Original Article

Identification of the key gene for hepatocellular carcinoma based on bioinformatics and machine learning and experimental verification

Jin Lu[1,2#], Junjie Ma[3#], Can Yu[4], Shaoyang Lu[4], Xueying Zhao[1], Lei Zhang[2,5]

[1]Department of Human Anatomy, Bengbu Medical University, Bengbu, China; [2]Key Laboratory of Digital Medicine and Smart Health, Bengbu Medical University, Bengbu, China; [3]Department of Hepatobiliary and Pancreatic Surgery, The Third Xiangya Hospital of Central South University, Changsha, China; [4]Department of Clinical Medicine, Bengbu Medical University, Bengbu, China; [5]Department of General Surgery, The Second Affiliated Hospital of Bengbu Medical University, Bengbu, China

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

[#]These authors contributed equally to this work.

Correspondence to: Dr. Lei Zhang, MD. Key Laboratory of Digital Medicine and Smart Health, Bengbu Medical University, Bengbu 233030, China; Department of General Surgery, The Second Affiliated Hospital of Bengbu Medical University, 633 Longhua Road, Bengbu 233000, China. Email: leizhang@bbmu.edu.cn.

Background: Hepatocellular carcinoma (HCC) is a severe hazard to human health and has a high fatality rate. While deregulated gene expression has been widely linked to hepatocarcinogenesis, many details of how these alterations drive tumor initiation and progression remain to be elucidated. We therefore combined bioinformatics and machine learning strategies to screen for and validate candidate driver genes in HCC.

Methods: Three datasets (GSE78737, GSE98383, and GSE121248) were obtained from the Gene Expression Omnibus (GEO) database. GSE78737 and GSE98383 were combined to form the training set, while GSE121248 was used as the validation set. Initially, differentially expressed genes (DEGs) between HCC and non-HCC (nHCC) in the training set were identified. Enrichment analysis of these DEGs was performed using Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Set Enrichment Analysis (GSEA). To identify diagnostic genes, machine learning algorithms including support vector machine-recursive feature elimination (SVM-RFE) and least absolute shrinkage and selection operator (LASSO) were applied. The validation set was employed to confirm the DEGs. Furthermore, immune cell infiltration differences between nHCC and HCC were analyzed using CIBERSORT. GEPIA2.0 was subsequently used to analyze the prognostic significance of the diagnostic genes in HCC, identifying key genes. Finally, the key genes were validated using data from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), as well as through immunohistochemistry (IHC) experiments, single-cell, and spatial transcriptomics analysis.

Results: A total of 80 DEGs were identified, with 8 upregulated and 72 downregulated. The GO pathways associated with these DEGs were primarily related to responses to alcohol, humoral immune response, vacuolar lumen, chemokine activity, and mannose binding. KEGG pathway analysis revealed that the DEGs were primarily focused on viral protein interactions with cytokines and cytokine receptors. GSEA indicated that the most active processes in HCC included DNA replication, cell cycle, and mismatch repair. Immune cell analysis showed significant overexpression of naive B cells, $CD8^+$ T cells, activated natural killer (NK) cells, M0 macrophages, and dendritic cells in HCC. In contrast, naive $CD4^+$ T cells, gamma delta T cells, and monocytes were significantly lower in HCC compared to nHCC. Machine learning and risk prognosis analysis identified *FAM83D* as a key gene in HCC, serving as an independent variable affecting HCC prognosis. Increased *FAM83D* mRNA and protein expression correlated with poor overall survival and prognosis in HCC patients. Additionally, *FAM83D* expression was significantly related to various immune

cells. Further single-cell analysis revealed that *FAM83D* is predominantly upregulated in malignant cells, and its high expression is strongly associated with poor response to immunotherapy in HCC patients.

Conclusions: *FAM83D* may be a key gene involved in the development and progression of HCC, contributing to early diagnosis and prognosis assessment. It has the potential to serve as a biomarker for HCC.

Keywords: Hepatocellular carcinoma (HCC); machine learning; immune cell infiltration; FAM83D; bioinformatics

Submitted May 21, 2025. Accepted for publication Sep 24, 2025. Published online Nov 26, 2025.
doi: 10.21037/tcr-2025-1067
View this article at: https://dx.doi.org/10.21037/tcr-2025-1067

Introduction

Hepatocellular carcinoma (HCC) is the most common form of liver cancer, accounting for approximately 90% of cases (1). It is reported to be the sixth most common tumor worldwide and the third leading cause of cancer-related mortality (2). Despite significant improvements in overall survival (OS) rates for HCC patients through surgical intervention, chemotherapy, and targeted therapies, the mortality rate remains high. This is primarily due to the fact that early symptoms of HCC are often subtle and difficult to detect, resulting in many patients being diagnosed at an advanced stage of the disease. Although the tumor-node-metastasis (TNM) and Barcelona Clinic Liver Cancer (BCLC) staging systems are still utilized in clinical practice to evaluate HCC prognosis, their accuracy is suboptimal (3,4). Furthermore, patients with HCC at similar stages may exhibit markedly different prognoses. Therefore, there is an urgent need for more accurate prognostic assessment techniques to develop tailored treatment strategies for HCC patients. Research has indicated that tumor diagnosis and survival prognosis are closely associated with cancer susceptibility genes (5). Accordingly, we employed machine learning and bioinformatics approaches to predict key genes in HCC, aiming to facilitate early diagnosis, effective treatment, and comprehensive prognostic evaluation of this condition.

Recently, machine learning has been used in a variety of medical fields (6-8). Machine learning has an advantage over most conventional statistical techniques in that it can discover and identify potential patterns in massive amounts of data (9). High-throughput sequencing technology has made it possible to identify cancer signature genes using machine learning (10). Therefore, researchers have used machine learning to discover cancer prognostic signature genes and categorize tumors. Koppad *et al.* (11) used machine learning to identify candidate colon cancer diagnosis genes. Najm *et al.* (12) created a machine-learning algorithm for predicting popular molecule protein drug targets. All these findings illustrate machine learning's enormous potential in precision medicine research. It can

Highlight box

Key findings

- *FAM83D* is highly expressed in hepatocellular carcinoma (HCC) malignant cells and positively correlates with tumor proliferation, migration, and an immunosuppressive microenvironment.
- Single-cell RNA sequencing revealed that *FAM83D*-high malignant cells display stronger receptor-ligand communication with cancer-associated fibroblasts (CAFs) and endothelial cells, predominantly via the Laminin and MK (midkine) pathways.
- *In vitro* knockdown of *FAM83D* markedly inhibited HCC cell proliferation, migration, and invasion, and reduced M2 macrophage polarization, indicating that *FAM83D* may serve as a target for reprogramming the immune niche.

What is known and what is new?

- *FAM83D* overexpression has been linked to poor prognosis in several solid tumors, but its role in shaping the HCC immune microenvironment has not been systematically defined.
- Our study is the first to demonstrate, at single-cell resolution, that *FAM83D* fosters an immunosuppressive milieu by enhancing malignant cell-CAF/endothelial crosstalk, and that targeting *FAM83D* simultaneously curbs tumor aggressiveness and immune evasion.

What is the implication, and what should change now?

- *FAM83D* may act as both a prognostic biomarker and a novel therapeutic target for HCC. Future pre-clinical studies should evaluate combined *FAM83D* inhibition plus immune-checkpoint blockade to improve patient outcomes.

Table 1 Details on microarray data

Set	Microarray	HCC (n)	nHCC (n)	Platforms
Training	GSE78737	37	66	(HGU133_Plus_2) Affymetrix Human Genome U133 Plus 2.0 Array
	GSE98383	16	58	
Validation	GSE121248	70	37	

HCC, hepatocellular carcinoma; nHCC, non-HCC.

learn to carry out particular classification tasks from high-dimensional gene expression data. However, there has been little research on machine learning aimed at trying to identify signature genes appropriate for HCC diagnosis. Therefore, more in-depth investigations are necessary.

The advent of next-generation sequencing (NGS) technologies and bioinformatics tools has greatly facilitated the identification of novel biomarkers for cancer diagnosis and precision medicine. In this study, we downloaded and analyzed three Gene Expression Omnibus (GEO) datasets using bioinformatics methods to identify differentially expressed genes (DEGs) in HCC. Subsequently, we conducted analyses on these DEGs, including Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Set Enrichment Analysis (GSEA), and immune cell infiltration analyses. Additionally, we employed the least absolute shrinkage and selection operator (LASSO) regression and support vector machine-recursive feature elimination (SVM-RFE) methods to identify key genes among the DEGs. Finally, we validated the transcriptional expression of these key genes using the Gene Expression Profiling Interactive Analysis (GEPIA), The Cancer Genome Atlas (TCGA), The International Cancer Genome Consortium (ICGC) and constructed a risk model. We also collected tissue samples from HCC patients and used immunohistochemistry (IHC) to validate the protein expression of key genes. In summary, our study identified key genes involved in the progression and prognosis of HCC, which are of significant importance for the diagnosis, treatment, and prognostic evaluation of HCC. We present this article in accordance with the TRIPOD and MDAR reporting checklists (available at https://tcr.amegroups.com/article/view/10.21037/tcr-2025-1067/rc).

Methods

Datasets source

The NCBI-GEO (https://www.ncbi.nlm.nih.gov/gds/) database, which has gene profiles, is open to the public and free to use. From the GEO database, we were able to retrieve three microarray datasets (GSE78737, GSE98383, and GSE121248) (*Table 1*). The samples of HCC and non-HCC tissues were then defined as HCC and non-HCC (nHCC), respectively. In addition, depending on the annotation data, we used the "Perl" programming language to switch the probe matrix into a gene matrix. Consequently, we created a training set by combining the GSE78737 and GSE98383 cohorts. Finally, the batch correction was performed, as well as the identification of various genes between the HCC and the nHCC by the R language's "sva" and "limma" packages. The validation set for the following validation was the GSE121248 cohort.

