Journal of Hepatocellular Carcinoma

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Identification of FDFT1 and PGRMC1 as New Biomarkers in Nonalcoholic Steatohepatitis (NASH)-Related Hepatocellular Carcinoma by Deep Learning

Qiqi Liu[1], Yinuo Yang[1], Yongshuai Wang[2], Shuhang Wei[1], Liu Yang[1], Tiantian Liu[1], Zhen Yu[3], Yuemin Feng[3], Ping Yao[4], Qiang Zhu[1,5]

[1]Department of Gastroenterology, Shandong Provincial Hospital, Cheeloo College of Medicine, Shandong University, Jinan, Shandong, 250021, People's Republic of China; [2]School of Computer Science and Technology, Shandong University, Qingdao, Shandong, 266237, People's Republic of China; [3]Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, 250021, People's Republic of China; [4]Department of Gastroenterology, the First Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang, 830000, People's Republic of China; [5]Department of Infectious Disease, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, 250021, People's Republic of China

Correspondence: Qiang Zhu, Department of Gastroenterology, Shandong Provincial Hospital, Cheeloo College of Medicine, Shandong University; Department of Infectious Disease, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, 250021, People's Republic of China, Email zhuqiang@sdu.edu.cn; Ping Yao, Department of Gastroenterology, the First Affiliated Hospital of Xinjiang Medical University, No. 137, Liyushan Road, Urumqi, Xinjiang, 830000, People's Republic of China, Email pingyaozh@xjmu.edu.cn

Background: With the global epidemic of obesity and diabetes, non-alcoholic fatty liver disease (NAFLD) is becoming the most common chronic liver disease, and NASH is increasingly becoming a major risk factor for hepatocellular carcinoma. Therefore, it is essential to explore novel biomarkers in NASH-related HCC.

Methods: Deep Learning (DL) methods are a promising and encouraging tool widely used in genomics by automatically applying neural networks (NNs). Therefore, DL, "limma package", weighted gene co-expression network analysis (WGCNA), and Protein–Protein Interaction Networks (PPI) were used to screen feature genes. Real-time quantitative PCR was used to validate the expression of feature genes in the NAFLD mice model. Enrichment and single-cell sequencing analyses of single genes were performed to investigate the role of feature genes in NASH-related HCC.

Results: Combined core genes screened by DL in NAFLD with important genes in metabolic syndrome, six feature genes (FDFT1, TNFSF10, DNAJC16, RDH11, PGRMC1, and MYC) were obtained. ROC analysis demonstrates the model's superiority with the AUC was 0.983 (0.9241–0.98885). Animal experiments based on NAFLD mouse models have also shown that FDFT1, TNFSF10, DNAJC16, RDH11, and PGRMC1 have a higher expression in NAFLD livers. Among the feature genes, FDFT1 and PGRMC1 showed significant expression trends and outstanding diagnosis value in NASH-HCC.

Conclusion: In conclusion, FDFT1 and PGRMC1 are key enzymes in the cholesterol synthesis pathway, our study validates the important role of cholesterol metabolism in NAFLD from another perspective, implying they may be new prognostic and diagnostic markers for NASH-HCC.

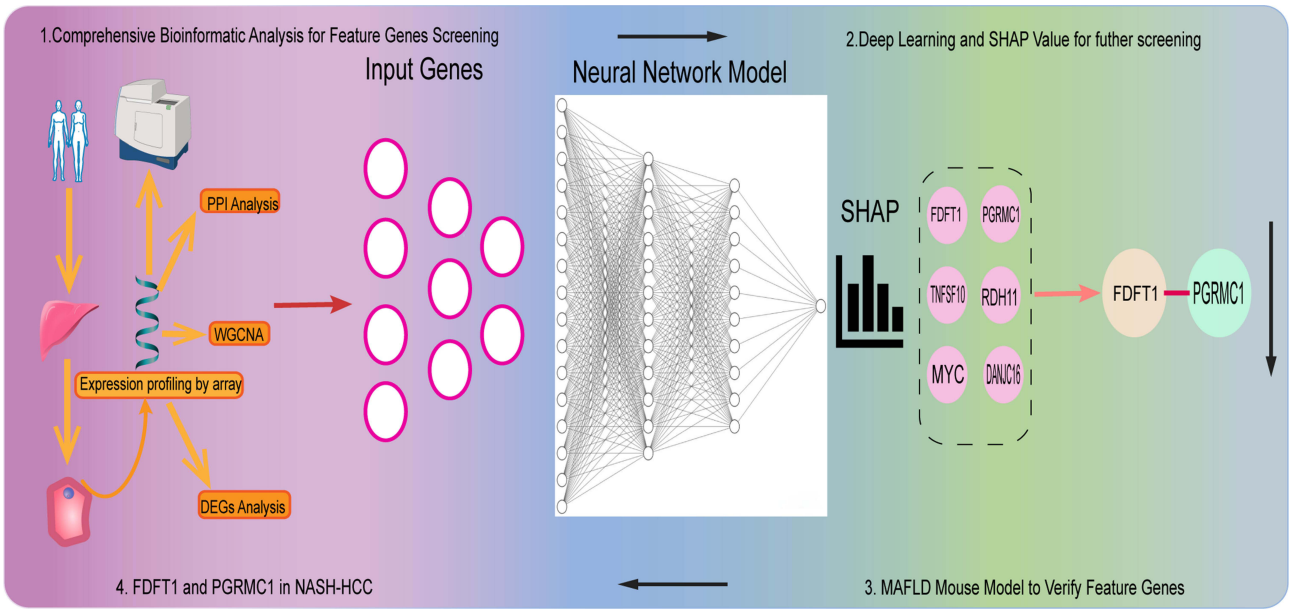
Keywords: NAFLD, NASH-HCC, deep learning, biomarkers, metabolic syndrome, cholesterol metabolism

