activated memory $CD4^+$ T cells. Conversely, it demonstrated that the decreased infiltration of other immune cell types, including activated dendritic cells, M2 macrophages and T regulatory cells (*Figure 8A*). Furthermore, most of the signature genes displayed negative correlations with the infiltration of M0 macrophages, plasma cells and activated memory $CD4^+$ T cells, while showing positive correlations with the infiltration of M2 macrophages, activated dendritic cells and T regulatory cells (*Figure 8B*).

RT-qPCR validation of signature gene expression

Oil Red O staining results indicated that treatment of HepG2 cells with 0.3 mmol/L sodium palmitate for 24 hours successfully established an NAFLD-HCC cell model characterized by excessive fat accumulation (Figure S4). The expression of six characteristic genes in NAFLD-HCC, NASH HCC and HCC cell models was compared, respectively. The results showed that compared to HCC, the expression level of *FCN2* (P=0.01) was elevated in NAFLD-HCC cells, with no significant difference in other genes (*Figure 9A*). In the NASH-HCC cell model constructed by knockout Nrf1, the expression levels of *CLEC4G* (P=0.002), *IGFBP3* (P=0.005), and *FCN2* (P=0.02) genes were decreased, while the expressions of *CXCL14* (P=0.02) and *VIPR1* (P=0.001) were increased, and there was no significant difference in *CCL14* gene expression (*Figure 9B*).

Discussion

In this study, we screened six signature genes significantly associated with NASH-HCC through comprehensive bioinformatics analyses of two datasets from the GEO database. Upon internal and external validation, all of these genes demonstrated high diagnostic efficacy and predictive ability, supporting their reliability as potential biomarkers. In addition, RT-qPCR experiments further confirmed that the expression levels of the *CLEC4G* and *IGFBP3* genes were significantly downregulated in the NASH-HCC cell model, suggesting that they may play a key role in the progression of NASH to HCC. This finding provides a basis for further investigation into the functional mechanisms of these genes in disease progression, and also provides potential new ideas for future clinical diagnosis and targeted therapy.

CLEC4G encodes liver sinusoidal endothelial cell lectin (LSECtin), a member of the C-type lectin receptor family. As a key immunomodulatory receptor, *CLEC4G* tends to be expressed at high levels in normal liver tissues, but its expression decreases in malignant liver diseases and eventually becomes undetectable (33,34). Abundant literature has demonstrated the perturbations of glucolipid metabolism in NASH. Hepatic *IGFBP3* mRNA is proportionally associated with glycemia and insulin resistance in NAFLD patients (35). A previous study has shown that the level of circulate IGFBP3 decreases progressively from the control group to NAFLD patients and then to NASH patients (36). Moreover, *IGFBP3* expression is significantly downregulated and correlated with multiple clinicopathological characteristics in HCC (37). All these results imply that therapeutic interventions aimed at the *CLEC4G* and *IGFBP3* could hold significant promise in the treatment of NASH-HCC.

The GSEA demonstrated that *CLEC4G* and *IGFBP3* had positive associations with the butanoate metabolism and fatty acid degradation pathways. The fatty acid metabolism is essentially absent in the HCC stage (38). Besides, the butyrate metabolism is also identified as one of the most disturbed lipid-related signals in NASH and HCC (39-41). The imbalance of the butyrate metabolism pathway may relate to the uptake of fatty acids and lipid accumulation