DEGs analysis

We used the "limma" package for screening the DEGs between HCC and nHCC in the training set, setting the screening condition to $|\log_2$ fold change (FC)$|$ >2 and P<0.05. If $\log_2$FC >2, it means that this gene is highly expressed in HCC. If $\log_2$FC <−2, the gene is poorly expressed in HCC. Finally, the results obtained from the analysis were presented via a heat map and a volcano map.

Bio-functional enrichment analysis

We determined the functional enrichment of DEGs for GO, KEGG, and GSEA by using three R packages: "clusterProfiler", " enrichplot", and "ggplot2". Therein, GO terms included biological process (BP), cellular component (CC), and molecular function (MF). For these functional enrichment analyses, we considered it statistically significant if P<0.05.

Immune cell infiltration analysis

We used the CIBERSORT program in R software to compare the infiltration of 22 types of immune cells

between nHCC and HCC. The analysis results are displayed as scatter plots, violin plots, and lollipop plots. Additionally, the association between key diagnostic genes and immune cell infiltration was evaluated.

Machine learning analysis

Utilizing the LASSO and SVM-RFE as machine learning algorithms, the precise HCC biomarkers were screened, and a reliable prediction system was obtained. LASSO regression is useful for processing high-dimensional data and is implemented by the "glmnet" package of R software. SVM-RFE has a gradual elimination of inter-variate interaction features to determine the prediction function in a cross-validated manner and is implemented by the "caret" package of R software. Finally, the receiver operating characteristic (ROC) curve was utilized to check the diagnostic gene accuracy.

Clinical prognostic and early-diagnostic analysis

GEPIA2.0 (http://gepia2.cancer-pku.cn/#index) was used to evaluate the influence of diagnostic genes on OS and prognosis of HCC patients, to determine the key diagnostic genes.

The GSE63898 dataset, including 228 HCC and 168 nHCC samples, was used to validate the performance of the *FAM83D* gene in diagnosing early-stage HCC. Early-stage HCC was defined as BCLC stage 0–A (13).

Risks and prognosis analysis

The key diagnostic genes for HCC were evaluated in relation to patient prognostic risk using data accessed from the TCGA (https://portal.gdc.cancer.gov). The TCGA dataset includes HCC RNA sequencing expression (level 3) data along with corresponding clinical information. Log-rank was used to test the survival differences between the two groups in Kaplan-Meier (K-M) survival analysis, and timeROC analysis was conducted to compare the prediction accuracy of *FAM83D* gene. To select the most significant variables for inclusion in the model, both univariate and multivariate Cox regression analyses were performed, presenting P values, hazard ratio (HR), and 95% confidence interval (CI) for each factor using the forestplot package in R software. A nomogram was generated based on the results of the multivariate Cox proportional hazards analysis, indicating the specific risk of OS for patients, as determined by scores related to each risk factor using the "rms" package in R. The predictive accuracy of mRNA for key diagnostic genes was then compared using ROC analysis.

ICGC validation analysis

In order to further verify the accuracy of the key diagnostic genes, we made an in-depth analysis by using the relevant data in ICGC (https://dcc.icgc.org/releases/current/Projects). RNA-seq data (level 3) and corresponding clinical information for 240 cases of HCC were obtained from the ICGC database. The log-rank test was employed to assess survival differences between the aforementioned two groups using K-M survival analysis. Additionally, timeROC analysis was conducted to evaluate the predictive accuracy of the *FAM83D* gene.

Single cell and spatial transcriptomics analysis

Single-cell and spatial transcriptomic data of HCC were obtained from GSE166635 and skrx2fz79n. Cells were filtered based on the following criteria: retaining genes expressed in a quantity of 500 to 10,000; removing genes expressed in fewer than 3 cells; retaining cells with gene expression levels between 1,000 and 100,000; and retaining cells with a mitochondrial gene percentage of less than 20%. Batch effects between different samples were eliminated using the "harmony" package. Clustering and dimensionality reduction analysis of cells was performed using the "Seurat" package. Receptor-ligand communication between different cell types was calculated using the "cellchat" package.

IHC validation analysis

In order to verify the accuracy of key diagnostic genes again, we collected tissue samples from HCC patients for IHC analysis. Controls were also collected as adjacent tissues (>3 cm from the HCC tissues edge). A total of 100 HCC samples were collected from January 2022 to December 2023 at the First Affiliated Hospital of Bengbu Medical University. All HCC tissue samples and corresponding paracancerous tissue samples were fixed in 4% formalin solution, paraffin-embedded, and consecutively sectioned at 4 μm intervals. Subsequent to sectioning, the samples underwent deparaffinization, were washed with xylene solution, and dehydrated using an ethanol gradient of varying concentrations. The study was conducted in

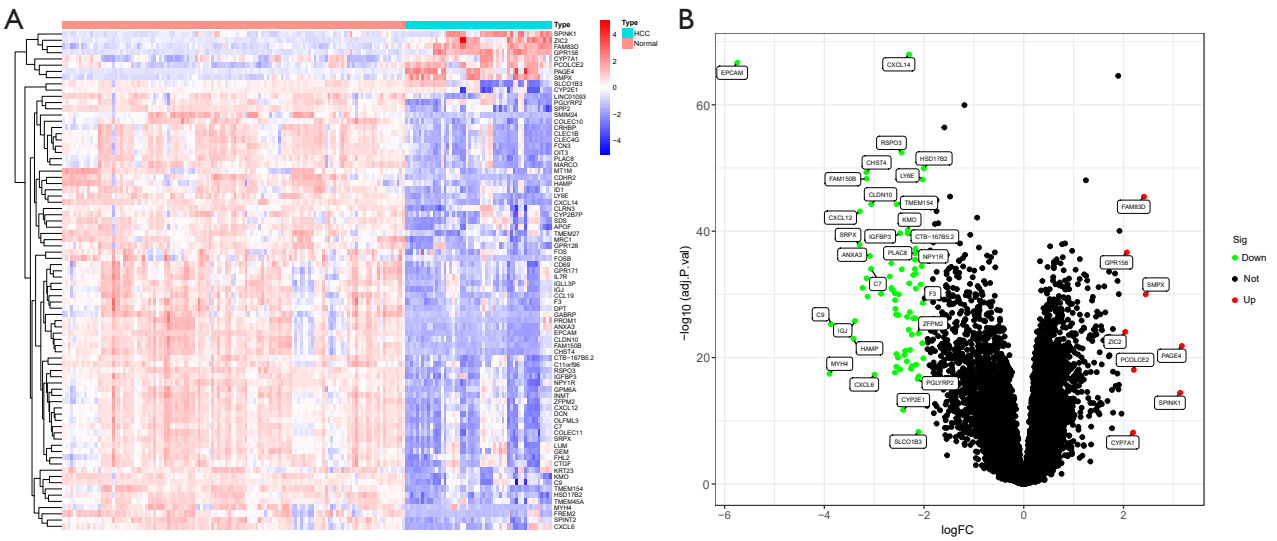

Figure 1 DEGs analysis of GSE78737 and GSE98383 datasets. (A) Heat map; (B) volcano map. DEGs, differentially expressed genes; FC, fold change; HCC, hepatocellular carcinoma.

accordance with the Declaration of Helsinki and its subsequent amendments. The study was approved by the Human Ethics Committee of Bengbu Medical University (No. 2024-156). Informed consent was not obtained due to the retrospective nature of the study.

FAM83D antibody (339-470AA, 1:200) was procured from Wuhan Huamei Bioengineering Co., Ltd. A known positive sample served as the positive control, whereas for the negative control, the samples were incubated in phosphate-buffered saline without the addition of the antibody. The proportion of positive cells and staining intensity was comprehensively scored. The percentage of stained cells among all counted cells was categorized as follows: 0 points for ≤5% positive cells; 1 point for 5–25% positive cells; 2 points for 26–50% positive cells; 3 points for 51–75% positive cells; and 4 points for >75% positive cells. Based on the appearance of the stained cells, the following scoring was used: 0 points for no yellow coloration, 1 point for light-yellow granules, 2 points for brown granules, 3 points for deep brown staining, and 4 points for very intense brown staining. For each section, 4 randomly selected fields at 200× magnification were examined, and each field was scored for the proportion of stained cells and staining intensity according to the following criteria: ≥6 points indicated high expression; <6 points indicated low expression.

Statistical analysis

In this study, differences between the two groups were compared using the *t*-test. Log-rank was used to test the survival differences between the two groups in K-M analysis. Cox regression analysis was employed to assess the impact of key genes on the prognosis of HCC patients, while ROC analysis evaluated the diagnostic accuracy of the genes. Additionally, Pearson correlation coefficient was used to analyze the correlation between key genes and immune cells. Statistical analyses were performed using R software (Version 4.3.2), with a two-sided P<0.05 considered significant.

Results

DEGs

A total of 80 DEGs were found, including 8 upregulated and 72 downregulated genes. Additionally, the volcano plot and heat map were made to visualize (*Figure 1*).

