Original Article

Construction and validation of a joint diagnosis model based on random forest and artificial intelligence network for hepatitis B-related hepatocellular carcinoma

Xili Jiang[1], Jiyun Hu[2], Shucai Xie[2]

[1]Department of Radiology, The Second People's Hospital of Hunan Province/Brain Hospital of Hunan Province, Changsha, China; [2]Department of Critical Care Medicine, National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China

Contributions: (I) Conception and design: All authors; (II) Administrative support: None; (III) Provision of study materials or patients: X Jiang, J Hu; (IV) Collection and assembly of data: All authors; (V) Data analysis and interpretation: S Xie, X Jiang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Shucai Xie, MD. Department of Critical Care Medicine, National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, No. 87 Xiangya Road, Kaifu District, Changsha 410008, China. Email: 282791444@qq.com.

Background: Hepatitis B virus (HBV) is the dominant pathogenic factor of hepatocellular carcinoma (HCC) in Asia and Africa. Early identification and clinical diagnosis are crucial for HBV-related HCC. Random forest (RF) and artificial neural network (ANN) were an innovative and highly effective supervised machine learning (ML) algorithm for the early diagnosis and screening of HBV-related HCC. This study aims to identify significant biomarkers and develop a novel genetic model for the efficient diagnosis of HBV-related HCC.

Methods: Gene Expression Omnibus (GEO) Series (GSE)19665, GSE55092, and GSE121248 were used to identify significant differentially expressed genes (DEGs). The enrichment analysis was performed on Metascape online tool. The RF algorithm and ANN were used to select the potential predictive gene panels and construct an HBV-related HCC diagnostic model. Subsequently, GSE17548, GSE104310, GSE44074, and GSE136247 were used to test the accuracy of the ANN model. Finally, the CIBERSORT algorithm was used to assess the abundance of immune infiltrates in all samples.

Results: First, 116 genes were identified as DEGs, and the DEGs were particularly enriched in cellular hormone metabolic process, monocarboxylic acid metabolic process, NABA extracellular matrix (ECM) AFFILIATED steroid metabolic process and metabolism of bile acid and bile salt. DNA topoisomerase II alpha (*TOP2A*), C-type lectin domain family 1 member B (*CLEC1B*), BUB1 mitotic checkpoint serine/threonine kinase B (*BUB1B*), ficolin 2 (*FCN2*), C-X-C motif chemokine ligand 14 (*CXCL14*), cyclase associated actin cytoskeleton regulatory protein 2 (*CAP2*), ficolin 3 (*FCN3*), kynurenine 3-monooxygenase (*KMO*) and cadherin related family member 2 (*CDHR2*) were available to develop an HBV-related HCC diagnostic model. After validation, the diagnostic model showed high sensitivity (88.5%, 90%, 88.5%, 76.5%) and specificity (100%, 81.8%, 89.5%, 72.2%), and the areas under the receiver operating characteristic (ROC) curves showed excellent efficiency (1, 0.927, 0.921, 0.833). Finally, the percentage of infiltrating immune cell types [B cells naïve, B cells memory, plasma cells, T cells CD8, T cells CD4 memory resting, T cells regulatory (Tregs), T cells gamma delta, natural killer (NK) cells resting, NK cells activated, Macrophages M0, Dendritic cells activated, Mast cells activated] for hepatitis B-related HCC were significantly different from that of non-cancerous liver tissue with HBV.

Conclusions: A novel early diagnostic model of HBV-related HCC was established, and the model showed better efficiency in distinguishing HBV-related HCC from other non-cancerous with HBV individuals.

Keywords: Hepatocellular carcinoma (HCC); hepatitis B virus (HBV); random forest (RF); artificial intelligence network; diagnostic model

Submitted Jul 11, 2023. Accepted for publication Nov 21, 2023. Published online Feb 26, 2024.
doi: 10.21037/tcr-23-1197
View this article at: https://dx.doi.org/10.21037/tcr-23-1197

Introduction

Hepatocellular carcinoma (HCC) is the most prevalent primary liver cancer (90%) and the fourth leading cause of cancer-related death worldwide (1). By 2025, it is estimated to threat the health of more than 0.8 million people annually, with Chinese patients accounting for more than half of the global HCC burden (2). Variations in the incidence rate for HCC globally are attributed to diversity in risk factors. The prevalence of hepatitis B and C virus infections, especially the hepatitis B virus (HBV), is responsible for the highest incidence of HCC in East Asia and sub-Saharan Africa, with HBV-induced HCCs accounting for ~60% of cases in Asia and Africa (1,3). Chronic HBV infection leads to persistent liver damage and impaired regeneration, a well-known driving force of liver fibrogenesis and carcinogenesis (4). Therefore, there is an urgent need to identify reliable diagnostic markers to distinguish HBV-related HCC from other non-cancerous individuals with HBV.

Conventionally, early clinical diagnosis of HCC is depended on clinical symptoms, serum alpha-fetoprotein (AFP) and imaging findings in patients with chronic hepatitis or cirrhosis. Despite significant improvement in the prevention, monitoring, early screening, diagnosis and therapy of HCC over the past decade, the prognosis for the vast majority of HCC patients is typically poor. In addition, most patients lack obvious clinical symptoms in the early stage of HCC, and tumor that located deep within the abdominal cavity makes early diagnosis even more challenging. This highlights the importance of cancer research in identifying effective biomarkers, which are an attractive alternative for surveillance and early diagnosis of HCC because of its objectivity and reproducibility. Previous studies have identified a few potential diagnostic markers for HBV-related HCC, including AFP, Des-Gamma Carboxy Prothrombin (DCP), Golgi Protein Complex 73 (GPC 73), Osteopontin (OPN), cell-free/circulating tumor DNA, tumor-associated microRNAs and extracellular vesicles (5-11).

Machine learning (ML), unlike traditional statistical methods, is not rule-based programming but rather learning from examples. ML is an emerging discipline based on the intersection of statistics and mathematical sciences. It builds a statistical model from learning from large massive datasets data to achieve accurate prediction and to guide future research efforts (12,13). The random forest (RF), an innovative and highly effective supervised ML algorithm, uses several different prediction features in the training samples to effectively classify unknown samples by constructing a series of decision trees (13). RF classifier is an integrated approach consisting of multiple decision trees that are independent of each other. Each decision tree processes samples and predicts output labels, and the final output of the model is determined by the class that receives the most votes from the individual trees (14). As RFs overcomes the common problem of over-fitting through the use of bootstrap aggregation, it appears to be more accurate in prediction than other algorithms (15). Another supervised ML algorithm is artificial neural network (ANN), which is based on the functioning of biological neural networks. A neural network is composed of a large number of nodes (or neurons) connected to each other. The connection between each two nodes represents a weighted value for the signal passing through the connection, which is called the weight (16). Usually, these neurons are grouped in layers and

Highlight box

Key findings

- A diagnostic model of hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC) was established, and the model showed better efficiency in distinguishing HBV-related HCC from other non-cancerous with HBV individuals.

What is known and what is new?

- In the previous studies, differentially expressed genes and the association pathways involved in HBV-induced HCCs were identified through integrated bioinformatics analysis using multiple datasets. Other types of diagnostic and predictive models for HBV-related HCC have also been established previously.
- Based on Gene Expression Omnibus (GEO) expression data, a diagnostic model of early HBV-related HCC was established.

What is the implication, and what should change now?

- The findings give a deeper and more comprehensive understanding of the occurrence and progression of HCC and its association with HBV and a valuable reference for the early screening and directions for improving the clinical efficacy of HBV-related HCC.

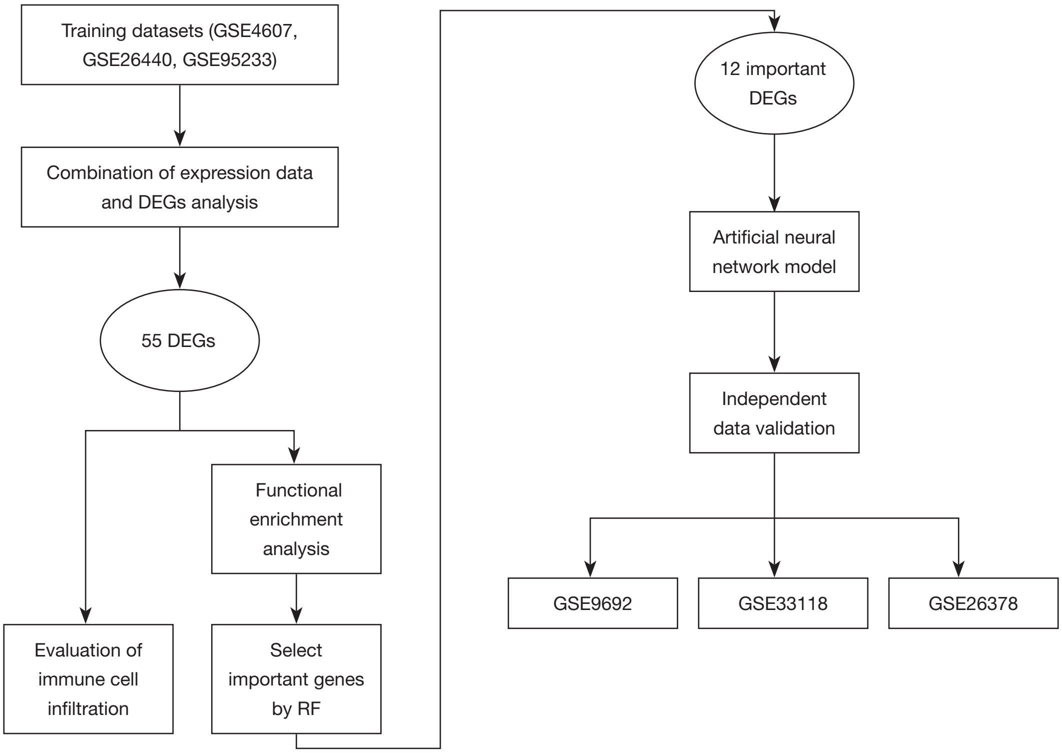
Training datasets (GSE4607, GSE26440, GSE95233)
Combination of expression data and DEGs analysis
55 DEGs
Functional enrichment analysis
Evaluation of immune cell infiltration
Select important genes by RF
12 important DEGs
Artificial neural network model
Independent data validation
GSE9692
GSE33118
GSE26378

Figure 1 Schematic illustration of the research design. GSE, Gene Expression Omnibus Series; DEGs, differentially expressed genes; RF, random forest.

process data in each layer, which are then passed forward to the next layers. Finally, the last layer responsible for making decisions and outputting results. The ANN is used to build a model of the complex relationship between input and output data and thus revealing the patterns (17). Compared to conventional programming, neural networks are available to deal with problems that algorithms could not solve, or the available solutions are too complex (17). ANN models are widely used in disease diagnosis, classification, prediction, and survival analysis because of their ability to handle linear and nonlinear relationship of data (18). It is well acknowledged that carcinogenesis and progression of HCC are closely related to mutation of genes, overexpression of various oncogenes and inactivation of tumor suppressor genes (19). With the rapid development in sequencing technology, huge volumes of gene expression profiling data related to cancer are generated for the identification of novel differential genes and diagnostic and prognostic biomarkers. In the previous studies, differentially expressed genes (DEGs) and the association pathways involved in HBV-induced HCCs were identified through integrated bioinformatics analysis using multiple datasets.