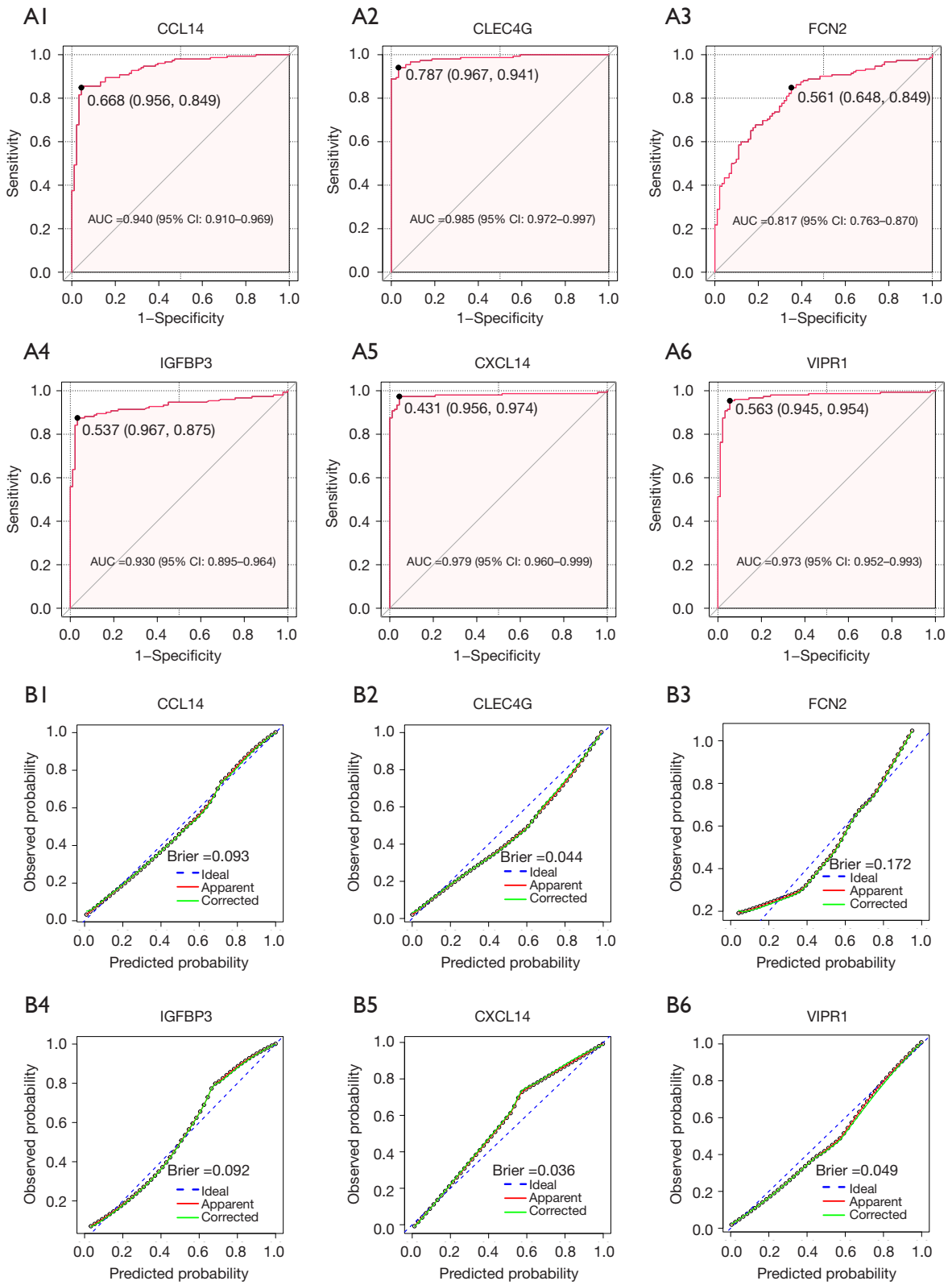

Figure 6 The diagnostic performance of the signature genes in GSE102079. (A1-A6) ROC curves. (B1-B6) Calibration curves. AUC, area under the curve; CI, confidence interval; ROC, receiver operating characteristic.