Bio-functional enrichment of DEGs

GO, KEGG, and GSEA analyses were performed using the clusterprofiler package based on the retrieved DEGs. The findings revealed that DEGs in BP were significant related to opsonization, humoral immune response, and complement activation. CC was strongly linked to collagen trimer, collagen-containing extracellular matrix, and serine-type peptidase complex. MF showed a significant association with chemokine activity, chemokine receptor binding, and

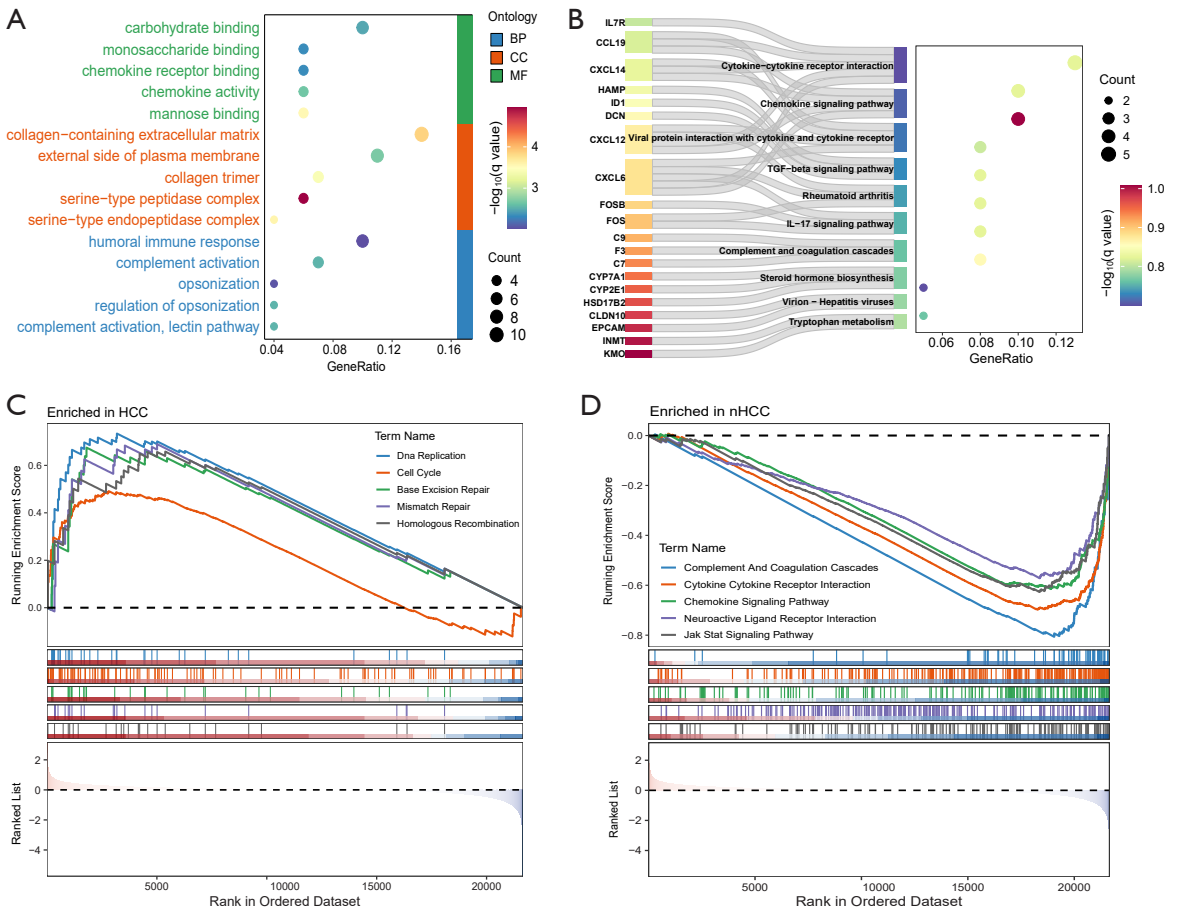

Figure 2 Enrichment analysis. (A) GO; (B) KEGG; (C,D) GSEA. BP, biological process; CC, cellular component; GO, Gene Ontology; GSEA, Gene Set Enrichment Analysis; HCC, hepatocellular carcinoma; KEGG, Kyoto Encyclopedia of Genes and Genomes; MF, molecular function; nHCC, non-HCC.

mannose-binding (*Figure 2A*). DEGs were predominantly focused on viral proteins' interaction with cytokines and cytokine receptors, TGF-β signaling pathway, and IL-17 signaling pathway in the KEGG pathway (*Figure 2B*). The GSEA enrichment analysis showed that complement and coagulation cascades, cytokine-cytokine receptor interaction, neuroactive ligand-receptor interaction, chemokine, and JAK-STAT signaling pathways, DNA replication, cell cycle, base excision repair, homologous recombination, mismatch repair were mainly active in nHCC and HCC (*Figure 2C,2D*).

Immune cell infiltration

Analysis of HCC and nHCC immune cell differential expression revealed significantly higher levels of naive B cells, $CD8^+$ T cells, T regulatory cells (Tregs), activated natural killer (NK) cells, M0 macrophages, resting dendritic cells, and activated dendritic cells. However, $CD4^+$ T cells naive, gamma delta T cells, and monocytes were significantly lower than in the nHCC (*Figure 3A*). The heat map was used to visualize immune cell correlations in tumor microenvironment. Red represents positive correlation and blue represents negative correlation (*Figure 3B*).

Machine learning

LASSO algorithm identified 13 genes, whereas the SVM-RFE method chose 16 genes (*Figure 4A,4B*). Seven genes, including *CXCL14*, *EPCAM*, *HSD17B2*, *LY6E*, *FAM83D*, *CDHR2*, and *SMIM24*, were obtained by taking their intersection (*Figure 4C*). P<0.05 indicates differences in

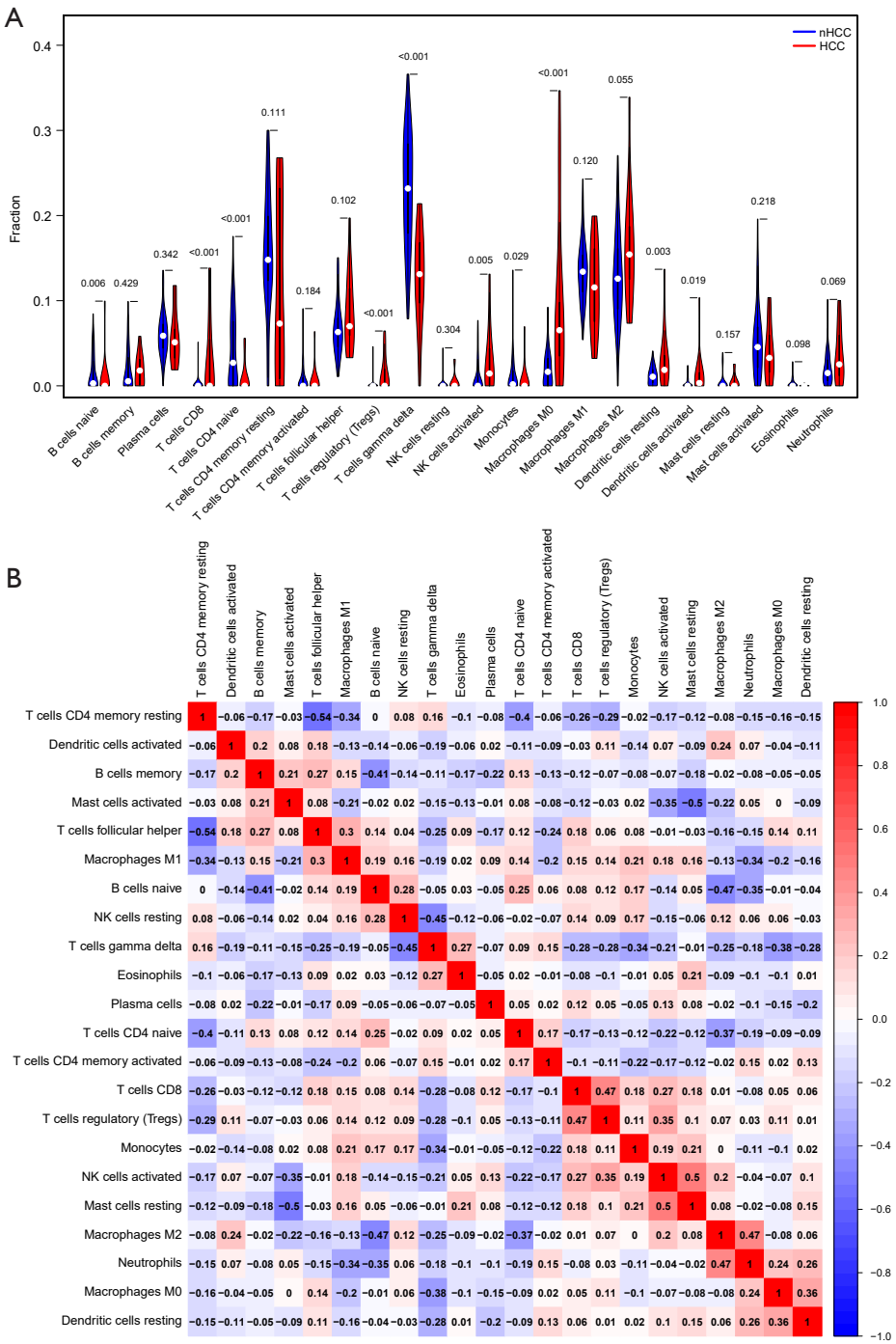

Figure 3 Immune cell infiltration analysis. (A) The violin plot shows differences in immune infiltration between nHCC and HCC; (B) the relationship between immune cells in the training set. HCC, hepatocellular carcinoma; nHCC, non-HCC.

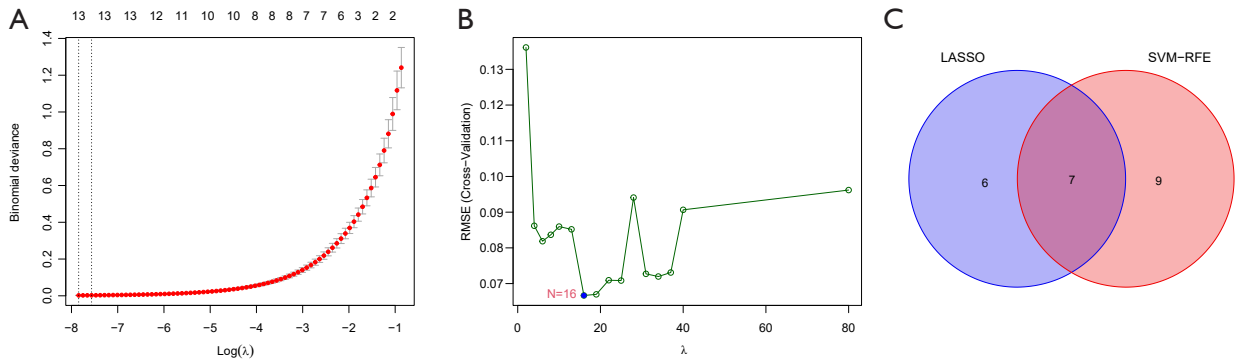

Figure 4 LASSO and SVM-RFE screening of HCC potential diagnostic genes in the training set. (A) LASSO; (B) SVM-RFE; (C) the two-algorithm intersection. HCC, hepatocellular carcinoma; LASSO, least absolute shrinkage and selection operator; RMSE, root mean square error; SVM-RFE, support vector machine-recursive feature elimination.