In this study, three datasets were merged. The RF algorithm was then used to identify the key genes expressed in HBV-related HCC, and ANNs constructed a genetic diagnostic model of HBV-related HCC. Finally, immune cell infiltration between HBV-related HCC samples and non-cancerous samples with HBV was evaluated. We present this article in accordance with the TRIPOD reporting checklist (available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-1197/rc).

Methods

Figure 1 shows the research framework of this study. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Gene expression data

Gene expression profiles of Gene Expression Omnibus (GEO) Series (GSE)19665 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19665) (20), GSE55092 (https://www.ncbi.nlm.nih.gov/geo/query/acc.

Table 1 Details of the GEO dataset

Dataset ID	Sample	Platform	Non-cancerous samples with HBV	HBV-related HCC	Classification	Country	Reference
GSE19665	HCC (HBV)	GPL570	5	5	Training sets	Japan	Deng *et al.* 2010, (20)
GSE55092	HCC (HBV)	GPL570	91	49	Training sets	USA	Melis *et al.* 2014, (21)
GSE121248	HCC (HBV)	GPL570	37	70	Training sets	Singapore	Wang *et al.* 2007, (22)
GSE17548	HCC (HBV)	GPL570	11	10	Test sets	Turkey	Yildiz *et al.* 2013, (23)
GSE104310	HCC (HBV)	GPL16791	7	9	Test sets	China	Yun *et al.* 2021, not published
GSE44074	HCC (HBV)	GPL13536	36	34	Test sets	Japan	Ueda *et al.* 2013, (24)
GSE136247	HCC (HBV)	GPL17586	19	26	Test sets	France	Cerapio *et al.* 2021, (25)

GEO, Gene Expression Omnibus; HBV, hepatitis B virus; HCC, hepatocellular carcinoma.

cgi?acc=GSE55092) (21), GSE121248 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121248) (22), GSE17548 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17548) (23), GSE104310 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104310), GSE44074 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44074) (24), and GSE136247 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136247) (25) were obtained from the GEO database of the National Centre for Biotechnology Information (https://www.ncbi.nlm.nih.gov/geo/) (raw data are available at https://cdn.amegroups.cn/static/public/tcr-23-1197-1.xlsx, *Table 1*). The seven datasets were divided into two groups: the training set, which including GSE19665, GSE55092, and GSE121248, and the remaining datasets were classified into the test set to verify the performance of the model. As previously described, the process of converting gene probe IDs to gene symbols was done using A Perl language command. The normalisation between arrays function was used to normalise the gene expression data, and the gene expression data were averaged when multiple probes correspond to a gene. Subsequently, the expression data of the three datasets in training sets were merged and used for the following analysis, the batch effect from the different datasets was removed, and the common genes were finally obtained. The gene expression data with a larger value were subjected to $\log_2$ transformation in the limma R package.

Identification of DEGs and enrichment analyses

The limma R package v.3.5.2 in R software was used to identify DEGs. The DEGs were selected based on the cut-off criterion that adjusted P value <0.05 and $|\log_2FC|$ >2. Metascape (http://metascape.org/gp/#/main/step1), a common integrated portal, contains functional enrichment, interactome analysis, gene annotation and membership search to provide a comprehensive gene list annotation and analysis resource for users to grasp biological characteristics (26). In present study, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were executed using Metascape (https://metascape.org/gp/index.html#/main/step1) online tool. P<0.05 was considered statistically significant.

RF screening for important genes

The RF software package (v.4.1.3) was used to filter out important variables and create an RF model that contributed most to the prediction of HBV-related HCC. First, the average model miscalculation rate of all genes based on out-of-band data was calculated. The best variable number for the binary tree at the node was set to 6, and 2000 was chosen as the best number of trees contained in the RF (27). Based on the point with the smallest error, the best RF model was then built, and the candidate genes for HBV-related HCC diagnosis were determined using the mean decrease Gini. Finally, for the subsequent model construction, the genes with a significance score greater than 4 were chosen as disease-specific genes.

Subsequently, scores were assigned to the expression data of the selected DEGs using the following rules: If an upregulated gene's log FC value for a sample was greater than the gene's median expression value across all samples, its score was automatically assigned as 1; otherwise, it was set to 0. If the log FC of the downregulated gene was greater than the mean expression value, its score was automatically assigned as 0; otherwise, it was set to 1. The heatmap of the selected DEGs was drawn to show their expression in the merged dataset.

Neural network to build the disease classification model

The R software package neural net (v.1.44.2) were available to develop an ANN model of the important variables. The weight of each gene was obtained and five hidden layers were set as the model parameters to build a classification model_of HBV-related HCC through the obtained gene score. The model accuracy results were obtained for HBV-related HCC samples and non-cancerous samples with HBV in the training set, and the receiver operating characteristic (ROC) software package was used to calculate the areas under the ROC curves (AUCs) classification performance verification results.

Validation of the predictive model

Four independent datasets (GSE17548, GSE104310, GSE44074, GSE136247) were used to verify the accuracy of the ANN model for classifying samples (HBV-related HCC or non-cancerous samples with HBV), and the ROC curves for each dataset were drawn using the pROC software package separately. At the same time, the optimal threshold in the ROC curve and the sensitivity and specificity in classifying cancer and normal samples under this threshold were calculated.

Evaluation of immune cell infiltration

The normalised gene expression data from the merged dataset was available to evaluate the abundance of immune infiltrates in all samples through the CIBERSORT algorithm. The percentages of 22 infiltrating immune cell types were calculated and output with the cutoff criterion that P value <0.05, and their correlations were displayed in a correlation heatmap drawn by the "corrplot" package (28). The ratios of infiltrating immune cells in non-cancerous liver tissues from HBV patients and HBV-related HCC tissues were visualised by a histogram, and the difference was shown by violin diagrams.

Statistical analysis

The limma R package v.3.5.2 in R software was used to identify DEGs. The DEGs were selected based on the cut-off criterion that adjusted P value <0.05 and $|\log_2FC| > 2$. The performance of ANN model was evaluated using ROCs, and the AUC, sensitivity, and specificity were determined. The correlation between 22 infiltrating immune cell types was assessed by calculating the Pearson correlation coefficient.

Results

Identification of DEGs in HCC

The samples in all datasets were strictly screened, and the samples without chronic HBV infection were excluded. GSE19665, GSE55092, and GSE121248 gene expression data were merged as a training dataset for subsequent analysis. A total of 133 non-cancerous liver tissues with HBV and 124 HBV-related HCC tissues were included in present analysis. As shown in the volcano graph (*Figure 2A*), 116 genes were identified as DEGs according to the cut-off criterion that adjusted P value <0.05 and $|\log_2FC| > 2$ (Table S1, Figure S1). *Figure 2B* shows a heatmap of the top 10 up- and downregulated genes.

Functional enrichment analysis of DEGs in the training dataset

To further investigate the biological functions of the 116 DEGs, GO analysis and KEGG pathway enrichment analysis were performed using online database Metascape. As previously described (29), the GO analysis consisted of three functional groups, namely, the biological process (BP) group, the cellular component (CC) group and the molecular function (MF) group. The results of GO analysis exhibited that the DEGs were particularly enriched in the BP (*Figure 3A*, Table S2), including monocarboxylic acid metabolic process, response to bacterium, response to peptide, regulation of growth, and cellular response to xenobiotic stimulus. For the CC (*Figure 3B*), the DEGs were mainly enriched in collagen-containing extracellular matrix, spindle, external side of plasma membrane, blood microparticle, and basolateral plasma membrane. In the MF (*Figure 3C*), the DEGs were principally enriched in oxidoreductase activity, protein homodimerization activity, carbohydrate binding, amide binding, and lipid transporter activity.

The results of KEGG pathway enrichment analysis revealed that the DEGs were particularly enriched in bile

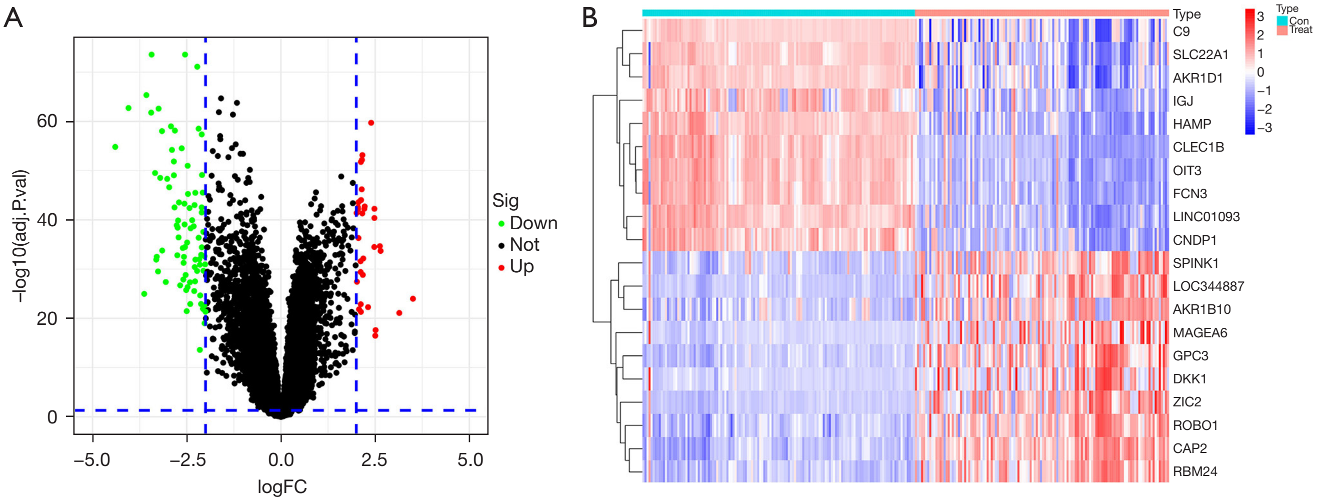

Figure 2 Identification of DEGs in HCC. (A) Volcano plot of differential expression analysis results. The abscissa is log$_2$FC and the ordinate is –log10 adjust P value. The red dots represent the upregulated genes based on an adjusted P<0.05 and log$_2$FC >2; the green dots represent the downregulated genes based on an adjusted P<0.05 and log$_2$FC <2; the black dots represent the remaining stable genes. (B) Heatmap of the top 10 up- and downregulated genes. Colours on the graph from red to blue indicate high to low expression. On the upper part of the heatmap, the blue band indicates the non-cancerous HBV samples and the red band indicates HBV-related HCC samples. FC, fold change; DEGs, differentially expressed genes; HCC, hepatocellular carcinoma; HBV, hepatitis B virus.

secretion, cytokine-cytokine receptor interaction, caffeine metabolism, tryptophan metabolism, and steroid hormone biosynthesis (*Figure 3D*, Table S3).