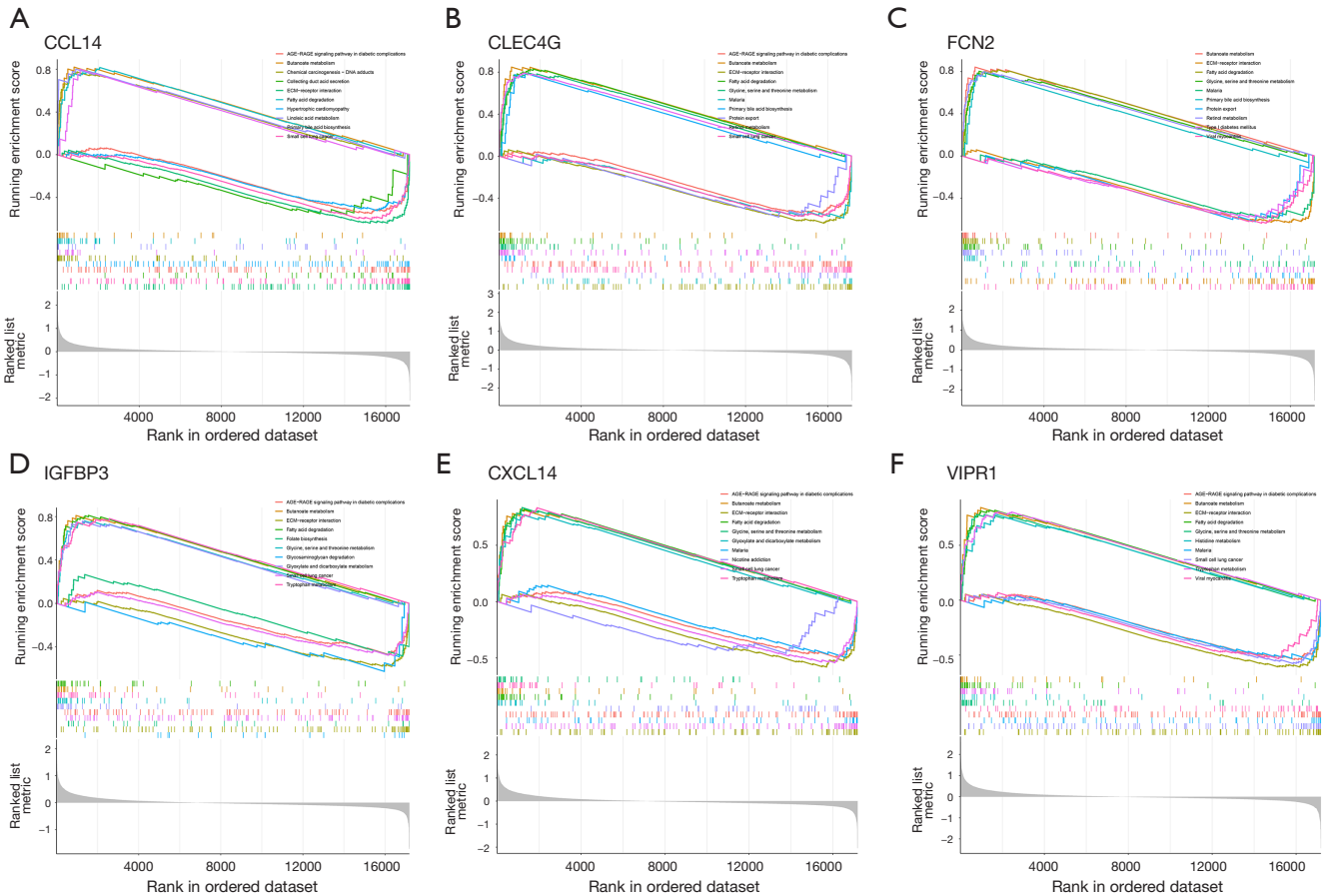

Figure 7 The signaling pathways associated with the signature genes. (A) *CCL14*; (B) *CLEC4G;* (C) *FCN2*; (D) *IGFBP3*; (E) *CXCL14*; (F) *VIPR1*.