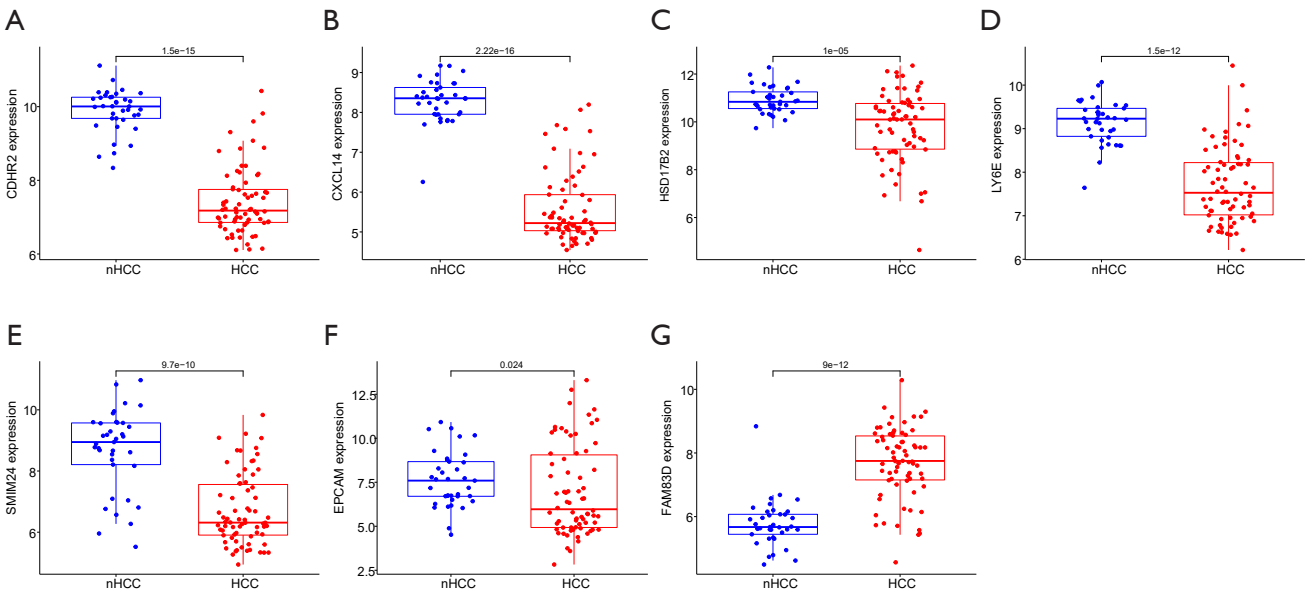

Figure 5 Diagnostic genes between HCC and nHCC in the training set (A-G). HCC, hepatocellular carcinoma; nHCC, non-HCC.

the seven diagnostic genes between nHCC and HCC. The HCC group is depicted in red, and the nHCC group in blue. In the training set, suspected HCC genes *CXCL14*, *HSD17B2*, *LY6E*, *CDHR2*, and *SMIM24* were putatively upregulated, while *EPCAM* and *FAM83D* were downregulated (*Figure 5A-5G*). The definitive diagnosis of HCC was assessed using ROC curves, and all seven gene area under the curve (AUC) was >0.95 (*Figure 6A-6G*). Moreover, seven genes (*CXCL14*, *HSD17B2*, *LY6E*, *CDHR2*, *SMIM24*, *EPCAM*, and *FAM83D*) had AUCs >0.9. In the validation set, the AUCs of *CXCL14*, *LY6E*, *CDHR2*, and *FAM83D* were greater than 0.9. However, the AUCs of *EPCAM*, *SMIM24*, and *HSD17B2* were less than 0.9 (*Figure 7A-7G*). Because of HCC's high lethality, we used the potential genes identified in the two sets. As a result, the *CXCL14*, *LY6E*, *CDHR2*, and *FAM83D* genes were all utilized as diagnostic genes.

Clinical and early-diagnostic prognostic

The four gene effects (*CXCL14*, *LY6E*, *CDHR2*, and *FAM83D*) on the OS and prognosis of HCC patients were

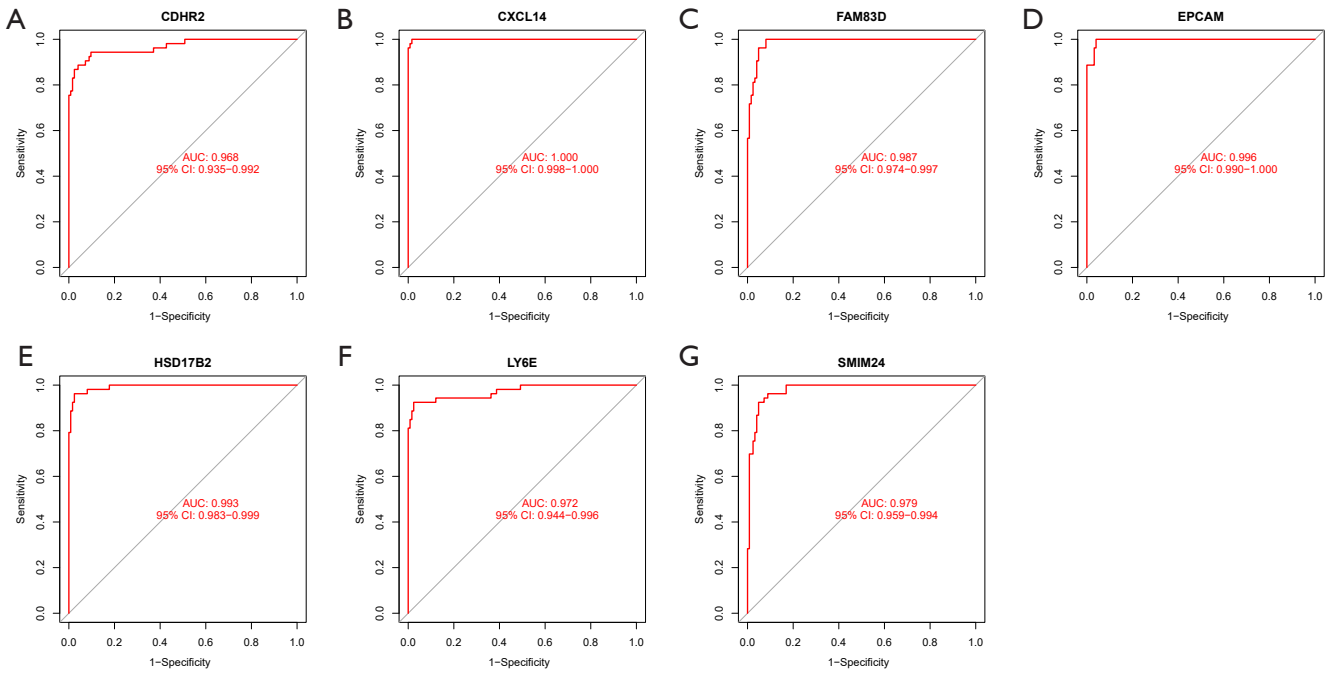

Figure 6 ROC curves of seven diagnostic genes for HCC in the training set (A-G). AUC, area under the curve; CI, confidence interval; HCC, hepatocellular carcinoma; ROC, receiver operating characteristic.

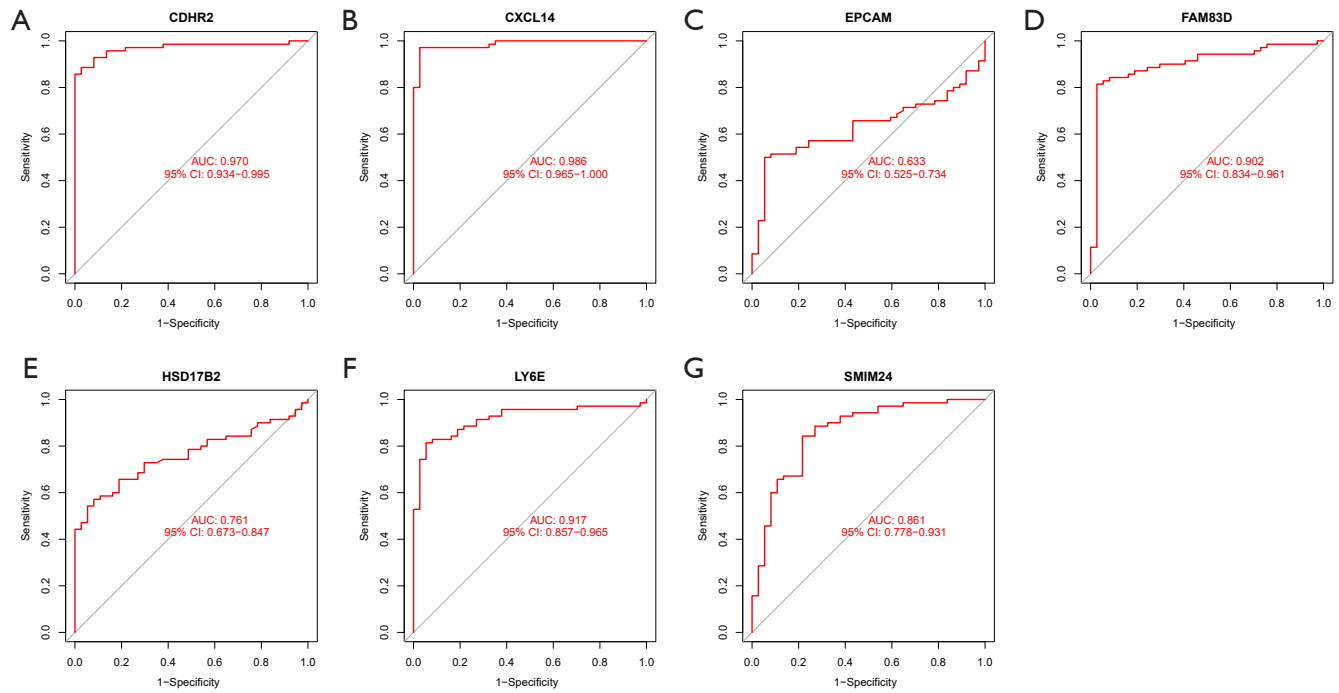

Figure 7 ROC curves of seven diagnostic genes for HCC in the validation set (A-G). AUC, area under the curve; CI, confidence interval; HCC, hepatocellular carcinoma; ROC, receiver operating characteristic.