RF screening for DEGs

All 116 DEGs were included in the RF classifier. *Figure 4A* shows the relationship between the model error and the number of decision trees. The final model showed a stable error when the number of decision trees was 2000. Therefore, the RF model was built with 2000 trees as the parameter of the final model. Genes with an importance score are shown in *Figure 4B*; genes with an importance score greater than 4 were selected as the candidate genes for subsequent analysis. Finally, nine genes were selected, DNA topoisomerase II alpha (*TOP2A*), C-type lectin domain family 1 member B (*CLEC1B*), BUB1 mitotic checkpoint serine/threonine kinase B (*BUB1B*), ficolin 2 (*FCN2*), C-X-C motif chemokine ligand 14 (*CXCL14*) and cyclase associated actin cytoskeleton regulatory protein 2 (*CAP2*) being the most important, followed by ficolin 3 (*FCN3*), kynurenine 3-monooxygenase (*KMO*) and cadherin related family member 2 (*CDHR2*). As shown in *Figure 4C*, *CAP2*, *TOP2A*, *BUB1B* were upregulated in HBV-related HCC samples, while *KMO*, *CDHR2*, *CXCL14*, *FCN2*, *CLEC1B* were upregulated in non-cancerous liver tissue with HBV.

Construction of the ANN model

Expression data for these nine genes in each sample were assigned a score of 1 or 0. Based on these nine important variables, an ANN model was constructed and used to distinguish HBV-related HCC tissues and non-cancerous liver tissue with HBV in 257 samples of the merge datasets (*Figure 5A*). As a result, the model could correctly predict 132 cases in the HBV-related HCC group with 99.2% (132/133) accuracy and 120 cases in the non-cancerous with HBV group with 96.8% (120/124) accuracy. The AUC of the model in the training dataset were close to 1 (average AUC >0.99), showing the highly stable of the model in diagnosing HBV-related HCC (*Figure 5B*).

Validation of the ANN model

Four independent datasets (GSE17548, GSE104310, GSE44074, GSE136247) were used to verify the performance of the ANN model to classify samples (HBV-related HCC tissues or non-cancerous liver tissues with HBV). As a result, the model could correctly predict 23 cases in the HBV-related HCC group with 88.5% (23/26)

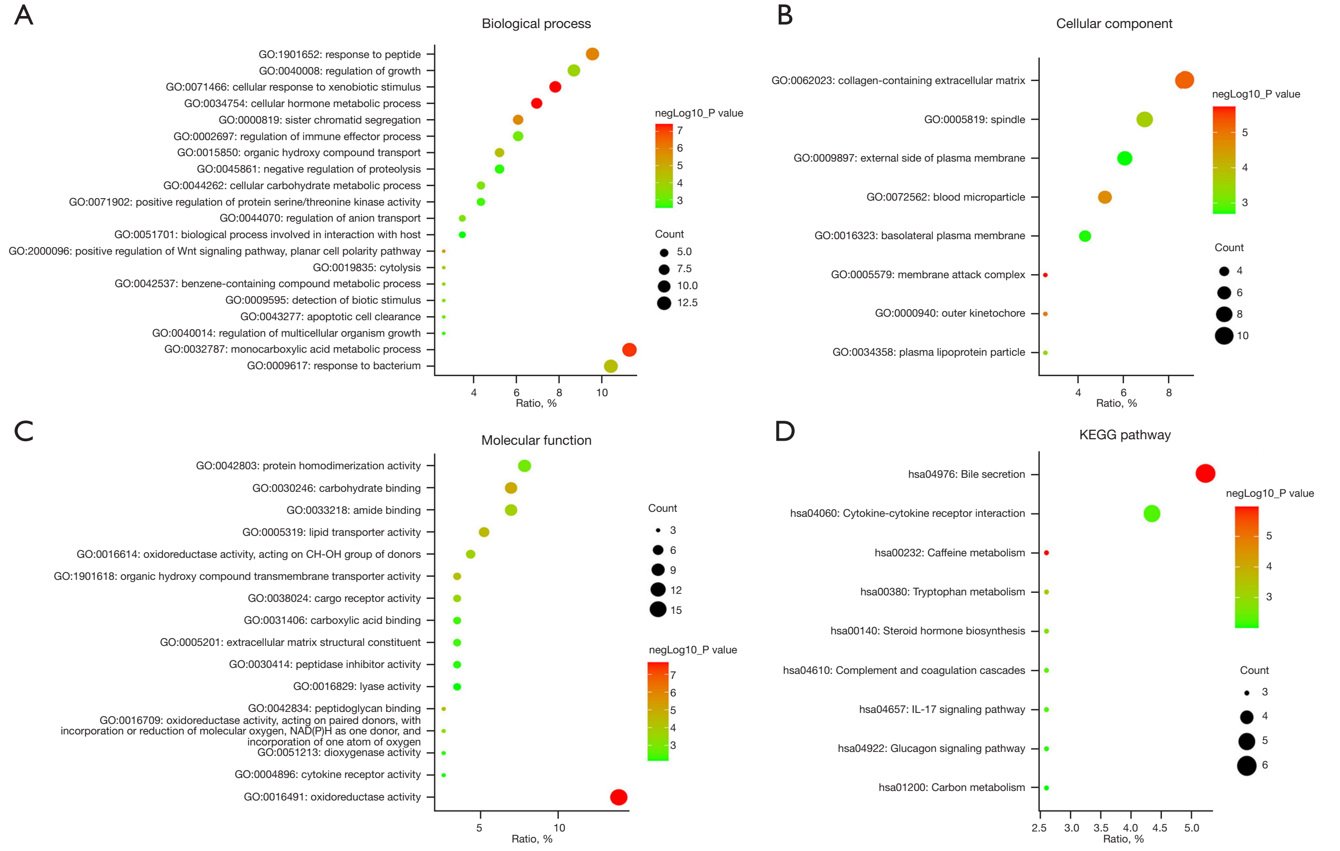

Figure 3 GO analysis and KEGG pathway enrichment analysis of DEGs using the online database Metascape. (A) Biological processes of GO analysis. (B) Cellular components of GO analysis. (C) Molecular functions of GO analysis. (D) KEGG pathway enrichment analysis. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes.

accuracy and 19 cases in the non-cancerous with HBV group with 100% (19/19) accuracy in GSE136247, 9 cases in the HBV-related HCC group with 90% (9/10) accuracy and 9 cases in the non-cancerous with HBV group with 81.8% (9/11) accuracy in GSE17548, 23 cases in the HBV-related HCC group with 88.5% (23/26) accuracy and 17 cases in the non-cancerous with HBV group with 89.5% (17/19) accuracy in GSE104310, and 26 cases in the HBV-related HCC group with 76.5% (26/34) accuracy and 26 cases in the non-cancerous with HBV group with 72.2% (26/36) accuracy in GSE44074. The AUCs of the model in the test dataset were 1 [95% confidence interval (CI): 1–1], 0.927 (95% CI: 0.791–1), 0.921 (95% CI: 0.738–1) and 0.833 (95% CI: 0.725–0.918), respectively (*Figure 6*).

Immune cell infiltration results

A total of 133 cases of non-cancerous liver tissues from HBV patients and 124 cases of HBV-related HCC tissues were selected for the immune cell infiltration analysis. Based on the cut-off criterion that $P<0.05$, 38 cases of HBV-related HCC tissues and 31 cases of non-cancerous liver tissues from HBV patients were selected for CIBERSORT analysis. First, the percentages of 22 kinds of immune cells in each sample were visualised in a histogram (*Figure 7A*). The correlations of 22 kinds of infiltrating immune cells between HBV-related HCC tissues and non-cancerous liver tissue with HBV were analysed (*Figure 7B*). For example, T follicular helper cells were positively correlated with T cells $CD8^+$ and macrophages M1. Natural killer (NK) cells resting were positively associated with neutrophils and T cells CD4 naïve. The Wilcoxon test was used to detect significantly different immune cell infiltrates between HBV-related HCC tissues and non-cancerous liver tissue with HBV. The results that presented 12 types (B cells naive, B cells memory, plasma cells, T cells CD8, T cells CD4