Introduction

Non-alcoholic fatty liver disease (NAFLD) refers to chronic liver disease (CLD) spectrum containing non-alcoholic fatty liver (NAFL), non-alcoholic steatohepatitis (NASH), and NASH-related cirrhosis or hepatocellular carcinoma (HCC) in sequence.[1] The epidemiology and demographic characteristics of NAFLD vary in different countries and regions, usually, simultaneously with the prevalence of obesity,[2] a result showed that the prevalence of NAFLD in the overweight population was 69.99%.[3] Additionally, the prevalence of NAFLD is about 25%, and the number is still rapidly growing

Received: 27 November 2024
Accepted: 27 March 2025
Published: 5 April 2025

Graphical Abstract



with the improvement in the quality of human life globally,[4] and approximately 2% of NASH patients progress to NASH-related HCC each year.[5]

The pathogenesis of NAFLD is indefinite, while the multiple-hit hypothesis is currently the most recognized, and such hits consist of insulin resistance, hormones secreted from the adipose tissue, nutritional factors, gut microbiota, genetic factors, and so on.[6] Among these, NASH is a progressive liver disease defined as the presence of ≥5% hepatic steatosis and concomitant inflammation with hepatocellular injury, such as ballooning degeneration, and inflammatory response, with or without fibrosis.[2] Lipotoxicity, fibrosis, reactivation of molecular pathways, and altered immune microenvironment are relevant to hepatocellular carcinoma induced by NASH.[5] Currently, hepatocellular carcinoma (HCC) has become the sixth most common cancer and the third cause of cancer-related deaths around the world. At the same time, new epidemiological data show a shift in hepatocellular carcinoma risk factors from viral-associated liver disease to non-viral liver disease, such as alcohol-associated and metabolic dysfunction-associated steatotic liver disease, which has implications for both preventive strategies and therapeutic options for hepatocellular carcinoma.[7] Therefore, it is necessary to explore the latent specific diagnostic and therapeutic targets for NAFLD, especially NASH-HCC.

As for diagnosis, liver biopsy as the gold standard for the diagnosis of NAFLD has been more and more widely used in clinical diagnosis and research but is limited by the occasionality of puncture and pathologist expertise.[8] Magnetic Resonance Proton Density Fat Fraction (MRI-PDFF) is responsible for quantifying liver fat content and calculating the degree of hepatic steatosis. It is an emerging and powerful magnetic resonance testing technique in recent years, which has been taken as the "gold standard" for non-invasive liver fat quantification in clinical applications.[9] However, this measurement is also not effectively used in clinical research due to limitations such as high cost and the prevalence of this technique in magnetic resonance equipment. Accordingly, for non-invasive modalities, the diagnosis of NAFLD mainly relies on computed tomography (CT) and ultrasound, however, neither is sensitive to detecting <30% steatosis. As for treatment in NAFLD, apart from the recently approved THR-β agonist, Rezdiffra, there are no other effective drugs approved. Drugs currently in development primarily target on regulation of lipid metabolism, alleviation of oxidative stress, alleviation of ER stress, and alleviation of inflammation.[1] It is critical and demanding to explore diagnostic models applied to the NAFLD disease spectrum from another perspective of feature gene expression, which is

also beneficial for new drug development and interpretation of pathogenesis in NAFLD. Therefore, deep learning and bioinformatics analysis are combined innovatively to identify effective biomarkers in NAFLD, including NASH-HCC.

Bioinformatics analysis has been widely used in new biomarker identification of NAFLD, while deep learning as a freshly emerging and powerful tool is not yet widely applied. Deep learning refers to the application of various machine learning algorithms on multi-layer neural networks to learn the intrinsic patterns and expression levels of sample data. A neural network was constructed and validated for predicting the occurrence of disease, the input layer was the expression of feature genes and the output layer was the probability of disease occurrence. Afterward, the SHapley Additive exPlanations (SHAP) value was applied to explain the diagnostic model constructed by deep learning, which can perform local and global interpretability simultaneously. The biggest advantage of SHAP is that SHAP value can reflect the influence of the features in each sample compared with other methods.[10]

In this study, deep learning and bioinformatic analysis are first combined to explore effective biological markers in the whole NAFLD disease spectrum. Eventually, FDFT1 and PGRMC1, the two genes in the cholesterol synthesis pathway were recognized, and a NAFLD prediction model was constructed based on multiple gene expressions.