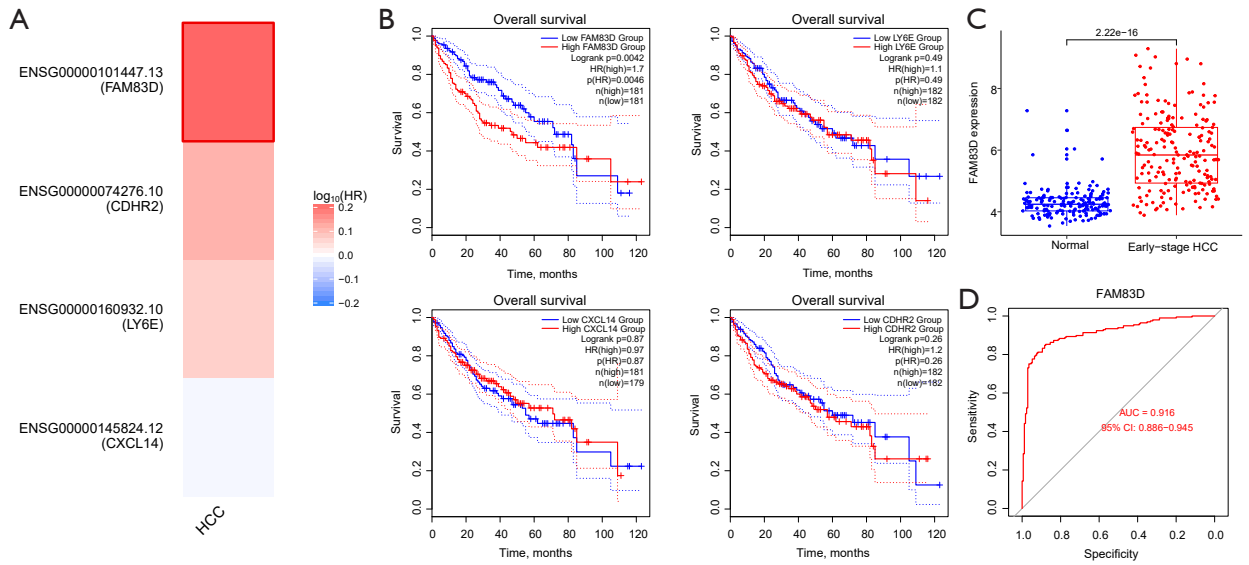

Figure 8 Prognostic and early-diagnostic analysis. (A) Survival map; (B) OS analysis of diagnostic genes; (C) differential expression of *FAM83D* between nHCC and early-stage HCC tissues; (D) performance of *FAM83D* in diagnosing early-stage HCC patients. AUC, area under the curve; CI, confidence interval; HCC, hepatocellular carcinoma; HR, hazard ratio; nHCC, non-HCC; OS, overall survival.

analyzed through the GEPIA 2.0 database, revealing that only *FAM83D* significantly affected HCC patient OS and prognosis (*Figure 8A,8B*). For early-stage HCC diagnosis, *FAM83D* expression was significantly up-regulated in early-stage HCC patients (P<0.05, *Figure 8C*). ROC-curve analysis showed an AUC of 0.916 (95% CI: 0.886–0.945) in discriminating early-stage HCC from nHCC patients, indicating a robust diagnostic performance (*Figure 8D*). Consequently, *FAM83D* is the key diagnostic gene finally determined.

Immune cell correlation of FAM83D

Subsequently, we looked at the connection between *FAM83D* and immune cell infiltration (*Figure 9A*). The analysis indicated that *FAM83D* and immune cells were significantly correlated and that the correlation between them was related to the size of circle representing the score and its color defining the test P value. Correlations with P<0.05 indicate significant relationships between immune cells and the target gene (shown in red). *FAM83D* had a significant positive correlation with immunosuppressive cells, such as Tregs (R=0.34, P<0.001), M2 macrophages (R=0.28, P=0.002), resting dendritic cells (R=0.27, P=0.003), resting mast cells (R=0.24, P=0.01), and M0 macrophages (R=0.20, P=0.03), and a significant negative relation to immune-stimulating cells, like gamma delta T cells (R=–0.20, P=0.02), activated mast cells (R=–0.21, P=0.02), plasma cells (R=–0.22, P=0.01), B cells naive (R=–0.23, P=0.01), and $CD4^+$ T cells naive (R=–0.37, P<0.001) (*Figure 9B-9M*).

Risk and prognosis of FAM83D

By analyzing the expression data of *FAM83D* and the corresponding clinical information of HCC patients from TCGA database, it was found that higher expression of *FAM83D* is associated with an increased risk of HCC, higher mortality, and poorer prognosis (*Figure 10*).

Furthermore, *FAM83D* was identified as an independent prognostic factor affecting the outcomes of HCC patients (*Figure 11A*). Risk nomograms and calibration curves demonstrate that patients with HCC have a worse prognosis when *FAM83D* expression is raised (*Figure 11B,11C*). The 1- and 3-year AUC values of patients with HCC diagnosed with *FAM83D* were 0.731 (95% CI: 0.674–0.788) and 0.65 (95% CI: 0.585–0.716), respectively (*Figure 11D*).

Verification of FAM83D

By analyzing the expression data of *FAM83D* and the corresponding clinical information of HCC patients from ICGC database, it was found that higher expression of *FAM83D* is associated with an increased risk of HCC,

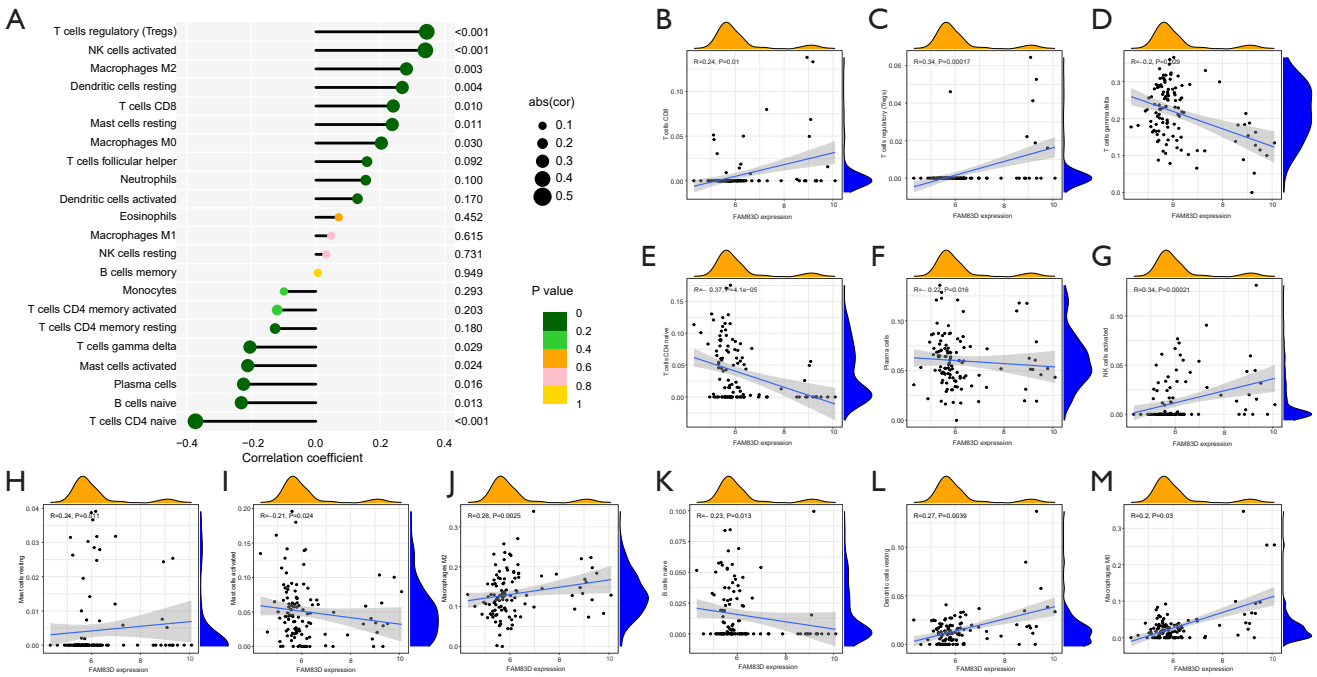

Figure 9 Immune correlation analysis of *FAM83D*. (A) *FAM83D* and lollipop diagram of immune cells; (B-M) correlation between *FAM83D* and immune cells.