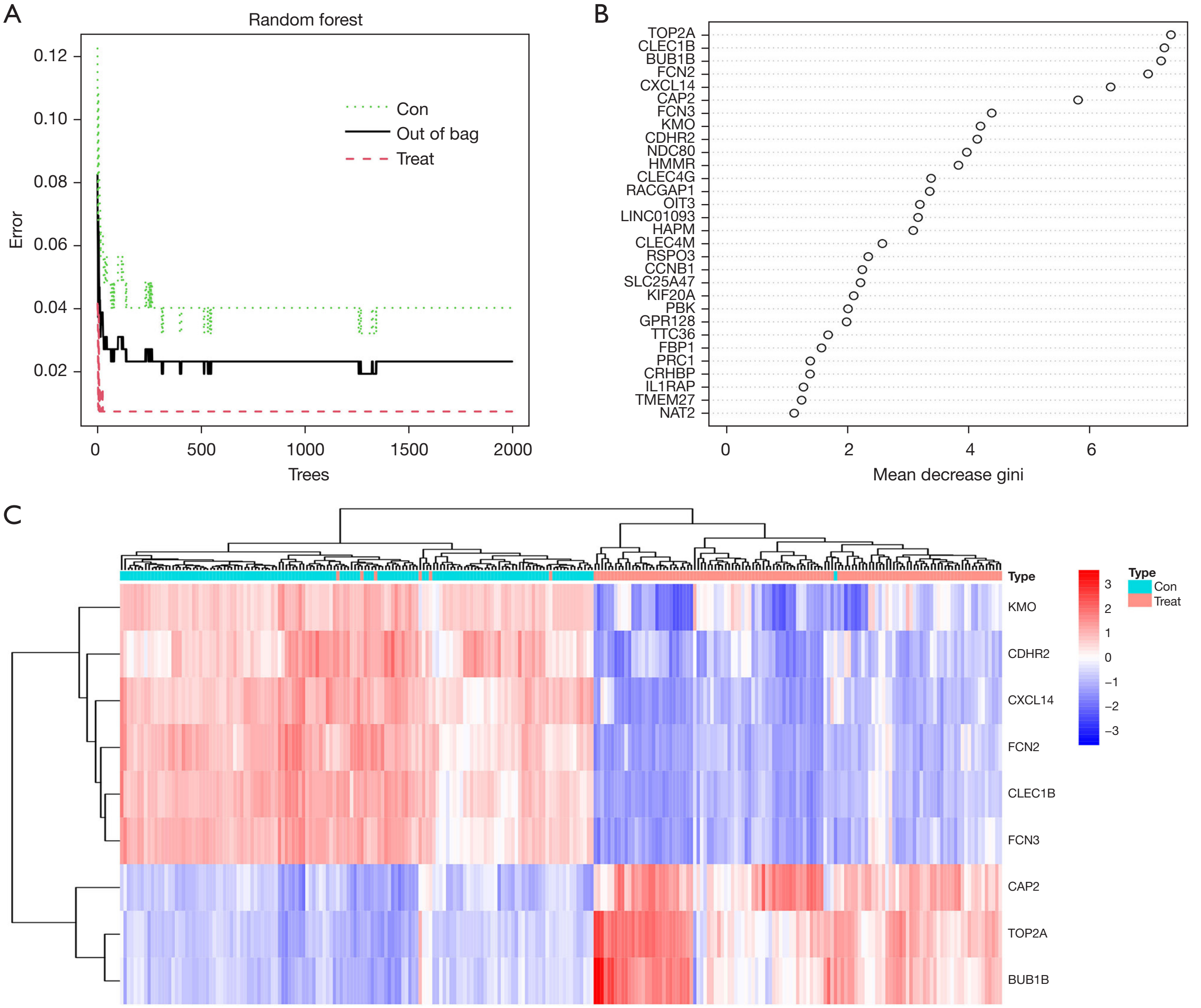

Figure 4 RF was used to screen differential genes, and nine genes were selected. (A) The influence of the number of decision trees on the error rate. The x-axis represents the number of decision trees, and the y-axis indicates the error rate. (B) The importance of the top 30 genes ranked by mean accuracy decreases. (C) Heatmap of the nine important genes generated by RF. The red colour indicates high expression genes in the samples, the blue colour indicates low expression genes in the samples, the red band on the upper side of the heatmap represents HBV-related HCC samples, and the blue band indicates non-cancerous liver tissue with HBV. RF, random forest; HBV, hepatitis B virus; HCC, hepatocellular carcinoma.

memory resting, Tregs, T cells gamma delta, NK cells resting, NK cells activated, Macrophages M0, Dendritic cells activated, Mast cells activated) of immune cells with P<0.05 are shown in a violin diagram in *Figure 7C*.

Discussion

This study aimed to establish an effective diagnostic model for HBV-related HCC based on gene expression data from GEO. The three datasets in the training group were from different countries, using the same sequencing platform, which minimised the effect of confounding factors to some extent, 116 DEGs were identified in the merged dataset formed from three HBV-related HCC datasets. Nine important candidate DEGs were acquired through the RF classifier, and a neural network model was created. Four

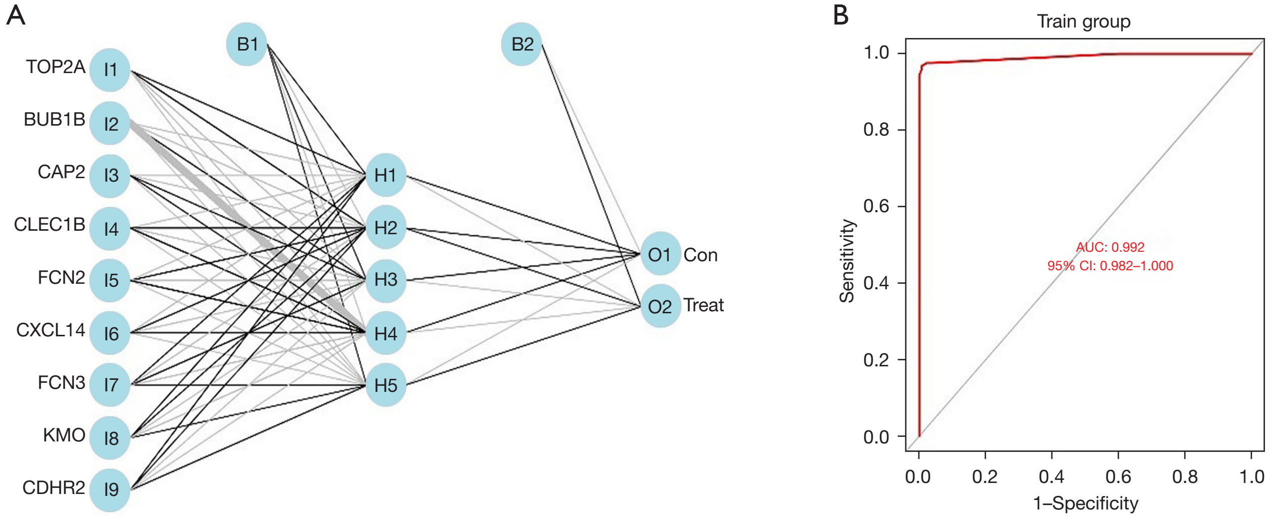

Figure 5 ANN model was constructed in the merged datasets. (A) Construction of a neural network: the neural network topology of the dataset with five hidden layers. (B) The ROC curve of the predictive model of the training dataset. AUC, area under the ROC curve; ROC, receiver operating characteristic; CI, confidence interval; ANN, artificial neural network.

independent datasets were used to verify the classification (HBV-related liver cancer or non-cancerous liver tissues with HBV) efficiency of the model, and the AUC (1, 0.927, 0.921, 0.833) efficiency showed excellent. Four independent datasets from different countries and regions were used to assess the performance of this diagnostic model, increasing the stability, usefulness and credibility of this model. The immune cell infiltration result shows that the percentages of 12 types of immune cells were significantly different between HBV-related HCC tissues and non-cancerous liver tissue with HBV.

RF and ANN are different types of algorithms. RF is an ensemble decision tree approach in which each decision tree processes a sample and predicts an output label. Decision trees in an ensemble are independent. ANN is composed of many layers of nodes that carry the signal and process it to make the final decision (30). An ANN model for the diagnosis and screening of HBV-related HCC was constructed based on nine important genes from RFs. Of these nine genes, TOP2A and BUB1B have been extensively studied in HCC (31-35). KMO (36,37), CDHR2 (38), CLEC1B (39), CXCL14 (40) and FCN2 (41) were significantly decreased in HCC tissues (or) and cell lines, overexpression of these genes exhibited tumor-inhibitory effects towards HCC (36,37), including inhibiting tumor formation and the growth of subcutaneous tumors, suppresseing proliferation, migration and invasion of HCC cells, epithelial-mesenchymal transition (EMT) and induced apoptosis. FCN3 expression was significantly lower in HCC tissues than in normal tissues (42). However, more *in vitro* and *in vivo* experiments are needed to further confirm its effect on HCC. KMO (37), CXCL14 (43), CAP2 (44) and FCN3 (45) were prognostic markers in HCC, and the combination of PD-L1$_{\mathrm{high}}$ and CLEC1B$_{\mathrm{low}}$ expression has been shown to predict worse outcomes (46).

CAP2 was a valuable molecular marker in the histological diagnosis of early HCC (47), and its overexpression might be related to multistage hepatocarcinogenesis (48). In addition, CAP2 transcriptional levels were significantly suppressed in silibinin-treated HCC cells. Silibinin could be a potential therapeutic agent against HCC, particularly for HBV-related HCCs (49). These findings indicate that CAP2 may play a critical role in the carcinogenesis or progression of HBV-related HCC. CXCL14 was markedly suppressed in HBV-related HCC tissues, and its polymorphisms were associated with advanced-stage chronic HBV infection (50). FCN2 is active in hepatitis B infection (51), and ficolin-2 serum levels and FCN2 haplotypes contribute to the outcome of HBV infection in a Vietnamese cohort (51). FCN2 was implied, which was implied to play a crucial role in innate immunity against HBV infection.

Other types of diagnostic and predictive models for HBV-related HCC have also been established previously. ATP binding cassette subfamily B member 6 (*ABCB6*), importin 7 (*IPO7*), translocase of inner mitochondrial membrane 9 (*TIMM9*), frizzled class receptor 7 (*FZD7*), and acetyl-CoA

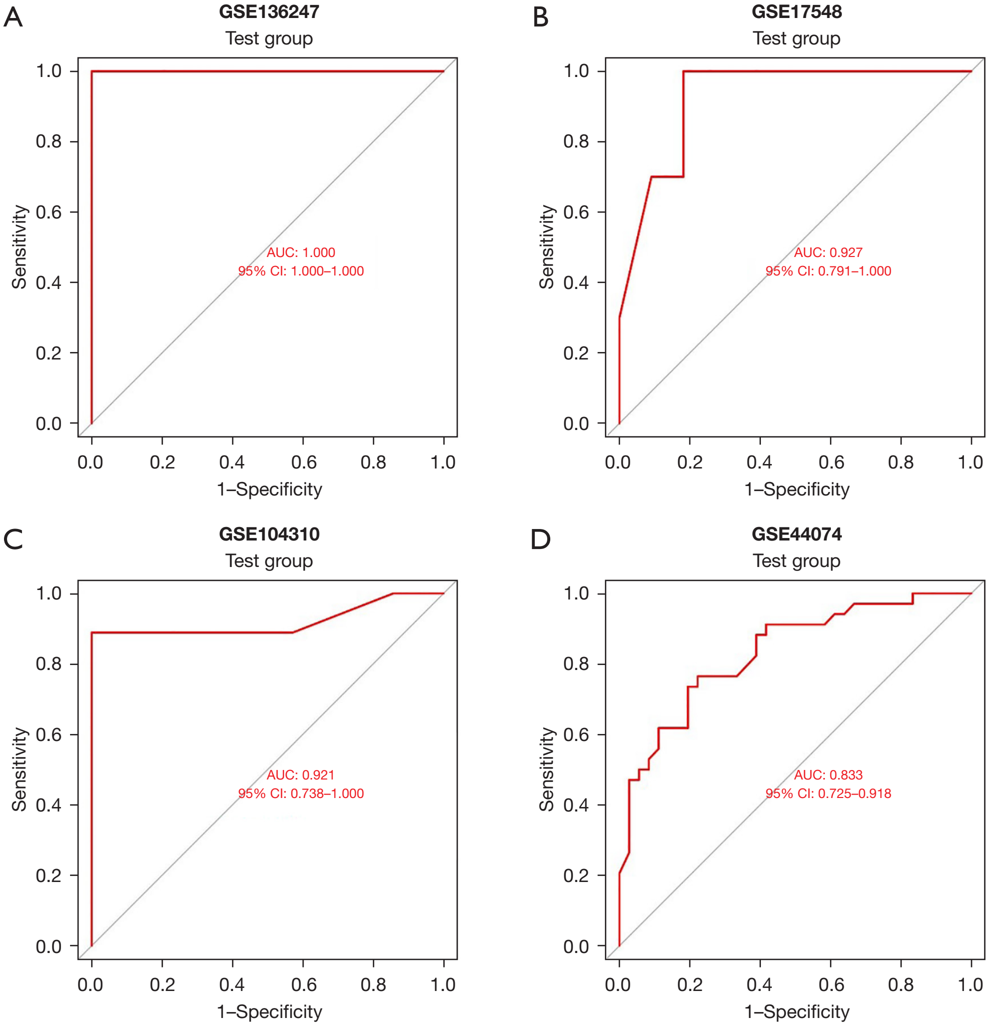

Figure 6 The ROC curve of the ANN model in the validation dataset. (A) GSE136247. (B) GSE17548. (C) GSE104310. (D) GSE44074. GSE, Gene Expression Omnibus Series; AUC, area under the ROC curve; ROC, receiver operating characteristic; CI, confidence interval; ANN, artificial neural network.