FDFT1, a key enzyme in cholesterol biosynthesis (catalyzing squalene synthesis), is implicated in NASH-associated metabolic dysregulation. Elevated FDFT1 expression promotes hepatic cholesterol accumulation, exacerbating lipotoxicity, oxidative stress, and inflammation—hallmarks of NASH progression. FDFT1-driven dyslipidemia may enhance oncogenic signaling, such as ER stress or mTOR pathways, creating a pro-tumorigenic microenvironment.[11] PGRMC1, a multifunctional membrane protein, contributes to HCC pathogenesis by modulating cell survival and proliferation.[12] Notably, PGRMC1 stabilizes hypoxia-inducible factors, enhancing tumor angiogenesis, and may confer chemoresistance in HCC.[13]

The interplay between FDFT1 and PGRMC1 likely synergizes to accelerate HCC development. Both proteins also intersect in regulating lipid raft composition, influencing growth factor receptor signaling. Targeting FDFT1 and PGRMC1 may hold therapeutic promise, offering novel avenues for intervention. Further research is needed to elucidate their crosstalk and validate the clinical efficacy in NASH-related HCC.

In summary, the role of FDFT1 and PGRMC1 in NASH-HCC is unclear, it is pivotal to investigate how they bridge metabolic dysregulation and carcinogenesis in NASH-associated HCC.

Materials and Methods

Data Collection and Processing

Figure 1 demonstrates the flow diagram of our data analysis process in general terms. Firstly, two NAFLD-associated datasets, for example, GSE89632 and GSE164760 and one Metabolic-Syndrome dataset, such as GSE98895, were screened from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), as shown in Table 1. Gene profile expression data and related annotation files are extracted and collated to obtain the corresponding expression of the genes by the "GEOquery" package. The GSE164760 dataset had 53 NASH-related HCCs, 74 NASH samples, 29 adjacent non-tumor NASH liver samples, as well as 14 control samples. The 74 NASH samples and 14 control samples in GSE164760 were selected and recruited as a training cohort. The GSE89632 dataset included 63 participants (20 Simple Steatosis, 19 NASH, 24 healthy Control) and their partial clinical information. Among them, 20 Simple Steatosis-24 healthy Control in GSE89632 (Va1) and 19 NASH-24Healthy Control in GSE89632 (Va2) were integrated as two validation cohorts. Then, 53 NASH-related HCCs, 14 controls, and 29 adjacent non-tumor NASH liver samples in GSE164760 were selected as the NASH-HCC cohort. The GSE98895 dataset had 20 control subjects and 20 metabolic syndrome (MS) patients. The gene expression at the transcript level was quantile normalized using the "normalizebetweenarrays" function in the "limma" package with a log2 transformation in these RNA-seq datasets. The TCGA-LIHC dataset (50 controls, 371 tumor samples) in the TCGA database (https://portal.gdc.cancer.gov/) was chosen.

Identification of Differentially Expressed Genes (DEGs)

Differentially expressed genes (DEGs) in the training cohort were identified with the "Limma" R package utilizing $|\log2$ fold change (FC)$| > 1$ and $p < 0.05$ as the screening criteria, where log FC > 1 and $p < 0.05$ were up-regulated genes and log FC < 1 and $p < 0.05$ were down-regulated genes. Heatmaps and volcano maps of DEGs were plotted using the