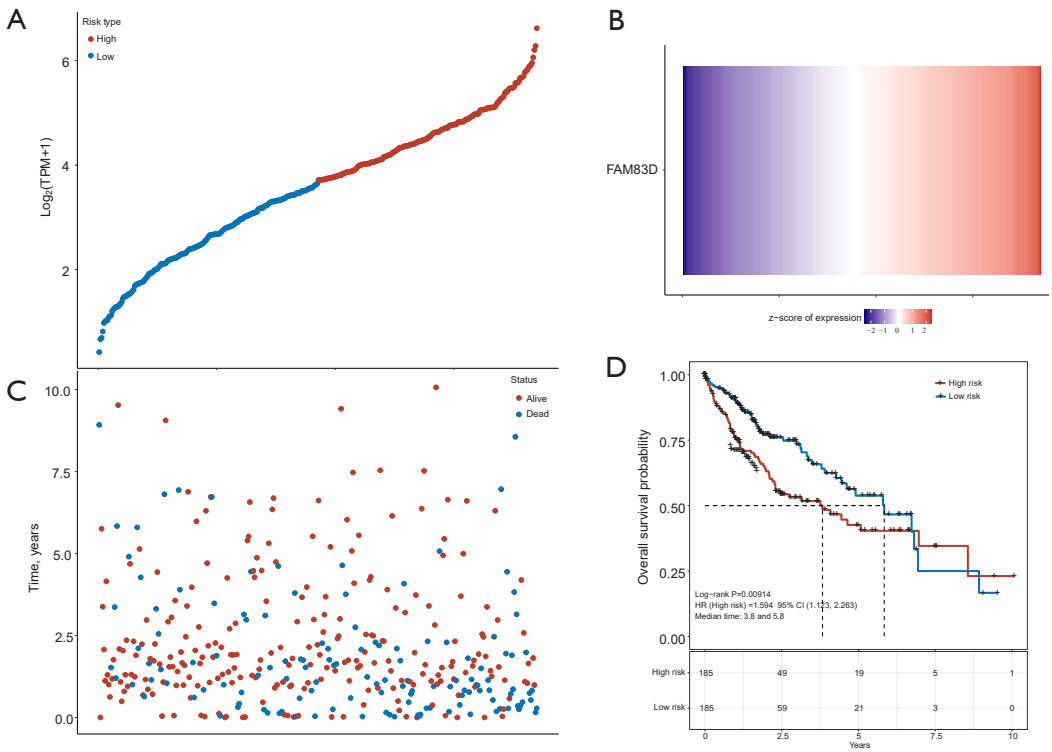

Figure 10 Risk model of *FAM83D*. (A) Correlation between *FAM83D* expression and survival time; (B) heat map of *FAM83D* expression; (C) correlation between *FAM83D* expression and survival status; (D) the Kaplan-Meier survival curve. CI, confidence interval; HR, hazard ratio; TPM, transcripts per million.

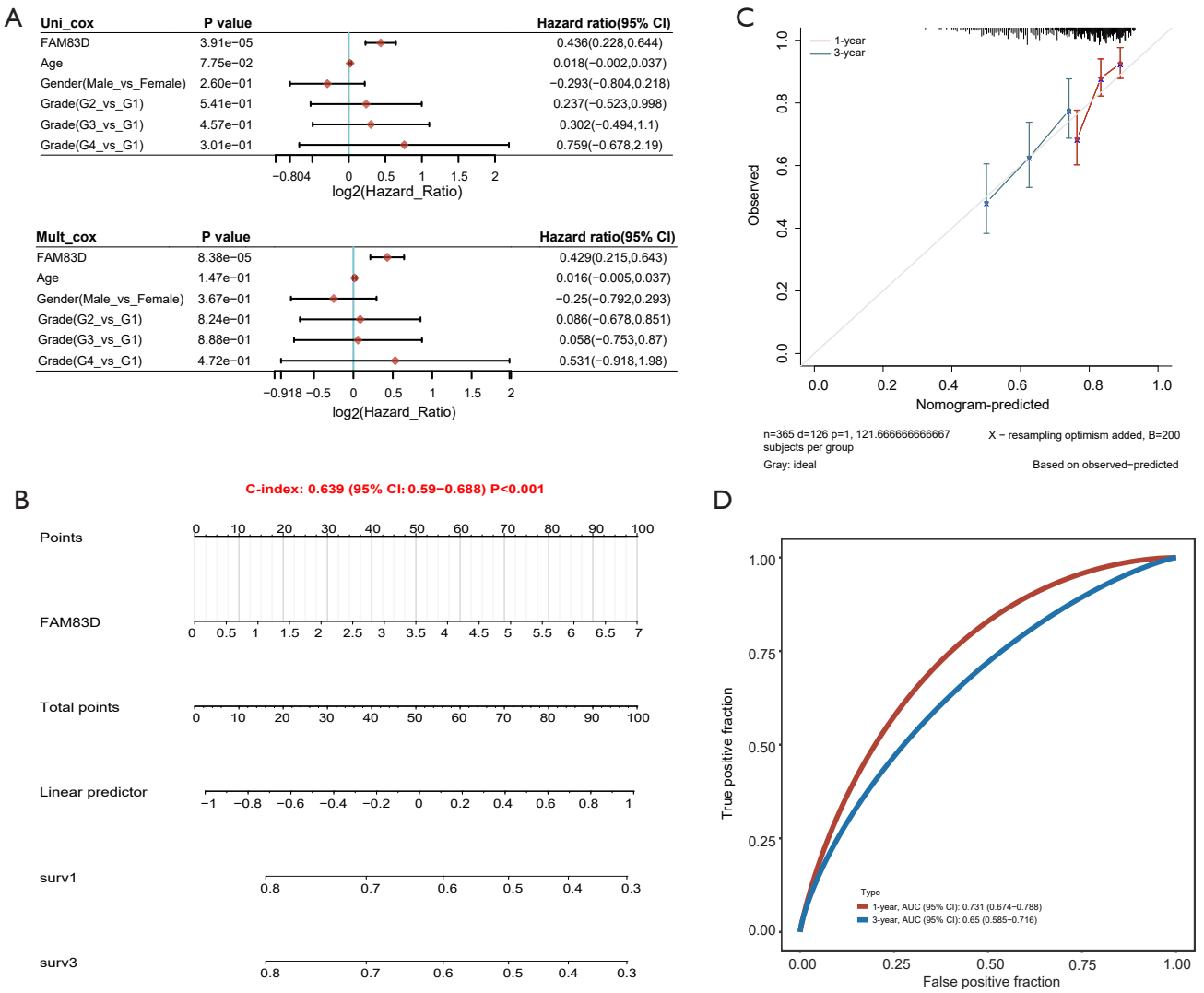

Figure 11 Prognostic model of *FAM83D*. (A) Cox regression analysis; (B) nomogram; (C) risk nomogram calibration curve; (D) ROC. AUC, area under the curve; CI, confidence interval; ROC, receiver operating characteristic.

higher mortality, and poorer prognosis (*Figure 12A,12B*). The 1-year and 3-year AUC values of patients with HCC diagnosed with *FAM83D* were 0.709 (95% CI: 0.628–0.789) and 0.693 (95% CI: 0.605–0.780), respectively (*Figure 12C*).

Expression of FAM83D

The expression of *FAM83D* is localized in the cytoplasm, with a high expression rate of 92.0% in 100 HCC tissues samples, whereas it is minimally expressed in adjacent tissues (*Figure 13A*), showing a significant difference.

We classified 10 different cell types (*Figure 13B*). Among them, *FAM83D* is mainly expressed in malignant cells, followed by hepatocytes (*Figure 13C,13D*). Further analysis of marker genes (*Figure 13E*) for each cell type and intercellular communication revealed that malignant cells have significantly stronger communication with cancer-associated fibroblasts (CAFs) and endothelial cells compared to other cells (*Figure 13F,13G*). The intensity of receptor-ligand communication is mainly concentrated on the laminin and MK pathways (*Figure 13H*). Therefore, *FAM83D* may promote tumor growth by enhancing these cellular communication processes.

Additionally, we further explored the potential relationship between *FAM83D* gene expression and immunotherapy in HCC using spatial transcriptome data.