acetyltransferase 1 (*ACAT1*), the five HBV-related genes were identified for constructing a prognostic model, which were capable of accurately differentiating HBV patients from non-HBV patients with HCC (52). Integrated analysis of the microbiome and host transcriptome revealed that six important microbial markers associated with the tumor immune microenvironment or bile acid metabolism showed good classification performance for discriminating 5-year survival and 2-year disease-free survival (53). LncRNA was also a potential diagnostic biomarker for HBV-related HCC, and AL356056.2, AL445524.1, TRIM52-AS1, AC093642.1, EHMT2-AS1, AC003991.1, AC008040.1, LINC00844 and LINC01018 were screened out by ML (54). Based on the data from the hospital authority data collaboration lab, 124,006 patients with chronic viral hepatitis (CVH) with complete data were included to build the models, and HCC ridge score (HCC-RS) from the ridge regression ML model accurately predicted HCC in patients with CVH (55). In addition, another study identified noninvasive biomarkers by applying a urinary proteomic strategy (56).

Infiltrating immune cells, a component of the tumor microenvironment, are involved in many processes, including tumor growth, invasion and metastasis. Accumulating evidence has shown that HCC tumors harbour a significant level of immune cell infiltration, and

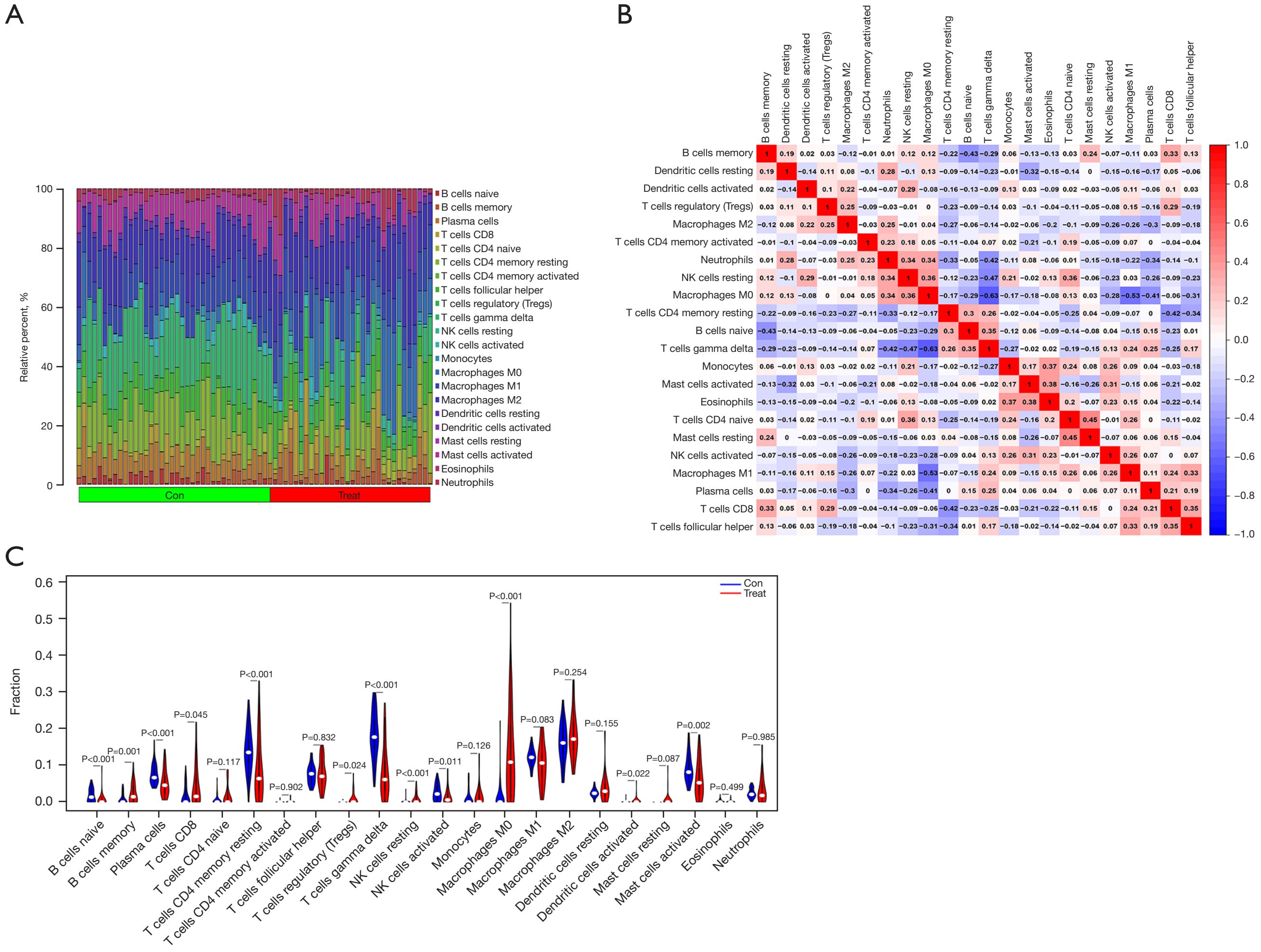

Figure 7 Immune cell infiltration in HBV-related HCC tissues and non-cancerous liver tissue with HBV. (A) The compositions of 22 immune cell types in each sample were shown in a histogram. (B) The correlations of 22 types of immune cells in HBV-related HCC tissues were evaluated. Red: positive correlation; blue: negative correlation. (C) Wilcoxon test was conducted to analyse the different immune cell infiltrates in HBV-related HCC and HBV non-cancerous liver tissues. NK, natural killer; HBV, hepatitis B virus; HCC, hepatocellular carcinoma.

the status of immune cell infiltration and its characteristics are usually associated with different prognostic outcomes (57,58). The ratio of group 2 innate lymphoid cells (ILC2s) to ILC1s increased from non-tumor to tumor tissue in the majority of the HCC patients, and the high ILC2/ILC1 ratio were correlated to better patient survival rates (59). In this study, the density of B cells memory, T cells CD8, Tregs, NK cells resting, macrophages M0, dendritic cells activated in tumor tissues significantly increased compared with non-cancerous liver tissues with HBV. In contrast, the density of B cells naïve, plasma cells, T cells CD4 memory resting, T cells gamma delta, NK cells activated, mast cells activated in HBV-related HCC tissues significantly decreased. T cells, B cells, NK cells, macrophages and mast cells have been previously reported to be present in immune cell infiltrates of HCC and play essential roles in the development, prognosis and immunotherapy treatment of HCC. High densities of naïve B cells and plasma cells were associated with superior survival (60). The antitumor or tumor-promoting effects of tumor-infiltrating lymphocytes depend on the proportion of the lymphocyte subsets constituent in the tumor microenvironment, and T lymphocytes are the primary tumor-infiltrating lymphocytes (TILs) cells in HCC (61). The mechanism of mast cell activation in HCC is unclear, but its activation facilitates immune escape and resultant tumor growth (58). More importantly, HBV-specific CD8$^+$ T cells, HBV-non-specific CD8$^+$, CD4$^+$ T, B and NK/NKT cells are all involved in the development of HBV-related HCC (62).

This study has some limitations. First, HCC exhibits high heterogeneity, which contains etiologic, geographic and molecular heterogeneity. Molecular heterogeneity can be further classified into interpatient, intertumor and intratumor heterogeneity (63). The HBV-related HCC diagnosis model using an ANN was solely based on gene expression data. Therefore, it is difficult to use a single model to accurately diagnose HCC at an early stage, although the model performed satisfactorily on the training and validation datasets. Second, the number of samples used for the construction and validation of this model was relatively small. Third, subsequent confirmatory experiments and clinical practice are needed to further monitor the accuracy and stability of the diagnostic model.

Conclusions

In conclusion, a combination of three datasets' expression data was used to select important variables through RF. An ANN model was formulated for the early diagnosis and screening of HBV-related HCC. Finally, the ratio of infiltrating immune cells in non-cancerous liver tissues from HBV patients and HBV-related HCC tissues was assessed. The findings give a deeper and more comprehensive understanding of the occurrence and progression of HCC and its association with HBV and a valuable reference for the early screening and directions for improving the clinical efficacy of HBV-related HCC.