 https://doi.org/10.2147/JHC.S505752

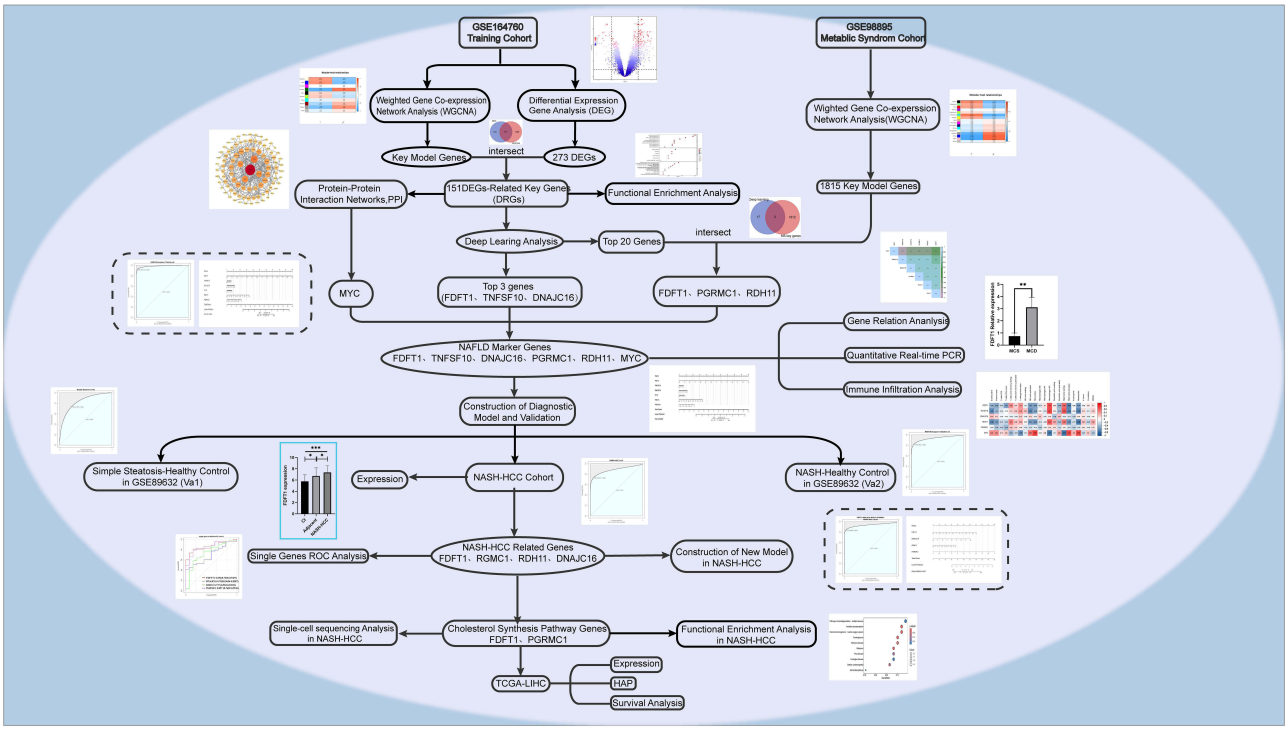


Figure 1 Flow chart of the analysis process. (Ns, P > 0.05; *P < 0.05; **P < 0.01; ***P < 0.001.).

"Pheatmap" R package and the "ggplot2" R package, respectively. After that, 231 up-regulated genes and 42 down-regulated genes were obtained in the training cohort.

Weighted Gene Co-Expression Network Analysis (WGCNA)

To understand the interrelationships among the selected genes and identify the gene modules and key genes that contribute to NAFLD, gene co-expression networks need to be constructed and weights of different modules need to be compared. Weighted Gene Co-expression Network Analysis (WGCNA) is performed to decode co-expression patterns among genes and effectively identify crucial genes as the latest and most popular method. Weighted gene co-expression networks were constructed using the "WGCNA" R package in the training cohort and metabolic syndrome cohort, respectively. All data was preprocessed by filtering out genes with low expression (approaching zero) in all samples. Then, "hclust" function was used to delete the outlier. The scale-free co-expression network was built, and optimal soft threshold was obtained by "pickSoftThreshold" function. The adjacency was calculated using "soft thresholding power" (β) derived from the co-expression similarity. Then, the adjacency matrix was converted into a topological overlap matrix

Table 1 Basic Information on GEO and TCGA Datasets Was Used in the Study

Cohorts	Source	Component
Training cohort	GSE164760	74 NASH and 14 control samples
Simple Steatosis cohort (Va1)	GSE89632	20 Simple Steatosis and 24 control samples
NASH cohort (Va2)	GSE89632	19 NASH and 24 control samples
NASH-HCC cohort	GSE164760	53 NASH-HCCs, 14 controls, and 29 adjacent non-tumor NASH liver samples
Metabolic Syndrom cohort	GSE98895	20 MS patients and 20 control samples
HCC cohort	TCGA-LIHC	50 controls, 371 tumor samples

(TOM), and the gene ratio and dissimilarity were determined. Using average linkage hierarchical clustering, genes with identical expression profiles were classified into the same gene module, with a TOM-based dissimilarity metric and a minimum gene group size (n = 30) for the gene dendrogram. The dissimilarity of module eigengenes was computed, and a cut line at 0.25 for the module dendrogram was chosen. Subsequently, several similar modules were combined for further investigation, and the eigengene network after integration was visualized. The correlation coefficient and P value between model genes and NAFLD were also computed.

Functional Enrichment Analysis

A total of 1381 key genes in the training cohort from four modules were identified by WGCNA analysis. A total of 151 DEGs-related key genes were obtained using the intersection of WGCNA key genes and DEGs. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were performed by "clusterProfiler" R package to screen biological processes (BP), molecular functions (MF), cellular components (CC), and gene-related signaling pathways. All data was analyzed after gene ID conversion and p-value < 0.05 significantly enriched results. Functional enrichment analysis results were presented through the "ggplot2" R package. To illustrate the up- and down-regulated gene set-related signal pathways based on the expression levels of the DEGs-related key gene, Gene Set Enrichment Analysis (GSEA) was performed using the "gsea" function and visualized by the "gseaplot2" function.