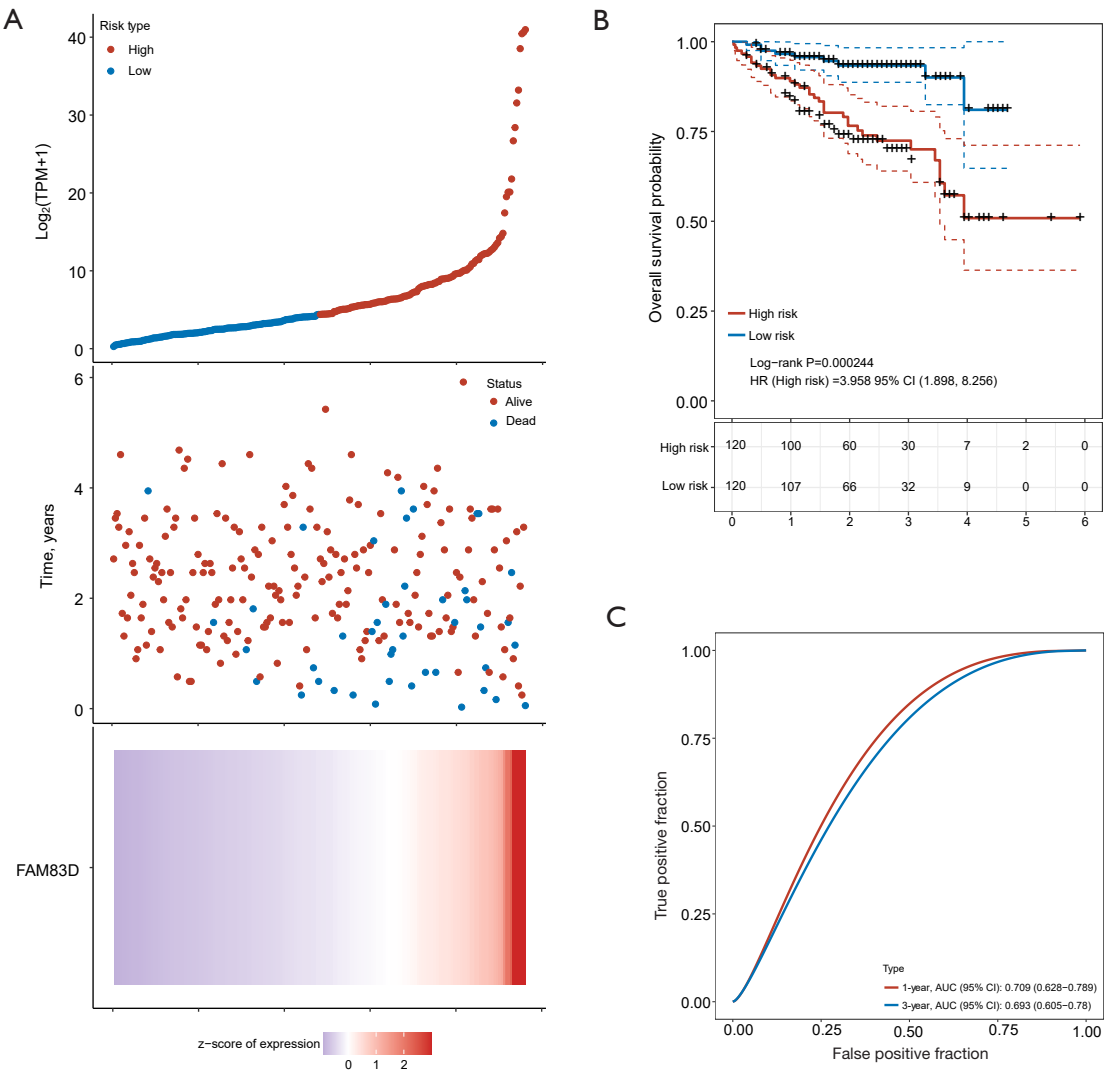

Figure 12 Verification of *FAM83D*. (A) Expression thermogram of *FAM83D* and its correlation with survival relationship and survival state; (B) the Kaplan-Meier survival curve; (C) ROC. AUC, area under the curve; CI, confidence interval; HR, hazard ratio; ROC, receiver operating characteristic.

The results indicated that *FAM83D* is highly expressed in malignant cells in both responder and non-responder to immunotherapy, with a significant increase in expression in non-responder (*Figure 13I*). This suggests that the expression of *FAM83D* may be closely associated with the nonresponse to immunotherapy in HCC patients.

Discussion

HCC is the most common primary liver malignancy and represents a major global health burden, being the third leading cause of cancer-related deaths worldwide (14). Despite advancements in surgical techniques and therapeutic interventions, the prognosis for HCC patients remains dismal, largely due to late-stage diagnosis and high recurrence rates (15,16). The molecular heterogeneity of HCC further complicates treatment strategies, necessitating the identification of reliable prognostic biomarkers and therapeutic targets. In this context, our study leveraged machine learning and bioinformatics approaches to systematically identify key prognostic genes associated with HCC, aiming to enhance the predictive accuracy of

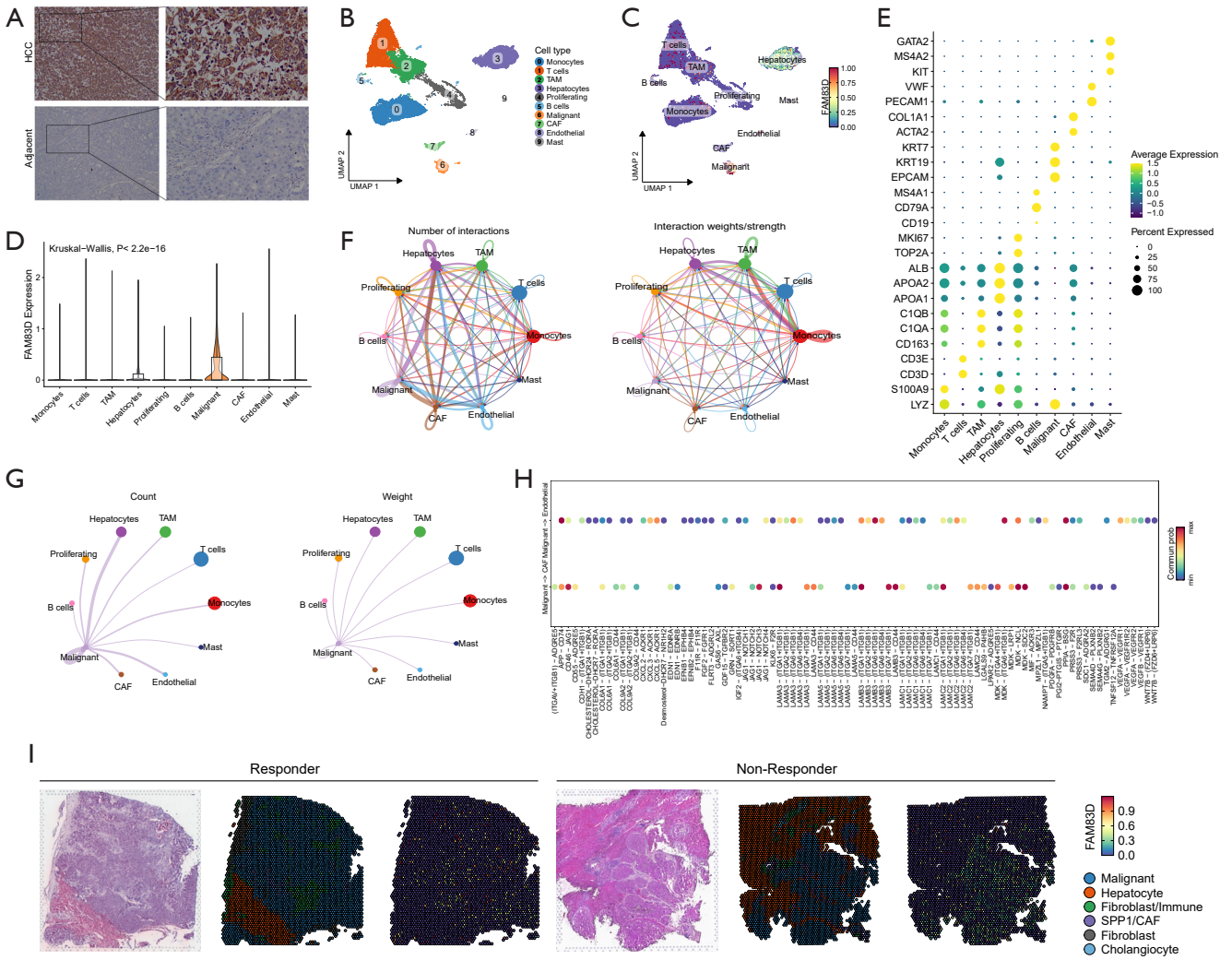

Figure 13 Expression of *FAM83D*. (A) The protein expression of *FAM83D* (immunohistochemical staining, left 200×, right 400×); (B) cell types distribution; (C,D) the expression of *FAM83D* in each cell type; (E) the expression of marker genes in each cell type; (F-G) the chord diagram of communication strength among cell types; (H) the communication probability of various receptor-ligand interactions; (I) the expression of *FAM83D* in immunotherapy responder and non-responder through spatial transcriptomics (left: H&E staining, 100×).

clinical outcomes and provide a foundation for personalized treatment strategies.

In this study, we utilized three GEO datasets and employed bioinformatics and machine learning methods to screen and identify the key gene *FAM83D* involved in the development and progression of HCC. To validate the accuracy of the identified key gene, we conducted validation analyses using the ICGC and TCGA datasets. Additionally, we performed IHC experiments on tissue samples collected from HCC patients to further confirm the accuracy of the key gene. Therefore, *FAM83D* may serve as a potential biomarker for the diagnosis and treatment of HCC in the future.

The identification of *FAM83D* as a key diagnostic and prognostic biomarker in HCC presents significant implications for understanding the molecular mechanisms of this malignancy and for developing potential therapeutic strategies. *FAM83D*, also known as CHICA, belongs to the FAM83 family and plays a crucial role in multiple tumor cell signaling pathways (17). This protein, encoded by the human chromosomal region 20q, is involved in spindle regulation and the maintenance of cell division (18). Consequently, *FAM83D* is associated with key regulators of mitosis and cytoplasmic division in tumors. *FAM83D* directs the protein kinase CK1α to

the mitotic spindle, which is essential for proper spindle localization (19). It is significantly upregulated in various cancers, including breast, ovarian, pancreatic, non-small cell lung cancers, and glioma, impacting the OS of patients (20-23). *FAM83D* can inhibit the FBXW7/MCL1 signaling pathway, induce the proliferation and migration of tumor cells, and inhibit apoptosis, thus promoting the progression of HCC (24). Moreover, *FAM83D* expression is linked to MCF-7 cell proliferation, suggesting that *FAM83D* could be a prospective prognostic biomarker across cancer types (25). Then, *FAM83D* is post-transcriptionally regulated by PRMT1, enhancing its interaction with other proteins (26). Additionally, the study found that miRNA-495 targeted *FAM83D* to inhibit the proliferation and migration of colorectal cancer cells (27). *FAM83D* influences HCC migration and invasion through the Notch1 and MEK/ERK signaling pathways (28,29). These findings suggest that *FAM83D* is involved in HCC development and is associated with patient OS. Overexpression of *FAM83D* in HCC leads to reduced OS in HCC patients and is related to several immune cells in the HCC microenvironment. Therefore, *FAM83D* may serve as a valuable prognostic biomarker for HCC.

A study has demonstrated that tumorigenesis, invasion, and metastasis are significantly influenced by the tumor microenvironment, which plays a crucial role in determining patient survival and prognosis (30). The tumor microenvironment is a complex ecosystem composed of tumor cells, immune cells, and non-immune cells, with intricate interactions among these cell types (31). Although immune cell infiltration in HCC has been previously reported, no study has employed machine learning algorithms and big data to analyze the relationship between prognostic diagnostic genes and immune cells in HCC. In this study, we utilized these tools, and our correlation findings may enhance the sensitivity of HCC diagnosis and contribute to understanding the impact of specific genes on immunotherapy outcomes in HCC patients.