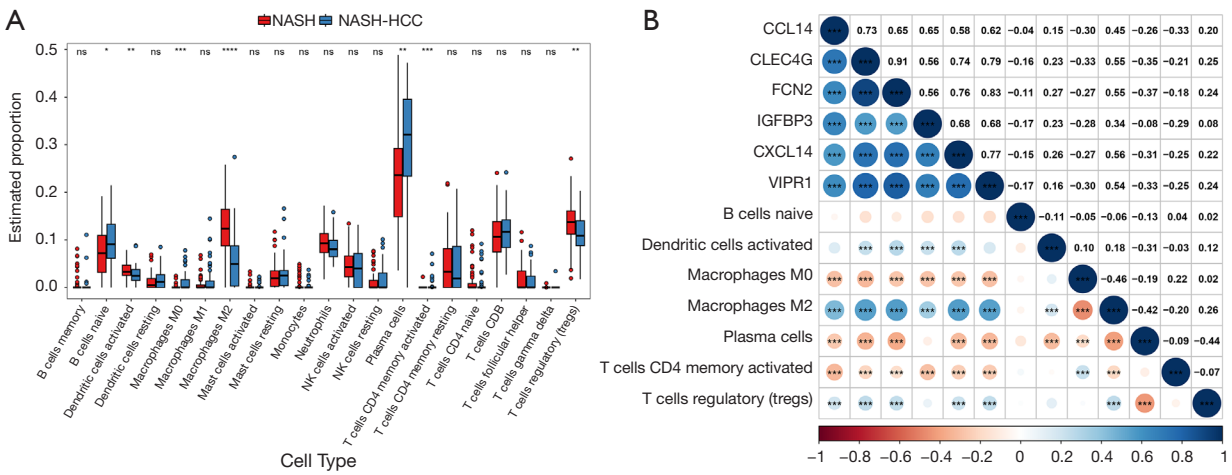

Figure 8 The association between immune cell infiltration and signature genes. (A) Comparison of immune cell infiltration between NASH and NASH-HCC group. (B) Correlation analysis between signature genes and significantly altered immune cell infiltration. The size of the dots and the shade of color indicate the value and significance of the correlation coefficient. ns, not significant; *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001. HCC, hepatocellular carcinoma; NASH, non-alcoholic steatohepatitis; NASH-HCC, NASH-associated HCC.

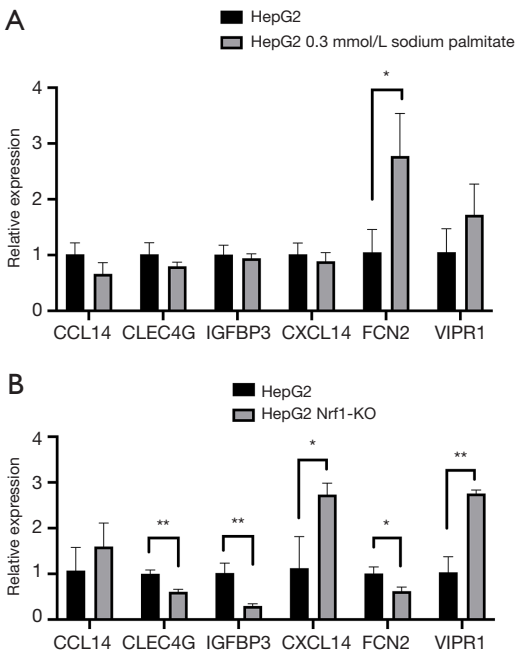

Figure 9 Expression of six signature genes in cell models. (A) NAFLD-HCC models; (B) NASH-HCC models. *, P<0.05; **, P<0.01. HCC, hepatocellular carcinoma; NAFLD, non-alcoholic fatty liver disease; NAFLD-HCC, NAFLD-associated HCC; NASH, non-alcoholic steatohepatitis; NASH-HCC, NASH-associated HCC.

in the liver, which accelerate the transition from NAFLD to NASH (42). In addition, *CLEC4G* and *IGFBP3* were inversely associated with the ECM-receptor interaction pathway, which enhanced HCC growth and development in both human and mouse studies (43,44). These findings give important information for future research into the metabolic abnormalities and etiology in NASH-HCC.