Protein–Protein Interaction Network Construction

Protein–Protein Interaction Networks (PPI) are composed of interactions among these proteins encoded by genes. Systematic analyses of protein interactions are important for understanding the functional linkages between proteins in disease states and are often associated with key gene screening. Using the String database (version 11.5; www.string-db.org), with the minimum required interaction score set at 0.400, and an MCODE plug-in was used to identify important interacted genes, the interaction network was constructed and plotted in Cytoscape software. The central genes were selected for subsequent analysis.

Deep Learning and SHAP Value

We use the 151 genes obtained from the previous screening as the main features for deep learning. The 151-dimensional gene features are processed through a two-layer fully connected layer[14] with layer dimensions 10 and 1. The input layer uses a rectified linear unit (ReLU) activation function[15] and the output layer activation function is Softmax. The output of the network is the possibility of suffering from Nonalcoholic Fatty Liver Disease. We use the classic Binary Cross Entropy Loss Function. The neural network was trained for 400 epochs using the Adam optimizer.[16] The training set consisted of 74 NASH samples and 14 control samples in GSE164760 from GEO database. Deep learning interpretability refers to the techniques used to explain and understand how deep learning models make predictions. The SHapley Additive exPlanations (SHAP) value[17] is employed to explain the deep learning model. The SHAP Deep Explainer is applied to assign predicted significance values to each gene and reveals which genes have the greatest impact, positively or negatively, on being sick or not.

Construction of Diagnostic Model and Assessment

FDFT1, TNFSF10, and DNAJC16 were the top three genes calculated by "SHAP Value" function. FDFT1, RDH11, and PGRMC1 were screened after intersecting the 1815 Metabolic Syndrome key genes and the top 20 significant genes. MYC was identified in the PPI network analysis. FDFT1, TNFSF10, DNAJC16, RDH11, PGRMC1, and MYC were united as marker genes in Non-alcoholic fatty liver disease (NAFLD) diagnosis. The diagnostic value of marker genes was evaluated using the diagnostic model, and a nomogram was constructed and visualized by "rms" R package to predict the occurrence of NAFLD. ROC analysis and curve description were performed by the "pROC" R package. The area under the curve (AUC), specificity, and sensitivity were calculated to assess the diagnostic model value. The calibration curve and Decision Curve Analysis (DCA) were performed to assess prediction accuracy and clinical benefit. Two validation cohorts and a NASH-HCC cohort were selected to validate the diagnostic model by ROC curve analysis, respectively. Furthermore, the diagnostic model was compared with clinical indicators in diagnosing NAFLD by ROC curve analysis.

NAFLD Mouse Model Construction and Histological Procedure

Eight-week-old male C57BL/6J mice were divided into control and NAFLD groups. They were fed a methionine- and choline-deficient diet (MCD) for the NAFLD group and a methionine- and choline-sufficient diet (MCS) as a control for 8 weeks. Finally, eight wild-type mice and eight NAFLD mice were enrolled in this study. The serum and pathological tests were used to confirm the successful NAFLD model. H&E staining, Masson's staining, and Oil red O staining were used to show the pathological change of the NAFLD model. Fresh liver samples were paraffin-embedded after fixation with 4% formaldehyde and stained for H&E using 4 mm sections according to the manufacturer's instructions. The hematoxylin staining solution and Ponceau S acid fuchsin stain were used for Masson staining, followed by washing with 2% glacial acetic acid aqueous solution. As for oil red O staining, frozen 8-mm sections were fixed with 10% calcium formaldehyde first. Then, the sections were washed with 60% isopropanol and finally stained with oil red O solution for 30 min at 37°C. During the experiments, all animals were approved by the Animal Experimental Ethics Committee of Shandong Provincial Hospital.

RNA Extraction, Quantitative Real-Time PCR

Trizol was used to extract the liver's RNA, and all RNA has a high purity accompanied by an A260/A280 ratio of 1.9 to 2.2. The extracted RNA was reverse transcribed using a reverse transcription kit to obtain cDNA. Primer sequences for the target genes were designed using Primer 5.0 (Supplementary Table 1). By utilizing the CFX96 Real-Time PCR Detection System and SYBR GREEN Premix Extaq, Real-time quantitative PCR was performed. Reaction system preparation for both reverse transcription and RT-qPCR was performed on ice.

Immune Infiltration Analysis

CIBERSORT uses the principle of linear support vector regression to estimate the abundance of immune cells by decomposing the expression matrix of immune cell subtypes. CIBERSORT provides data on the expression of 22 common immune-infiltrating cells, including immune cells of different cell types and functional states. We used the "CIBERSORT" R package to assess the number of immune cell infiltration from the NAFLD gene expression profile. The "ggplot2" R package was used to show the abundance and proportion of the immune infiltration with a stacked histogram. The Wilcoxon test was used to distinguish differences in the proportions of 22 types of immune cells between NAFLD and control samples, where $p < 0.05$ was considered as the criterion for statistical significance, and the results were presented as bar graphs based on the "ggplot2" package. Thereafter, the Spearman rank correlation coefficient was then applied for correlation analysis between the expression of diagnostic marker genes and the content of 22 infiltrating immune cells, then $P < 0.05$ was considered statistically significant.