The immune cell infiltration analysis provided insights into the tumor microenvironment of HCC. Our analysis revealed significant overexpression of $CD8^+$ T cells, Tregs, activated NK cells, M0 macrophages, and resting dendritic cells in the HCC microenvironment. Conversely, naive $CD4^+$ T cells and gamma delta T cells were significantly underexpressed. Therefore, the final key diagnostic gene screened, *FAM83D*, may play an immune-related role in causing and treating HCC. Ma *et al.* (32) concluded that *FAM83D* had a significant prognostic value in pancreatic ductal adenocarcinoma (PDAC), which may be crucial in controlling tumor progression and immune cell infiltration after discovering through bioinformatics analysis that *FAM83D* was related to $CD8^+$ T cells, gamma delta T cells, and $CD4^+$ T cell infiltration in PDAC. Moreover, *FAM83D* is significantly overexpressed in gastric adenocarcinoma (STAD). It has a significant correlation to tumor stage in STAD patients and with B cells, $CD8^+$ T cells, $CD4^+$ T cells, macrophages, and dendritic cell infiltration (33). However, due to the tumor microenvironment complexity, the individual immune cell function in tumorigenesis, invasion, and metastasis cannot be explored in isolation and requires additional analysis of their interactions.

Although we employed bioinformatics and machine learning algorithms to identify the diagnostically valuable key gene *FAM83D* in HCC and conducted extensive internal and external validations, our study has certain limitations at the genomic level. Utilizing multi-omics data or immune-related non-coding RNA signals could provide a more comprehensive understanding of the pathogenic mechanisms of HCC and improve survival rate predictions. In addition, we lack further functional experiments to prove how *FAM83D* affects the occurrence and development of HCC.

Conclusions

In summary, we successfully identified the key gene *FAM83D* in HCC using bioinformatics and machine learning algorithms, and conducted multiple validation analyses to demonstrate that *FAM83D* could serve as a critical biomarker for HCC. Furthermore, we discovered that the expression of *FAM83D* is closely associated with the infiltration of immune cells within the HCC microenvironment. Therefore, our findings may also have significant implications for advancing immunotherapy in HCC. These findings contribute to the growing body of knowledge on HCC and offer promising avenues for future research and clinical applications. Further exploration of *FAM83D* and its role in HCC could lead to new insights into the molecular mechanisms of this malignancy and the development of novel therapeutic strategies.

Acknowledgments

The authors appreciate the GEO, TCGA, ICGC databases for data support.

Footnote

Reporting Checklist: The authors have completed the TRIPOD and MDAR reporting checklists. Available at https://tcr.amegroups.com/article/view/10.21037/tcr-2025-1067/rc

Data Sharing Statement: Available at https://tcr.amegroups.com/article/view/10.21037/tcr-2025-1067/dss

Peer Review File: Available at https://tcr.amegroups.com/article/view/10.21037/tcr-2025-1067/prf

Funding: This work was supported by the Key Project of Natural Science Research of Education Department of Anhui Province (No. 2023AH051929), General Project of Anhui Provincial Health Commission (No. 2024Aa20053), and the Natural Science Major Program for Universities in Anhui Province (No. 2024AH040192), and the Key Laboratory Open Project of Anhui Province (No. AHCM2023Z002).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tcr.amegroups.com/article/view/10.21037/tcr-2025-1067/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 and its subsequent amendments. The study was approved by the Human Ethics Committee of Bengbu Medical University (No. 2024-156). Informed consent was not obtained due to the retrospective nature of the study.

References

1. Arnett A, Siegel DA, Dai S, et al. Incidence and survival of pediatric and adult hepatocellular carcinoma, United States, 2001-2020. Cancer Epidemiol 2024;92:102610.
2. Xie L, Shu Y, Ye M, et al. Identification of MTCH1 as a novel prognostic indicator and therapeutic target in hepatocellular carcinoma. Pathol Res Pract 2024;259:155358.
3. Regnault H, Chalaye J, Galetto-Pregliasco A, et al. Selective internal radiation therapy for unresectable HCC: The SIRT downstaging study. Hepatol Commun 2024;8:e0475.
4. Trevisani F, Vitale A, Kudo M, et al. Merits and boundaries of the BCLC staging and treatment algorithm: Learning from the past to improve the future with a novel proposal. J Hepatol 2024;80:661-9.
5. Panigrahi G, Ambs S. How Comorbidities Shape Cancer Biology and Survival. Trends Cancer 2021;7:488-95.
6. Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N Engl J Med 2023;388:1201-8.
7. Salas M, Petracek J, Yalamanchili P, et al. The Use of Artificial Intelligence in Pharmacovigilance: A Systematic Review of the Literature. Pharmaceut Med 2022;36:295-306.
8. Ma J, An S, Cao M, et al. Integrated machine learning and deep learning for predicting diabetic nephropathy model construction, validation, and interpretability. Endocrine 2024;85:615-25.
9. Cresta Morgado P, Carusso M, Alonso Alemany L, et al. Practical foundations of machine learning for addiction research. Part I. Methods and techniques. Am J Drug Alcohol Abuse 2022;48:260-71.
10. Painuli D, Bhardwaj S, Köse U. Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review. Comput Biol Med 2022;146:105580.
11. Koppad S, Basava A, Nash K, et al. Machine Learning-Based Identification of Colon Cancer Candidate Diagnostics Genes. Biology (Basel) 2022;11:365.
12. Najm M, Azencott CA, Playe B, et al. Drug Target Identification with Machine Learning: How to Choose Negative Examples. Int J Mol Sci 2021;22:5118.
13. Zuo D, Xiao J, An H, et al. Screening for Lipid-Metabolism-Related Genes and Identifying the Diagnostic Potential of ANGPTL6 for HBV-Related Early-Stage

Hepatocellular Carcinoma. Biomolecules 2022;12:1700.
14. 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.
15. Kulik L, El-Serag HB. Epidemiology and Management of Hepatocellular Carcinoma. Gastroenterology 2019;156:477-491.e1.
16. Javan H, Dayyani F, Abi-Jaoudeh N. Therapy in Advanced Hepatocellular Carcinoma. Semin Intervent Radiol 2020;37:466-74.
17. Liu ZM, Yuan Y, Jin L. FAM83D acts as an oncogene by regulating cell cycle progression via multiple pathways in synovial sarcoma: a potential novel downstream target oncogene of anlotinib. Discov Oncol 2024;15:82.
18. Geng Y, Liu J, Wang Z, et al. Systematic analysis of the oncogenic role of FAM83D across cancers based on data mining. Cell Cycle 2023;22:1005-19.
19. Fulcher LJ, He Z, Mei L, et al. FAM83D directs protein kinase CK1α to the mitotic spindle for proper spindle positioning. EMBO Rep 2019;20:e47495.
20. Zhang Q, Yu S, Lok SIS, et al. FAM83D promotes ovarian cancer progression and its potential application in diagnosis of invasive ovarian cancer. J Cell Mol Med 2019;23:4569-81.
21. Hua YQ, Zhang K, Sheng J, et al. Fam83D promotes tumorigenesis and gemcitabine resistance of pancreatic adenocarcinoma through the Wnt/β-catenin pathway. Life Sci 2021;287:119205.
22. Yin C, Lin X, Wang Y, et al. FAM83D promotes epithelial-mesenchymal transition, invasion and cisplatin resistance through regulating the AKT/mTOR pathway in non-small-cell lung cancer. Cell Oncol (Dordr) 2020;43:395-407.
23. Wang J, Quan Y, Lv J, et al. Inhibition of FAM83D displays antitumor effects in glioblastoma via down-regulation of the AKT/Wnt/β-catenin pathway. Environ Toxicol 2022;37:1343-56.
24. Nie J, Lu L, Du C, et al. FAM83D promotes the proliferation and migration of hepatocellular carcinoma cells by inhibiting the FBXW7/MCL1 pathway. Transl Cancer Res 2022;11:3790-802.
25. Yu H, Chen Q, Wang Z, et al. Pan-cancer and single-cell analysis reveals FAM83D expression as a cancer prognostic biomarker. Front Genet 2022;13:1009325.
26. Snijders AM, Lee SY, Hang B, et al. FAM83 family oncogenes are broadly involved in human cancers: an integrative multi-omics approach. Mol Oncol 2017;11:167-79.
27. Yan L, Yao J, Qiu J. miRNA-495 suppresses proliferation and migration of colorectal cancer cells by targeting FAM83D. Biomed Pharmacother 2017;96:974-81.
28. Mu Y, Zou H, Chen B, et al. FAM83D knockdown regulates proliferation, migration and invasion of colorectal cancer through inhibiting FBXW7/Notch-1 signalling pathway. Biomed Pharmacother 2017;90:548-54.
29. Wang D, Han S, Peng R, et al. FAM83D activates the MEK/ERK signaling pathway and promotes cell proliferation in hepatocellular carcinoma. Biochem Biophys Res Commun 2015;458:313-20.
30. Xue R, Zhang Q, Cao Q, et al. Liver tumour immune microenvironment subtypes and neutrophil heterogeneity. Nature 2022;612:141-7.
31. Petroni G, Buqué A, Coussens LM, et al. Targeting oncogene and non-oncogene addiction to inflame the tumour microenvironment. Nat Rev Drug Discov 2022;21:440-62.
32. Ma Z, Zhou Z, Zhuang H, et al. Identification of Prognostic and Therapeutic Biomarkers among FAM83 Family Members for Pancreatic Ductal Adenocarcinoma. Dis Markers 2021;2021:6682697.
33. Zhang T, Lai S, Cai Y, et al. Comprehensive Analysis and Identification of Prognostic Biomarkers and Therapeutic Targets Among FAM83 Family Members for Gastric Cancer. Front Cell Dev Biol 2021;9:719613.

Cite this article as: Lu J, Ma J, Yu C, Lu S, Zhao X, Zhang L. Identification of the key gene for hepatocellular carcinoma based on bioinformatics and machine learning and experimental verification. Transl Cancer Res 2025;14(11):7995-8011. doi: 10.21037/tcr-2025-1067