Immune infiltration status influences the choice of immunotherapy for HCC. Consistent with the findings of this study, Tang *et al.* [2009] (45) found that in a T cell-mediated acute liver injury mouse model, *CLEC4G* deficiency led to enhanced T cell immune response, thereby aggravating liver injury. Administration of exogenous recombinant CLEC4G protein or *CLEC4G* plasmid suppressed T cell activation and alleviated liver damage. These results further indicate that *CLEC4G* plays a key role in hepatic immune tolerance and may serve as a potential target for regulating T cell immune responses. Similarly, a study by Scully *et al.* (46) on *IGFBP3* and its role in obesity-induced breast cancer found that in *IGFBP3*-deficient mice, infiltration of $CD3^+$ T cells was significantly increased. This suggests that IGFBP3 may play a role in immune suppression or tumor immune evasion by inhibiting T cell tumor infiltration. In current study, it was observed that NASH-HCC patients exhibited elevated levels of immune cell infiltration, including naive B cells, M0-type macrophages, plasma cells and activated memory $CD4^+$ T cells, within their tumor tissues when compared to NASH patients. These findings are in concordance with previous research outcomes in HCC patients compared to healthy control (47). Notably, a recent study has presented varying results when comparing immune cell infiltration in healthy controls and NASH patients. These discrepancies involve cell types such as naive B cells, M2-type macrophages and activated dendritic cells (48). The diverge from these findings suggest a potential reversal in the stage of immune cell infiltration in NASH-HCC. Further functional research endeavors are required to assist the immunotherapy of NASH-HCC. Low expression of *CLEC4G* and *IGFBP3* combined with increased M2 macrophage infiltration and decreased plasma cell infiltration may have potential as biomarkers or even therapeutic targets in NASH-HCC oncology. Given that *CLEC4G* is associated with immune tolerance and *IGFBP3* with tumor suppression, their downregulation, along with shifts in immune cell composition, may contribute to immune evasion and tumor progression in NASH-HCC. These insights emphasize the importance of immune profiling in developing targeted immunotherapeutic approaches for NASH-HCC.

The biomarker screening studies for NAFLD and HCC can be mainly categorized into two types: the first category focuses on biomarkers for diagnosing NAFLD/NASH (49-52), while the second category targets biomarkers for diagnosing HCC related to NAFLD/NASH (53-55). Notably, the study by Meng *et al.* (49) found that *IGFBP2* is a potential biomarker for NASH, which is similar to the *IGFBP3* identified in this study. Both belong to the insulin-like growth factor binding protein (IGFBP) family and regulate the biological activity of insulin-like growth factor (IGF), thereby influencing cell growth, differentiation, and survival. This finding suggests that biomarkers may show consistent changes as one progresses from a healthy population to NAFLD/NASH patients, and further to NASH-HCC. Although the detection range expands, the specificity decreases, as NASH and NASH-HCC face significantly different clinical treatments. In addition, in the research on HCC-related biomarkers, Cai *et al.* (53) and Wang *et al.* (55) constructed protein-protein interaction

networks to screen for key genes. Although their findings differ from those of this study, they each identified cyclin-dependent kinase inhibitor 2A (*CDKN2A*) and cyclin-dependent kinase 1 (*CDK1*), which are closely related to cell cycle regulation and share some similarities. Overall, while these studies employed similar analytical methods, the differences in their findings may be attributed to variations in the study population, sample size, disease progression, and gene measurement methods.

While this study employed a comprehensive analytical approach that gradually focused on specific genes and identified significant signature genes, it was crucial to acknowledge its limitations. Firstly, future studies should expand the sample sizes to reduce individual random errors. Additionally, due to the small sample size, this study did not include an independent validation set during the feature gene selection process using machine learning methods, which may have affected the robustness of the findings. Secondly, the datasets used in the current study did not include data on demographic characteristics of the study population, routine clinical test metrics and prognosis, a limitation that hampered our ability to extend the association between these genes and NASH-HCC survival, thus limiting their potential clinical applicability. Lastly, the HepG2 cell culture model cannot fully replicate real-life NAFLD/NASH conditions. Liver fibrosis, which is marked by an excessive deposition of extracellular matrix components, is primarily a consequence of long-term and continuous liver injury (56). Thus, it was difficult to observe fibrosis formation in the cell culture model. To validate the roles of *CLEC4G* and *IGFBP3* in NAFLD/NASH-HCC, further *in vivo* validation via animal experiments is essential for the selected genes. Moreover, large-scale clinical research should be conducted to avoid racial bias in subsequent investigations. This comprehensive approach will help to elucidate the underlying mechanisms.