ScLiverDB Database

ScLiverDB (http://bioinfo.life.hust.edu.cn/liverdb) is a single-cell transcriptome database of human and mouse livers for revealing liver cellular composition, cellular heterogeneity, and gene expression at the single-cell level in various diseases of the liver, in multiple cell types, and at different developmental stages. The difference in NAFLD marker gene expression at the single-cell level in diseases and healthy controls was analyzed.

Statistical Analysis

The expression profiles in the databases were analyzed using optimal R packages in R software (version 4.3.1.) and Python Version 3.11.1. The experimental data in this study was processed using GraphPad Prism 8.0, and the Shapiro–Wilk normality test was used to check for the normal distribution of the data. Mann–Whitney *U*-test was performed in the absence of normal distribution. Independent samples *t*-test was applied to compare the two groups. The Analysis of Variance (ANOVA) test was used to determine whether three or more groups were significantly different. Correlation analysis was performed via Spearman's rank test. And $p < 0.05$ was considered statistically significant. All experiments were repeated 3 or more times. Ns ($p > 0.05$), *($p < 0.05$), **($p < 0.01$), ***($p < 0.001$).

Results

Differentially Expressed Genes (DEGs) of NAFLD

Firstly, after screening the appropriate dataset, the GSE164760 dataset was chosen and downloaded from the NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/) database, from which 74 NASH samples and 14 control samples were selected as the training cohort. Differentially expressed genes (DEGs) in the training cohort were recognized utilizing the "Limma" R package with |log2 fold change (FC)| > 1, $p < 0.05$ as a screening criterion. Consequently, a total of 273 genes were screened with 231 up-regulate genes and 42 down-regulate genes (Figure 2A).

Weighted Gene Co-Expression Network Analysis (WGCNA) in NAFLD and Metabolic System

The "WGCNA" R package was used to identify significantly important gene models in the training cohort and MS dataset (GSE98895). After deleting the outlier in the training cohort, the Optimal SoftThreshold (β) calculated by the "pickSoftThreshold" function was 9 (Figure 2B). Finally, the topological overlap matrix (TOM) was constructed, and Gene Significance (GS) for NAFLD vs Module Membership (MM) was calculated. The module clustered dendrogram was clipped at a height of 0.25 to detect and merge similar modules (Figure 2C), then 10 NAFLD-related co-expression modules were screened (Figure 2D), with correlation coefficient and P value between model genes and NAFLD was figured (Figure 2E). Genes in "Elightyellow, Meblue, Edarkgreen, and Megrey" models were selected as important genes with the correlation coefficient above 0.5 as a criterion and the correlations between module membership and gene significance in these modules were presented individually (Figure 2F). Accordingly, a total of 1381 genes in the training cohort were screened for further analysis. Similarly, with the optimal soft threshold power (β) being 10, the same process was performed in the metabolic syndrome cohort (GSE98895), "blue" and "pink" models (Supplementary Figure 1D)