Acknowledgments

Funding: This work was supported by the China Postdoctoral Science Foundation (No. 2022M713535), the Provincial Natural Science Foundation of Hunan (No. 2023JJ41005), and the Health Research Project of Hunan Provincial Health Commission (No. B202309017571).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-1197/rc

Peer Review File: Available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-1197/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tcr.amegroups.com/article/view/10.21037/tcr-23-1197/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. Hepatocellular carcinoma. Nat Rev Dis Primers 2021;7:7.
2. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
3. El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. Gastroenterology 2012;142:1264-1273.e1.
4. Li J, Cheng L, Jia H, et al. IFN-γ facilitates liver fibrogenesis by CD161+CD4+ T cells through a regenerative IL-23/IL-17 axis in chronic hepatitis B virus infection. Clin Transl Immunology 2021;10:e1353.
5. Pandyarajan V, Govalan R, Yang JD. Risk Factors and Biomarkers for Chronic Hepatitis B Associated Hepatocellular Carcinoma. Int J Mol Sci 2021;22:479.
6. European Association for the Study of the Liver. Electronic address: easloffice@easloffice; . EASL Clinical Practice Guidelines on haemochromatosis. J Hepatol 2022;77:479-502.
7. Bertino G, Neri S, Bruno CM, et al. Diagnostic and

prognostic value of alpha-fetoprotein, des-γ-carboxy prothrombin and squamous cell carcinoma antigen immunoglobulin M complexes in hepatocellular carcinoma. Minerva Med 2011;102:363-71.
8. Dai M, Chen X, Liu X, et al. Diagnostic Value of the Combination of Golgi Protein 73 and Alpha-Fetoprotein in Hepatocellular Carcinoma: A Meta-Analysis. PLoS One 2015;10:e0140067.
9. Shang S, Plymoth A, Ge S, et al. Identification of osteopontin as a novel marker for early hepatocellular carcinoma. Hepatology 2012;55:483-90.
10. Ahn JC, Teng PC, Chen PJ, et al. Detection of Circulating Tumor Cells and Their Implications as a Biomarker for Diagnosis, Prognostication, and Therapeutic Monitoring in Hepatocellular Carcinoma. Hepatology 2021;73:422-36.
11. Beudeker BJB, Boonstra A. Circulating biomarkers for early detection of hepatocellular carcinoma. Therap Adv Gastroenterol 2020;13:1756284820931734.
12. Greener JG, Kandathil SM, Moffat L, et al. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2022;23:40-55.
13. Van Calster B, Wynants L. Machine Learning in Medicine. N Engl J Med 2019;380:2588.
14. Inturi AR, Manikandan VM, Kumar MN, et al. Synergistic Integration of Skeletal Kinematic Features for Vision-Based Fall Detection. Sensors (Basel) 2023;23:6283.
15. Zhao N, Charland K, Carabali M, et al. Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. PLoS Negl Trop Dis 2020;14:e0008056.
16. Sampa MB, Hossain MN, Hoque MR, et al. Blood Uric Acid Prediction With Machine Learning: Model Development and Performance Comparison. JMIR Med Inform 2020;8:e18331.
17. Azimi P, Mohammadi HR, Benzel EC, et al. Artificial neural networks in neurosurgery. J Neurol Neurosurg Psychiatry 2015;86:251-6.
18. Renganathan V. Overview of artificial neural network models in the biomedical domain. Bratisl Lek Listy 2019;120:536-40.
19. Nakonieczna S, Grabarska A, Kukula-Koch W. The Potential Anticancer Activity of Phytoconstituents against Gastric Cancer-A Review on In Vitro, In Vivo, and Clinical Studies. Int J Mol Sci 2020;21:8307.
20. Deng YB, Nagae G, Midorikawa Y, et al. Identification of genes preferentially methylated in hepatitis C virus-related hepatocellular carcinoma. Cancer Sci 2010;101:1501-10.
21. Melis M, Diaz G, Kleiner DE, et al. Viral expression and molecular profiling in liver tissue versus microdissected hepatocytes in hepatitis B virus-associated hepatocellular carcinoma. J Transl Med 2014;12:230.
22. Wang SM, Ooi LL, Hui KM. Identification and validation of a novel gene signature associated with the recurrence of human hepatocellular carcinoma. Clin Cancer Res 2007;13:6275-83.
23. Yildiz G, Arslan-Ergul A, Bagislar S, et al. Genome-wide transcriptional reorganization associated with senescence-to-immortality switch during human hepatocellular carcinogenesis. PLoS One 2013;8:e64016.
24. Ueda T, Honda M, Horimoto K, et al. Gene expression profiling of hepatitis B- and hepatitis C-related hepatocellular carcinoma using graphical Gaussian modeling. Genomics 2013;101:238-48.
25. Cerapio JP, Marchio A, Cano L, et al. Global DNA hypermethylation pattern and unique gene expression signature in liver cancer from patients with Indigenous American ancestry. Oncotarget 2021;12:475-92.
26. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 2019;10:1523.
27. Tian Y, Yang J, Lan M, et al. Construction and analysis of a joint diagnosis model of random forest and artificial neural network for heart failure. Aging (Albany NY) 2020;12:26221-35.
28. Friendly M. Corrgrams: Exploratory Displays for Correlation Matrices. The American Statistician 2002;56:316-24.
29. Xie S, Jiang X, Zhang J, et al. Identification of significant gene and pathways involved in HBV-related hepatocellular carcinoma by bioinformatics analysis. PeerJ 2019;7:e7408.
30. Dey P. Artificial neural network in diagnostic cytology. Cytojournal 2022;19:27.
31. Sha M, Cao J, Zong ZP, et al. Identification of genes predicting unfavorable prognosis in hepatitis B virus-associated hepatocellular carcinoma. Ann Transl Med 2021;9:975.
32. Liao X, Yu T, Yang C, et al. Comprehensive investigation of key biomarkers and pathways in hepatitis B virus-related hepatocellular carcinoma. J Cancer 2019;10:5689-704.
33. Chen X, Liao L, Li Y, et al. Screening and Functional Prediction of Key Candidate Genes in Hepatitis B Virus-Associated Hepatocellular Carcinoma. Biomed Res Int 2020;2020:7653506.
34. Qiang R, Zhao Z, Tang L, et al. Identification of 5

Hub Genes Related to the Early Diagnosis, Tumour Stage, and Poor Outcomes of Hepatitis B Virus-Related Hepatocellular Carcinoma by Bioinformatics Analysis. Comput Math Methods Med 2021;2021:9991255.

35. Yu M, Xu W, Jie Y, et al. Identification and validation of three core genes in p53 signaling pathway in hepatitis B virus-related hepatocellular carcinoma. World J Surg Oncol 2021;19:66.
36. Shi Z, Gan G, Gao X, et al. Kynurenine catabolic enzyme KMO regulates HCC growth. Clin Transl Med 2022;12:e697.
37. Jin H, Zhang Y, You H, et al. Prognostic significance of kynurenine 3-monooxygenase and effects on proliferation, migration, and invasion of human hepatocellular carcinoma. Sci Rep 2015;5:10466.
38. Xia Z, Huang M, Zhu Q, et al. Cadherin Related Family Member 2 Acts As A Tumor Suppressor By Inactivating AKT In Human Hepatocellular Carcinoma. J Cancer 2019;10:864-73.
39. Zhang G, Su L, Lv X, et al. A novel tumor doubling time-related immune gene signature for prognosis prediction in hepatocellular carcinoma. Cancer Cell Int 2021;21:522.
40. Wang W, Huang P, Zhang L, et al. Antitumor efficacy of C-X-C motif chemokine ligand 14 in hepatocellular carcinoma in vitro and in vivo. Cancer Sci 2013;104:1523-31.
41. Yang G, Liang Y, Zheng T, et al. FCN2 inhibits epithelial-mesenchymal transition-induced metastasis of hepatocellular carcinoma via TGF-β/Smad signaling. Cancer Lett 2016;378:80-6.
42. Wang S, Song Z, Tan B, et al. Identification and Validation of Hub Genes Associated With Hepatocellular Carcinoma Via Integrated Bioinformatics Analysis. Front Oncol 2021;11:614531.
43. Lin T, Zhang E, Mai PP, et al. CXCL2/10/12/14 are prognostic biomarkers and correlated with immune infiltration in hepatocellular carcinoma. Biosci Rep 2021;41:BSR20204312.
44. Fu J, Li M, Wu DC, et al. Increased Expression of CAP2 Indicates Poor Prognosis in Hepatocellular Carcinoma. Transl Oncol 2015;8:400-6.
45. Lai X, Wu YK, Hong GQ, et al. A Novel Gene Signature Based on CDC20 and FCN3 for Prediction of Prognosis and Immune Features in Patients with Hepatocellular Carcinoma. J Immunol Res 2022;2022:9117205.
46. Hu K, Wang ZM, Li JN, et al. CLEC1B Expression and PD-L1 Expression Predict Clinical Outcome in Hepatocellular Carcinoma with Tumor Hemorrhage. Transl Oncol 2018;11:552-8.
47. Sakamoto M, Mori T, Masugi Y, et al. Candidate molecular markers for histological diagnosis of early hepatocellular carcinoma. Intervirology 2008;51 Suppl 1:42-5.
48. Shibata R, Mori T, Du W, et al. Overexpression of cyclase-associated protein 2 in multistage hepatocarcinogenesis. Clin Cancer Res 2006;12:5363-8.
49. Ghasemi R, Ghaffari SH, Momeny M, et al. Multitargeting and antimetastatic potentials of silibinin in human HepG-2 and PLC/PRF/5 hepatoma cells. Nutr Cancer 2013;65:590-9.
50. Lin Y, Chen BM, Yu XL, et al. Suppressed Expression of CXCL14 in Hepatocellular Carcinoma Tissues and Its Reduction in the Advanced Stage of Chronic HBV Infection. Cancer Manag Res 2019;11:10435-43.
51. Hoang TV, Toan NL, Song le H, et al. Ficolin-2 levels and FCN2 haplotypes influence hepatitis B infection outcome in Vietnamese patients. PLoS One 2011;6:e28113.
52. Ma K, Wu H, Ji L. Construction of HBV gene-related prognostic and diagnostic models for hepatocellular carcinoma. Front Genet 2023;13:1065644.
53. Huang H, Ren Z, Gao X, et al. Integrated analysis of microbiome and host transcriptome reveals correlations between gut microbiota and clinical outcomes in HBV-related hepatocellular carcinoma. Genome Med 2020;12:102.
54. Nong S, Chen X, Wang Z, et al. Potential lncRNA Biomarkers for HBV-Related Hepatocellular Carcinoma Diagnosis Revealed by Analysis on Coexpression Network. Biomed Res Int 2021;2021:9972011.
55. Wong GL, Hui VW, Tan Q, et al. Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis. JHEP Rep 2022;4:100441.
56. Zhao Y, Li Y, Liu W, et al. Identification of noninvasive diagnostic biomarkers for hepatocellular carcinoma by urinary proteomics. J Proteomics 2020;225:103780.
57. Zheng C, Zheng L, Yoo JK, et al. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell 2017;169:1342-1356.e16.
58. Rohr-Udilova N, Klinglmüller F, Schulte-Hermann R, et al. Deviations of the immune cell landscape between healthy liver and hepatocellular carcinoma. Sci Rep 2018;8:6220.
59. Heinrich B, Gertz EM, Schäffer AA, et al. The tumour microenvironment shapes innate lymphoid cells in patients with hepatocellular carcinoma. Gut 2022;71:1161-75.
60. Zhang Z, Ma L, Goswami S, et al. Landscape of

infiltrating B cells and their clinical significance in human hepatocellular carcinoma. Oncoimmunology 2019;8:e1571388.

61. Zheng X, Jin W, Wang S, et al. Progression on the Roles and Mechanisms of Tumor-Infiltrating T Lymphocytes in Patients With Hepatocellular Carcinoma. Front Immunol 2021;12:729705.