Conclusions

In summary, the two identified genes in this study—namely *CLEC4G* and *IGFBP3*—demonstrate remarkable accuracy and reliability in distinguishing between NASH and NASH-HCC. They hold promise as potential biomarkers for NASH-HCC patients. Moreover, the functions associated with these genes provide valuable insights for interpreting the pathogenesis of NASH-HCC. The findings necessitate further validation through more extensive clinical population studies and *in vitro* cellular experiments to confirm the potential utility of the identified signature genes as reliable biomarkers and targets for therapy.

Acknowledgments

We extend our profound and sincere gratitude to Professor Si-Wang Yu for the gracious donation of Nrf1 knock out HepG2 cell.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://tcr.amegroups.com/article/view/10.21037/tcr-2024-2291/rc

Peer Review File: Available at https://tcr.amegroups.com/article/view/10.21037/tcr-2024-2291/prf

Funding: This study was supported by R&D Program of Beijing Municipal Education Commission (No. KM202210017010) and China Postdoctoral Science Foundation (No. 2018M641116).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tcr.amegroups.com/article/view/10.21037/tcr-2024-2291/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Global Burden of Disease Liver Cancer Collaboration;

Akinyemiju T, Abera S, et al. The Burden of Primary Liver Cancer and Underlying Etiologies From 1990 to 2015 at the Global, Regional, and National Level: Results From the Global Burden of Disease Study 2015. JAMA Oncol 2017;3:1683-91.
2. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
3. Zeigerer A. NAFLD - A rising metabolic disease. Mol Metab 2021;50:101274.
4. Huang DQ, El-Serag HB, Loomba R. Global epidemiology of NAFLD-related HCC: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol 2021;18:223-38.
5. Younossi Z, Stepanova M, Ong JP, et al. Nonalcoholic Steatohepatitis Is the Fastest Growing Cause of Hepatocellular Carcinoma in Liver Transplant Candidates. Clin Gastroenterol Hepatol 2019;17:748-755.e3.
6. Estes C, Anstee QM, Arias-Loste MT, et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016-2030. J Hepatol 2018;69:896-904.
7. Kim GA, Lee HC, Choe J, et al. Association between non-alcoholic fatty liver disease and cancer incidence rate. J Hepatol 2017;S0168-8278(17)32294-8.
8. Pais R, Fartoux L, Goumard C, et al. Temporal trends, clinical patterns and outcomes of NAFLD-related HCC in patients undergoing liver resection over a 20-year period. Aliment Pharmacol Ther 2017;46:856-63.
9. Desai A, Sandhu S, Lai JP, et al. Hepatocellular carcinoma in non-cirrhotic liver: A comprehensive review. World J Hepatol 2019;11:1-18.
10. Quek J, Tan DJH, Chan KE, et al. Quality Assessment of Ultrasound and Magnetic Resonance Imaging for Hepatocellular Carcinoma Surveillance: A Systematic Review and Meta-Analysis. Dig Dis 2023;41:757-66.
11. Geh D, Anstee QM, Reeves HL. NAFLD-Associated HCC: Progress and Opportunities. J Hepatocell Carcinoma 2021;8:223-39.
12. Duan XY, Zhang L, Fan JG, et al. NAFLD leads to liver cancer: do we have sufficient evidence? Cancer Lett 2014;345:230-4.
13. Younossi ZM, Koenig AB, Abdelatif D, et al. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology 2016;64:73-84.
14. Ruf B, Heinrich B, Greten TF. Immunobiology and immunotherapy of HCC: spotlight on innate and innate-like immune cells. Cell Mol Immunol 2021;18:112-27.
15. Chew V, Chen J, Lee D, et al. Chemokine-driven lymphocyte infiltration: an early intratumoural event determining long-term survival in resectable hepatocellular carcinoma. Gut 2012;61:427-38.
16. Liu X, Wang Y, Li T, et al. Identification of Hub Genes and Immune Infiltration in Non-alcoholic Fatty Liver Disease -Related Hepatocellular Carcinoma by Bioinformatics Analysis. Turk J Gastroenterol 2023;34:383-93.
17. Pinyol R, Torrecilla S, Wang H, et al. Molecular characterisation of hepatocellular carcinoma in patients with non-alcoholic steatohepatitis. J Hepatol 2021;75:865-78.
18. Chiyonobu N, Shimada S, Akiyama Y, et al. Fatty Acid Binding Protein 4 (FABP4) Overexpression in Intratumoral Hepatic Stellate Cells within Hepatocellular Carcinoma with Metabolic Risk Factors. Am J Pathol 2018;188:1213-24.
19. Hatano M, Akiyama Y, Shimada S, et al. Loss of KDM6B epigenetically confers resistance to lipotoxicity in nonalcoholic fatty liver disease-related HCC. Hepatol Commun 2023;7:e0277.
20. 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.
21. Fan J, Shi S, Qiu Y, et al. Analysis of signature genes and association with immune cells infiltration in pediatric septic shock. Front Immunol 2022;13:1056750.
22. Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284-7.
23. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559.
24. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997;16:385-95.
25. Izmirlian G. Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. Ann N Y Acad Sci 2004;1020:154-74.
26. Huang S, Cai N, Pacheco PP, et al. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics 2018;15:41-51.
27. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl