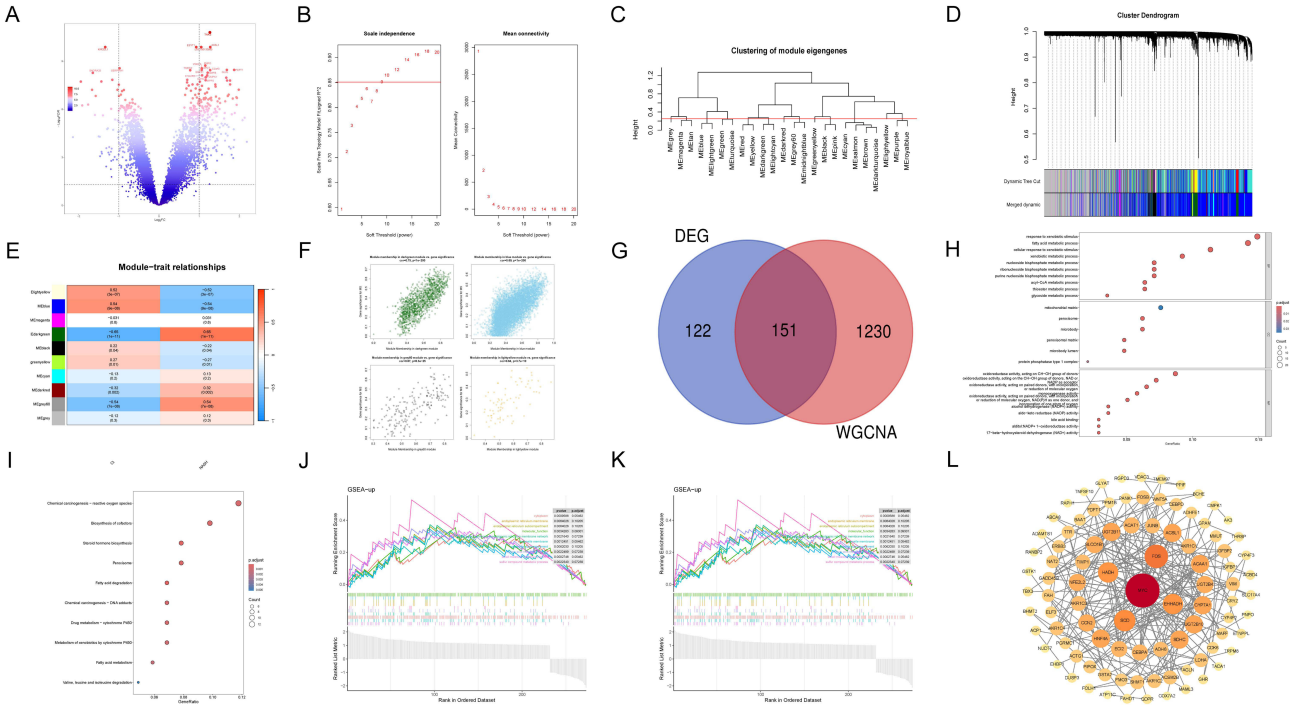


Figure 2 Screening for differential genes and critical gene modules, functional enrichment analyses. (**A**) DEGs are represented by volcano plots, with red representing genes with up-regulated expression and blue representing genes with down-regulated expression. (**B**) Analysis of network topology for various soft-thresholding powers. The left panel shows the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). The right panel displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis). The Optimal SoftThreshold (β) was 9. (**C**) Module clustered dendrogram was cut at a height of 0.25 to detect and combine similar modules. (**D**) Original and combined modules under the clustering tree. (**E**) Heat map of module–trait relationships, each cell contains the corresponding correlation. (**F**) MM vs GS scatter plot for key modules. (**G**) Venn diagram to represent the intersection of DEGs and key module genes. (**H**) GO enrichment analysis related to 151 DEGs-related key genes (DKGs). (**I**) KEGG enrichment analysis. (**J**) Up-regulated gene sets based on GSEA. (**K**) Down-regulated gene sets based on GSEA. (**L**) Protein–protein interaction analysis of 151 DEGs-related key genes (DKGs).

https://doi.org/10.2147/JHC.S505752

were screened as crucial models showing the highest correlation coefficient, which is 0.74. A total of 1815 genes were selected for further analysis (Supplementary Figure 1).

Functional Enrichment Analysis

A total of 151 DEGs-related key genes (DKGs) were obtained using the intersection of WGCNA key genes and DEGs (Figure 2G). Functional enrichment analysis was performed based on DKGs via Gene Ontology (GO) analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses, and Gene Set Enrichment Analysis (GSEA). In the GO analysis (Figure 2H), DKGs were significantly enriched in "response to xenobiotic stimulus", "fatty acid metabolic process" and other bisphosphate metabolic processes (BP). In the Cellular Component (CC) term, DKGs are mainly located in the mitochondrial matrix. In the Molecular Function (MF) term, DKGs are mainly enriched in "oxidoreductase activity" and "NADP/NAD+ activity". Furthermore, KEGG analysis (Figure 2I) showed that DKGs were ample in "Chemical carcinogenesis − reactive oxygen species", "Biosynthesis of cofactors", "Steroid hormone biosynthesis", "Peroxisome", "Fatty acid degradation" and other pathways. GSEA revealed the DKGs were significantly abundant in up-regulated gene sets (Figure 2J), like "cytoplasm", "organelle membrane", and "small molecule metabolic process", and in down-regulated gene sets (Figure 2K), such as "cell surface receptor signaling pathway", "negative regulation of apoptotic process", and "response to abiotic stimulus", respectively.

Screening Hub Genes by Deep Learning and PPI Networks

A neural network model was constructed using deep learning methods to identify feature genes critical for NAFLD diagnosis (Figure 3A). The model architecture consisted of a two-layer fully connected neural network implemented in Python using the PyTorch framework. The input layer processed 151-dimensional gene expression features (from the 151 DEGs-related key genes). The first hidden layer comprised 10 neurons with ReLU (Rectified Linear Unit) activation to capture non-linear relationships, followed by a dropout layer (rate = 0.2) to mitigate overfitting. The output layer used a sigmoid activation function to generate a probability score (0–1) for NAFLD classification. Binary cross-entropy loss was employed as the loss function, and the Adam optimizer (learning rate = 0.0001, β1 = 0.9, β2 = 0.999) was used for training over 400 epochs. The