62. Chen Y, Tian Z. HBV-Induced Immune Imbalance in the Development of HCC. Front Immunol 2019;10:2048.

63. Dhanasekaran R. Deciphering Tumor Heterogeneity in Hepatocellular Carcinoma (HCC)-Multi-Omic and Singulomic Approaches. Semin Liver Dis 2021;41:9-18.

Cite this article as: Jiang X, Hu J, Xie S. Construction and validation of a joint diagnosis model based on random forest and artificial intelligence network for hepatitis B-related hepatocellular carcinoma. Transl Cancer Res 2024;13(2):1068-1082. doi: 10.21037/tcr-23-1197

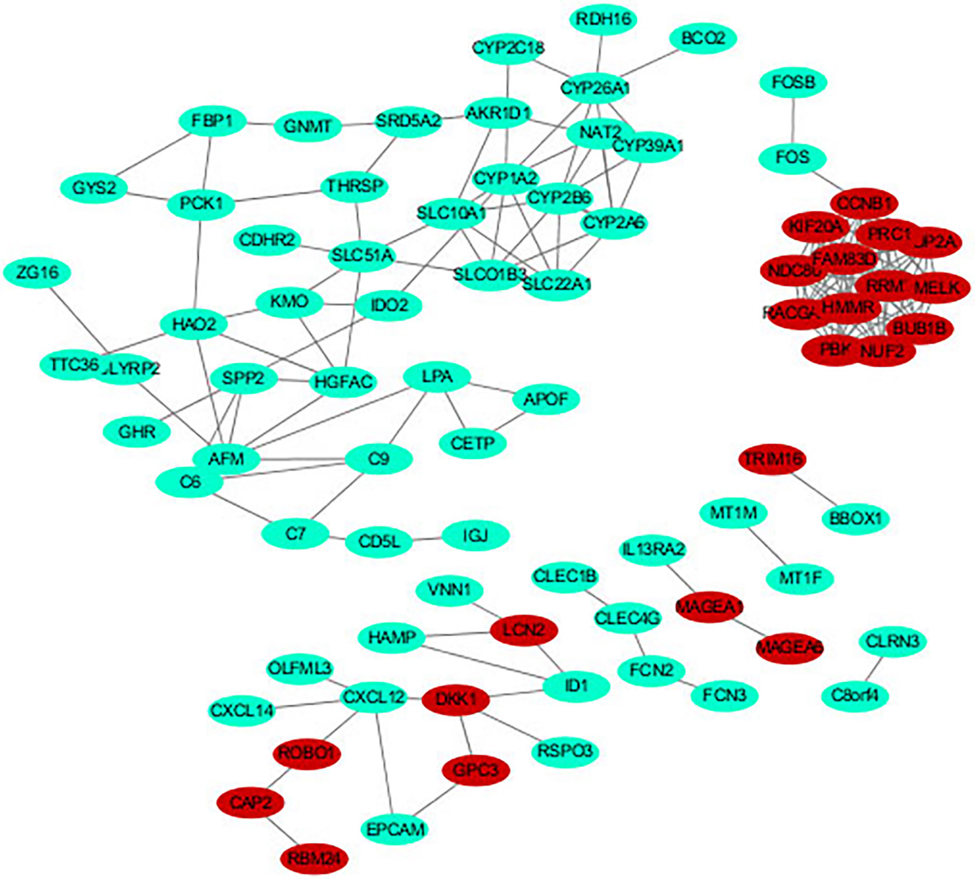

Figure S1 PPI network of differentially expressed genes. Circles represent genes (the red represents log FC >0, the green represents log FC <0), lines represent interactions between gene encoded proteins and line colors represent evidence of interactions between proteins. PPI, protein-protein interaction; FC, fold change.

https://dx.doi.org/10.21037/tcr-23-1197

Table S1 DEGs were identified in the dataset merged by GSE19665, GSE55092, and GSE121248

ID	logFC	AveExpr	t	P value	adj. P value	B
CXCL14	–2.544576	6.34945308	–27.423644	1.90E–78	2.59E–74	168.068268
CLEC1B	–3.4329001	7.42679483	–27.390276	2.40E–78	2.59E–74	167.836565
FCN2	–2.2206992	6.39988521	–26.530286	1.05E–75	7.59E–72	161.815421
OIT3	–3.5715519	7.36700563	–24.658246	8.19E–70	4.43E–66	148.382336
LINC01093	–4.0451304	7.76034823	–23.768316	6.01E–67	1.86E–63	141.842915
CLEC4G	–3.2462654	6.74354921	–23.709637	9.32E–67	2.52E–63	141.408335
FCN3	–3.4437076	8.8527026	–23.430768	7.52E–66	1.63E–62	139.337366
CAP2	2.39426019	8.04539907	22.7784117	1.03E–63	1.86E–60	134.456812
GPR128	–2.9182295	6.68192639	–22.546616	6.01E–63	9.99E–60	132.710751
RSPO3	–2.1847306	6.30050748	–22.391964	1.95E–62	3.01E–59	131.542389
CLEC4M	–2.8142277	6.64806351	–22.266048	5.10E–62	7.35E–59	130.589135
CRHBP	–3.156771	6.81679025	–22.228427	6.80E–62	9.19E–59	130.30398
CDHR2	–2.1015227	8.03715736	–22.019731	3.36E–61	4.27E–58	128.719267
HAMP	–4.3980972	9.36791144	–21.232471	1.45E–58	1.49E–55	122.698906
SLC25A47	–2.6340965	8.7318427	–21.135064	3.09E–58	2.90E–55	121.94949
TTC36	–2.8921356	8.35113777	–21.033574	6.78E–58	6.11E–55	121.167633
RACGAP1	2.16183945	7.03931053	20.7021333	8.95E–57	6.92E–54	118.607137
TOP2A	2.14988566	5.27272419	20.4023272	9.32E–56	6.50E–53	116.281836
CYP26A1	–2.8416724	7.09686776	–20.303384	2.02E–55	1.37E–52	115.512569
HMMR	2.12203833	6.27459095	20.27307	2.57E–55	1.68E–52	115.276705
PLAC8	–2.4756285	7.01370729	–20.040852	1.59E–54	1.01E–51	113.467093
CNDP1	–3.3412824	7.86986649	–19.605404	4.91E–53	2.95E–50	110.060993
IL1RAP	–2.0942222	8.89466727	–19.481031	1.31E–52	7.66E–50	109.085212
IDO2	–2.8412497	7.18719372	–19.45993	1.55E–52	8.82E–50	108.919536
KCNN2	–3.2047402	6.40628195	–19.29655	5.65E–52	2.91E–49	107.635561
CYP1A2	–3.018836	9.90220702	–19.230363	9.54E–52	4.69E–49	107.114809
CXCL12	–2.9655968	9.16289723	–18.719083	5.54E–50	2.35E–47	103.081168
NDC80	2.14393493	5.78888634	18.5945591	1.49E–49	6.22E–47	102.095982
CETP	–2.0759883	6.90653043	–18.39339	7.44E–49	2.87E–46	100.502276
CD5L	–2.2727505	7.70742354	–18.379736	8.30E–49	3.15E–46	100.394011
KMO	–2.4672064	8.69821451	–18.311186	1.43E–48	5.35E–46	99.8503035
PRC1	2.12772385	6.72268565	17.9403021	2.79E–47	9.42E–45	96.9039011
CCNB1	2.11571785	5.74060892	17.9170064	3.36E–47	1.08E–44	96.7185817
FAM83D	2.05580848	6.8791897	17.8029773	8.37E–47	2.66E–44	95.8110672
TMEM27	–2.7061758	7.40032287	–17.792946	9.08E–47	2.84E–44	95.7312007
MARCO	–2.5606435	7.67943728	–17.711678	1.74E–46	5.16E–44	95.0839883
ZG16	–2.2926421	7.94482155	–17.628254	3.40E–46	9.81E–44	94.4192761
RRM2	2.2228061	7.73375779	17.5456553	6.60E–46	1.83E–43	93.7608261
GPR158	2.04736143	5.3607137	17.4914294	1.02E–45	2.79E–43	93.3283943
NPY1R	–2.1036549	5.88291012	–17.483499	1.09E–45	2.94E–43	93.2651382
HAO2	–2.8246959	8.80197804	–17.480559	1.11E–45	2.97E–43	93.2416906
NAT2	–2.4261428	9.04103488	–17.44686	1.46E–45	3.85E–43	92.9728729
PBK	2.21313299	5.36001211	17.4019865	2.09E–45	5.45E–43	92.6148493
ROBO1	2.4804005	9.2531166	17.3987845	2.15E–45	5.53E–43	92.5892988
MELK	2.04803433	7.29807767	17.3634664	2.85E–45	7.17E–43	92.3074491
OLFML3	–2.1026019	7.52988506	–17.168831	1.36E–44	3.14E–42	90.7533463
KIF20A	2.00793763	5.72635009	17.1368028	1.76E–44	4.02E–42	90.4974816
BUB1B	2.15908291	6.9202473	17.1126416	2.14E–44	4.78E–42	90.3044404
RBM24	2.4794942	6.2873214	16.8344272	2.01E–43	4.19E–41	88.0802986
HGFAC	–2.3563959	7.65007814	–16.697045	6.08E–43	1.17E–40	86.981248
APOF	–2.7303888	9.09444321	–16.680151	6.97E–43	1.31E–40	86.8460663
FBP1	–2.5565585	10.9963788	–16.438405	4.89E–42	8.46E–40	84.9110724
FOS	–2.7657023	8.91756115	–16.385063	7.51E–42	1.26E–39	84.4839885
BCO2	–2.3159779	8.21043643	–16.326855	1.20E–41	1.97E–39	84.0178905
FREM2	–2.7410914	5.7221795	–16.241833	2.38E–41	3.79E–39	83.3370196
MT1F	–2.396978	11.0882789	–16.20254	3.27E–41	5.09E–39	83.0223371
C8orf4	–2.0608662	9.57693862	–15.982202	1.93E–40	2.79E–38	81.2575406
SRPX	–2.7167506	7.13361498	–15.6451	2.93E–39	3.64E–37	78.5575769
CYP39A1	–2.1837453	7.46419048	–15.622612	3.51E–39	4.29E–37	78.3774957
NUF2	2.05738451	4.52875921	15.6027624	4.12E–39	4.96E–37	78.2185389
LPA	–2.4902639	8.38327464	–15.329274	3.74E–38	3.91E–36	76.0290637
ZIC2	2.62567386	4.69785113	15.1071722	2.24E–37	2.16E–35	74.2521216
GPC3	2.47710687	6.43662826	15.0515581	3.50E–37	3.25E–35	73.8073865
GHR	–2.5465291	10.322801	–15.04758	3.61E–37	3.34E–35	73.7755768
GRAMD1C	–2.1008286	8.06686749	–15.033167	4.06E–37	3.72E–35	73.6603369
FOSB	–2.6044033	8.196666	–14.987522	5.86E–37	5.31E–35	73.2954236
MT1M	–3.1518953	7.84322898	–14.836298	1.98E–36	1.71E–34	72.0869752
LOC344887	2.64539651	5.44903398	14.8151245	2.35E–36	2.00E–34	71.91784
ID1	–2.1167992	8.30653243	–14.579245	1.56E–35	1.22E–33	70.0349784
SPP2	–2.7502875	10.270496	–14.545955	2.04E–35	1.56E–33	69.7694576
SRD5A2	–2.0287757	8.02767535	–14.483538	3.37E–35	2.51E–33	69.2717771
AKR1D1	–3.2961038	10.1091412	–14.458953	4.11E–35	3.03E–33	69.0758127
C7	–2.5741731	8.09244358	–14.45236	4.33E–35	3.16E–33	69.023258
COL15A1	2.18513587	5.95514173	14.3498401	9.85E–35	6.87E–33	68.2064634
ANXA10	–2.1660479	9.13134905	–14.295249	1.53E–34	1.05E–32	67.7717762
SLC22A1	–3.3089567	10.0527123	–14.277241	1.76E–34	1.19E–32	67.6284219
CA2	–2.2707491	9.53539549	–14.276749	1.77E–34	1.19E–32	67.6245102
CR936796	2.11220051	5.93023271	14.1771121	3.93E–34	2.58E–32	66.8317385
PGLYRP2	–2.5835523	9.09962977	–14.078964	8.63E–34	5.38E–32	66.0514629
CYP2C18	–2.0125866	9.28734037	–13.983148	1.86E–33	1.10E–31	65.2903835
PRG4	–2.1506923	6.586506	–13.962553	2.19E–33	1.29E–31	65.1268907
GBA3	–2.2881345	9.43008856	–13.817351	6.99E–33	3.91E–31	63.9751069
RDH16	–2.0574012	10.7701176	–13.596486	4.07E–32	2.11E–30	62.2265497
IL13RA2	–2.225249	6.13242995	–13.582648	4.54E–32	2.33E–30	62.1171389
IGJ	–3.2619562	7.8710689	–13.548463	5.96E–32	2.98E–30	61.846938
TRIM16	2.11603793	6.69760625	13.5108216	8.04E–32	3.98E–30	61.5495497
CRNDE	2.17311722	5.71512974	13.3359465	3.23E–31	1.49E–29	60.1697846
CLRN3	–2.5289685	8.46410842	–13.332428	3.32E–31	1.53E–29	60.1420575
BBOX1	–2.5907476	8.12384816	–13.048445	3.15E–30	1.32E–28	57.9085754
THRSP	–2.279467	8.37565341	–12.933737	7.80E–30	3.09E–28	57.0091035
SMPX	2.017613	5.41069708	12.910697	9.36E–30	3.65E–28	56.8286399
SLCO1B3	–3.0558913	9.40973033	–12.891284	1.09E–29	4.19E–28	56.6766329
CYP2A6	–2.4142187	10.2796799	–12.870158	1.29E–29	4.86E–28	56.5112771
CNTN3	–2.6745936	6.62062451	–12.683901	5.58E–29	1.98E–27	55.0559304
GYS2	–2.4771606	9.22389918	–12.569326	1.37E–28	4.53E–27	54.1630753
VNN1	–2.5145344	9.53649745	–12.384047	5.86E–28	1.81E–26	52.7233548
GNMT	–2.3020348	9.38614604	–12.321662	9.54E–28	2.87E–26	52.2397798
C9	–3.6280301	10.7365082	–12.142858	3.84E–27	1.07E–25	50.857302
ACOT12	–2.1425298	9.4699936	–12.044998	8.23E–27	2.22E–25	50.102961
SPINK1	3.50352604	8.1481923	11.8345504	4.20E–26	1.06E–24	48.4865586
FAM110C	–2.1205849	8.59651321	–11.501186	5.46E–25	1.21E–23	45.9434806
PCK1	–2.407529	11.7446286	–11.496659	5.65E–25	1.25E–23	45.909104
COX7B2	2.11849345	4.84578431	11.4824161	6.30E–25	1.39E–23	45.8009686
SDS	–2.0349595	8.58071055	–11.384891	1.33E–24	2.82E–23	45.0617115
LCN2	2.31229931	7.70959633	11.2987206	2.56E–24	5.30E–23	44.4102426
C6	–2.0718186	11.5259624	–11.19729	5.54E–24	1.10E–22	43.6455413
MAGEA1	2.07293912	5.77535978	11.1558954	7.59E–24	1.47E–22	43.3341359
AFM	–2.0500275	10.6946934	–11.15097	7.88E–24	1.52E–22	43.297106
CYP2B7P	–2.5012243	8.32072183	–11.035516	1.89E–23	3.51E–22	42.4308315
SLC10A1	–2.0012838	11.0153519	–10.999316	2.48E–23	4.54E–22	42.1598712
LINC01419	2.12198445	4.02663356	10.9909537	2.64E–23	4.82E–22	42.097327
AKR1B10	3.14085593	9.32353039	10.917218	4.61E–23	8.17E–22	41.5465557
SLC51A	–2.0259855	9.58535298	–10.239472	7.15E–21	1.01E–19	36.5512572
MAGEA6	2.51197733	5.06016824	9.78109646	2.01E–19	2.44E–18	33.2490772
DKK1	2.50349755	5.58884284	9.3967322	3.14E–18	3.35E–17	30.5337099
EPCAM	–2.151953	7.91948837	–8.3883066	3.26E–15	2.54E–14	23.6767844