Acad Sci U S A 2005;102:15545-50.
28. Chen B, Khodadoust MS, Liu CL, et al. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods Mol Biol 2018;1711:243-59.
29. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453-7.
30. Zhu H, Zhao T, Zhao S, et al. O-GlcNAcylation promotes the progression of nonalcoholic fatty liver disease by upregulating the expression and function of CD36. Metabolism 2024;156:155914.
31. Xu Z, Chen L, Leung L, et al. Liver-specific inactivation of the Nrf1 gene in adult mouse leads to nonalcoholic steatohepatitis and hepatic neoplasia. Proc Natl Acad Sci U S A 2005;102:4120-5.
32. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 2001;25:402-8.
33. Chen X, Li S, Jiang ZM, et al. C-type Lectin Domain Family 4 Member G (CLEC4G) Is a Negative Marker for CD34 in the Evolution of Liver Pathogenesis. Ann Clin Lab Sci 2023;53:516-28.
34. Zhang Y, Wei H, Fan L, et al. CLEC4s as Potential Therapeutic Targets in Hepatocellular Carcinoma Microenvironment. Front Cell Dev Biol 2021;9:681372.
35. Stanley TL, Fourman LT, Zheng I, et al. Relationship of IGF-1 and IGF-Binding Proteins to Disease Severity and Glycemia in Nonalcoholic Fatty Liver Disease. J Clin Endocrinol Metab 2021;106:e520-33.
36. Min HK, Maruyama H, Jang BK, et al. Suppression of IGF binding protein-3 by palmitate promotes hepatic inflammatory responses. FASEB J 2016;30:4071-82.
37. Chen J, Zhuang W, Xia Y, et al. Construction and validation of a novel IGFBP3-related signature to predict prognosis and therapeutic decision making for Hepatocellular Carcinoma. PeerJ 2023;11:e15554.
38. Balakrishnan K. Hepatocellular carcinoma stage: an almost loss of fatty acid metabolism and gain of glucose metabolic pathways dysregulation. Med Oncol 2022;39:247.
39. Pirola CJ, Sookoian S. The lipidome in nonalcoholic fatty liver disease: actionable targets. J Lipid Res 2021;62:100073.
40. Chuanbing Z, Zhengle Z, Ruili D, et al. Genes Modulating Butyrate Metabolism for Assessing Clinical Prognosis and Responses to Systematic Therapies in Hepatocellular Carcinoma. Biomolecules 2022;13:52.
41. Liu F, Li H, Chang H, et al. Identification of hepatocellular carcinoma-associated hub genes and pathways by integrated microarray analysis. Tumori 2015;101:206-14.
42. Zhou Y, Orešič M, Leivonen M, et al. Noninvasive Detection of Nonalcoholic Steatohepatitis Using Clinical Markers and Circulating Levels of Lipids and Metabolites. Clin Gastroenterol Hepatol 2016;14:1463-1472.e6.
43. Chandrashekar DS, Golonka RM, Yeoh BS, et al. Fermentable fiber-induced hepatocellular carcinoma in mice recapitulates gene signatures found in human liver cancer. PLoS One 2020;15:e0234726.
44. Zhang QJ, Li DZ, Lin BY, et al. SNHG16 promotes hepatocellular carcinoma development via activating ECM receptor interaction pathway. Hepatobiliary Pancreat Dis Int 2022;21:41-9.
45. Tang L, Yang J, Liu W, et al. Liver sinusoidal endothelial cell lectin, LSECtin, negatively regulates hepatic T-cell immune response. Gastroenterology 2009;137:1498-508.e1-5.
46. Scully T, Firth SM, Scott CD, et al. Insulin-like growth factor binding protein-3 links obesity and breast cancer progression. Oncotarget 2016;7:55491-505.
47. Chen W, Bi K, Zhang X, et al. In-depth characterization of the biomarkers based on tumor-infiltrated immune cells reveals implications for diagnosis and prognosis in hepatocellular carcinoma. J Transl Autoimmun 2020;3:100067.
48. Zhang Z, Wang S, Zhu Z, et al. Identification of potential feature genes in non-alcoholic fatty liver disease using bioinformatics analysis and machine learning strategies. Comput Biol Med 2023;157:106724.
49. Meng Q, Li X, Xiong X. Identification of Hub Genes Associated With Non-alcoholic Steatohepatitis Using Integrated Bioinformatics Analysis. Front Genet 2022;13:872518.
50. Zhang JJ, Shen Y, Chen XY, et al. Integrative network-based analysis on multiple Gene Expression Omnibus datasets identifies novel immune molecular markers implicated in non-alcoholic steatohepatitis. Front Endocrinol (Lausanne) 2023;14:1115890.
51. Huang S, Sun C, Hou Y, et al. A comprehensive bioinformatics analysis on multiple Gene Expression Omnibus datasets of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. Sci Rep 2018;8:7630.
52. Li LY, Wu JX. Analysis of hub genes and molecular mechanisms in non-alcoholic steatohepatitis based on the gene expression omnibus database. Zhonghua Yi Xue Za Zhi 2021;101:3317-22.
53. Cai C, Song X, Yu C. Identification of genes in

hepatocellular carcinoma induced by non-alcoholic fatty liver disease. Cancer Biomark 2020;29:69-78.
54. Dreval K, Tryndyak V, de Conti A, et al. Gene Expression and DNA Methylation Alterations During Non-alcoholic Steatohepatitis-Associated Liver Carcinogenesis. Front Genet 2019;10:486.
55. Wang B, Zhang Y, Gai L, et al. Identification of Key Biomarkers in Hepatocellular Carcinoma Induced by Non-alcoholic steatohepatitis or Metabolic Syndrome via Integrated Bioinformatics Analysis. Cell Mol Biol (Noisy-le-grand) 2023;69:174-80.
56. Parola M, Pinzani M. Liver fibrosis in NAFLD/NASH: from pathophysiology towards diagnostic and therapeutic strategies. Mol Aspects Med 2024;95:101231.

Cite this article as: Wang X, Gao M, Zhang Z, Ao X, Luo A, Wen Z, Pan X, Sun M, Wang T, Jia Z. Potential diagnostic marker gene set for non-alcoholic steatohepatitis associated hepatocellular carcinoma with lymphocyte infiltration. Transl Cancer Res 2025;14(4):2274-2289. doi: 10.21037/tcr-2024-2291