DEGs, differentially expressed genes; GSE, Gene Expression Omnibus Series.

Table S2 Top GO enrichment terms of differentially expressed genes associated with hepatitis B-related hepatocellular carcinoma

Category	Term	Count	%	P value
BP	GO:0032787: monocarboxylic acid metabolic process	13	11.3	6.76083E−08
BP	GO:0009617: response to bacterium	12	10.43	3.38844E−05
BP	GO:1901652: response to peptide	11	9.57	1.28825E−06
BP	GO:0040008: regulation of growth	10	8.7	0.000141254
BP	GO:0071466: cellular response to xenobiotic stimulus	9	7.83	3.71535E−08
BP	GO:0034754: cellular hormone metabolic process	8	6.96	3.80189E−08
BP	GO:0000819: sister chromatid segregation	7	6.09	1.65959E−06
BP	GO:0002697: regulation of immune effector process	7	6.09	0.000645654
BP	GO:0015850: organic hydroxy compound transport	6	5.22	3.63078E−05
BP	GO:0045861: negative regulation of proteolysis	6	5.22	0.002344229
BP	GO:0044262: cellular carbohydrate metabolic process	5	4.35	0.000446684
BP	GO:0071902: positive regulation of protein serine/threonine kinase activity	5	4.35	0.001862087
BP	GO:0044070: regulation of anion transport	4	3.48	0.000446684
BP	GO:0051701: biological process involved in interaction with host	4	3.48	0.002951209
BP	GO:2000096: positive regulation of Wnt signaling pathway, planar cell polarity pathway	3	2.61	4.46684E−06
BP	GO:0019835: cytolysis	3	2.61	6.76083E−05
BP	GO:0042537: benzene-containing compound metabolic process	3	2.61	0.000147911
BP	GO:0009595: detection of biotic stimulus	3	2.61	0.000446684
BP	GO:0043277: apoptotic cell clearance	3	2.61	0.000630957
BP	GO:0040014: regulation of multicellular organism growth	3	2.61	0.001995262
CC	GO:0062023: collagen-containing extracellular matrix	10	8.7	6.45654E−06
CC	GO:0005819: spindle	8	6.96	0.000234423
CC	GO:0009897: external side of plasma membrane	7	6.09	0.001995262
CC	GO:0072562: blood microparticle	6	5.22	2.04174E−05
CC	GO:0016323: basolateral plasma membrane	5	4.35	0.001862087
CC	GO:0005579: membrane attack complex	3	2.61	1.86209E−06
CC	GO:0000940: outer kinetochore	3	2.61	1.14815E−05
CC	GO:0034358: plasma lipoprotein particle	3	2.61	0.000354813
MF	GO:0016491: oxidoreductase activity	16	13.91	2.0893E−08
MF	GO:0042803: protein homodimerization activity	9	7.83	0.00134896
MF	GO:0030246: carbohydrate binding	8	6.96	1.0965E−05
MF	GO:0033218: amide binding	8	6.96	0.00017378
MF	GO:0005319: lipid transporter activity	6	5.22	3.6308E−05
MF	GO:0016614: oxidoreductase activity, acting on CH-OH group of donors	5	4.35	0.00019055
MF	GO:1901618: organic hydroxy compound transmembrane transporter activity	4	3.48	5.1286E−05
MF	GO:0038024: cargo receptor activity	4	3.48	0.00028184
MF	GO:0031406: carboxylic acid binding	4	3.48	0.0042658
MF	GO:0005201: extracellular matrix structural constituent	4	3.48	0.00436516
MF	GO:0030414: peptidase inhibitor activity	4	3.48	0.0057544
MF	GO:0016829: lyase activity	4	3.48	0.00724436
MF	GO:0042834: peptidoglycan binding	3	2.61	4.2658E−05
MF	GO:0016709: oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NAD(P)H as one donor, and incorporation of one atom of oxygen	3	2.61	0.00047863
MF	GO:0051213: dioxygenase activity	3	2.61	0.00549541
MF	GO:0004896: cytokine receptor activity	3	2.61	0.00616595

GO, Gene Ontology; BP, biological process; CC, cellular component; MF, molecular function.

Table S3 KEGG pathway analysis of differentially expressed genes associated with hepatitis B-related hepatocellular carcinoma

Term	Count	%	P value
hsa04976: Bile secretion	6	5.22	1.20226E–06
hsa04060: Cytokine-cytokine receptor interaction	5	4.35	0.005495409
hsa00232: Caffeine metabolism	3	2.61	1.07152E–06
hsa00380: Tryptophan metabolism	3	2.61	0.000549541
hsa00140: Steroid hormone biosynthesis	3	2.61	0.001659587
hsa04610: Complement and coagulation cascades	3	2.61	0.004265795
hsa04657: IL-17 signaling pathway	3	2.61	0.005623413
hsa04922: Glucagon signaling pathway	3	2.61	0.007943282
hsa01200: Carbon metabolism	3	2.61	0.009772372

KEGG, Kyoto Encyclopedia of Genes and Genomes.