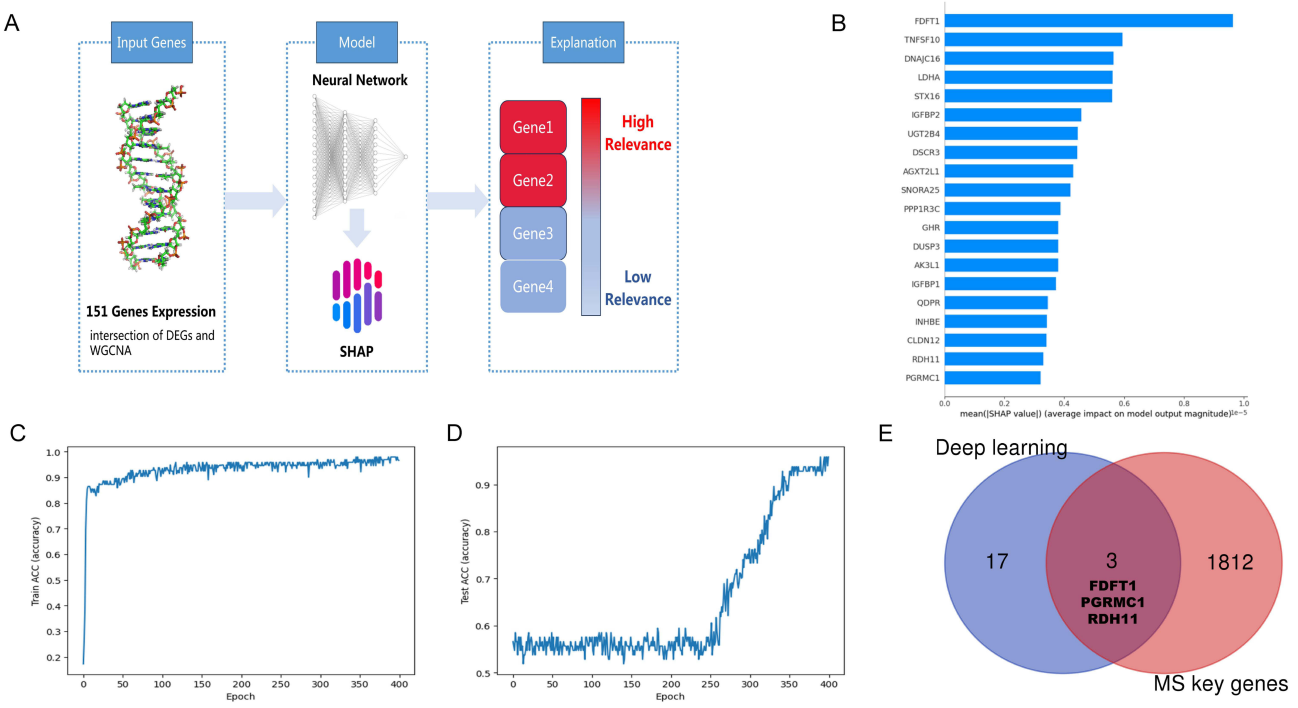


Figure 3 Deep learning for feature gene selection. (**A**) Conceptual diagram of the deep learning model in NAFLD diagnosis to identify feature genes. (**B**) The x-position of the bar chart is the SHapley Additive exPlanations (SHAP) value, that is, the impact on the model's quality. The top 20 most important genes are shown. (**C** and **D**) The training process of the deep learning model in training (**C**) and test (**D**) cohorts. (**E**) Venn diagram to represent the intersection of MS key module genes and top-20 deep learning genes.

model was trained on the GSE164760 dataset (74 NASH and 14 control samples) and validated on two external cohorts from GSE89632 (Va1: 20 Simple Steatosis vs 24 controls; Va2: 19 NASH vs 24 controls). Batch normalization was applied to stabilize training, and early stopping monitored validation loss to prevent overfitting.

To interpret the model, SHapley Additive exPlanations (SHAP) values were calculated using the DeepExplainer algorithm. SHAP values quantify the contribution of each gene to the model's prediction by evaluating marginal effects across all possible feature combinations. The top 20 genes with the highest mean absolute SHAP values were selected as candidates (Figure 3B). These genes exhibited the strongest influence on the model's output, reflecting their diagnostic importance.

Subsequently, the 151 NAFLD key genes were validated by retraining the neural network on the training set and evaluating prediction accuracy on the validation sets (Figure 3C and D). The model achieved a precision of 0.906 on the training set, demonstrating the high prediction accuracy. To ensure generalizability, 5-fold cross-validation was performed, yielding an average precision of 0.886 on the validation sets.

Three genes (FDFT1, RDH11, and PGRMC1) were identified by intersecting the top 20 SHAP-ranked genes with 1815 metabolic syndrome-associated genes from GSE98895 (Figure 3E). MYC was further prioritized via PPI network analysis (Figure 2L), where it served as a central hub with the highest connectivity (degree = 15) in the StringDB network (interaction score > 0.9). Consequently, FDFT1, TNFSF10, DNAJC16, RDH11, PGRMC1, and MYC were united as feature genes for NAFLD diagnosis.

This approach leveraged deep learning's ability to handle high-dimensional genomic data and SHAP's interpretability to bridge mechanistic insights with predictive accuracy, ensuring both reliability and biological relevance.

Construction of Diagnostic Model Based on Six Feature Genes

The diagnostic value of hub genes was evaluated using the diagnostic model, and a nomogram was constructed and visualized by "rms" R package to predict the occurrence of NAFLD (Figure 4A). To further validate the model, two external cohorts were utilized to evaluate its efficacy, which are 20 Simple Steatosis-24 healthy Control in GSE89632

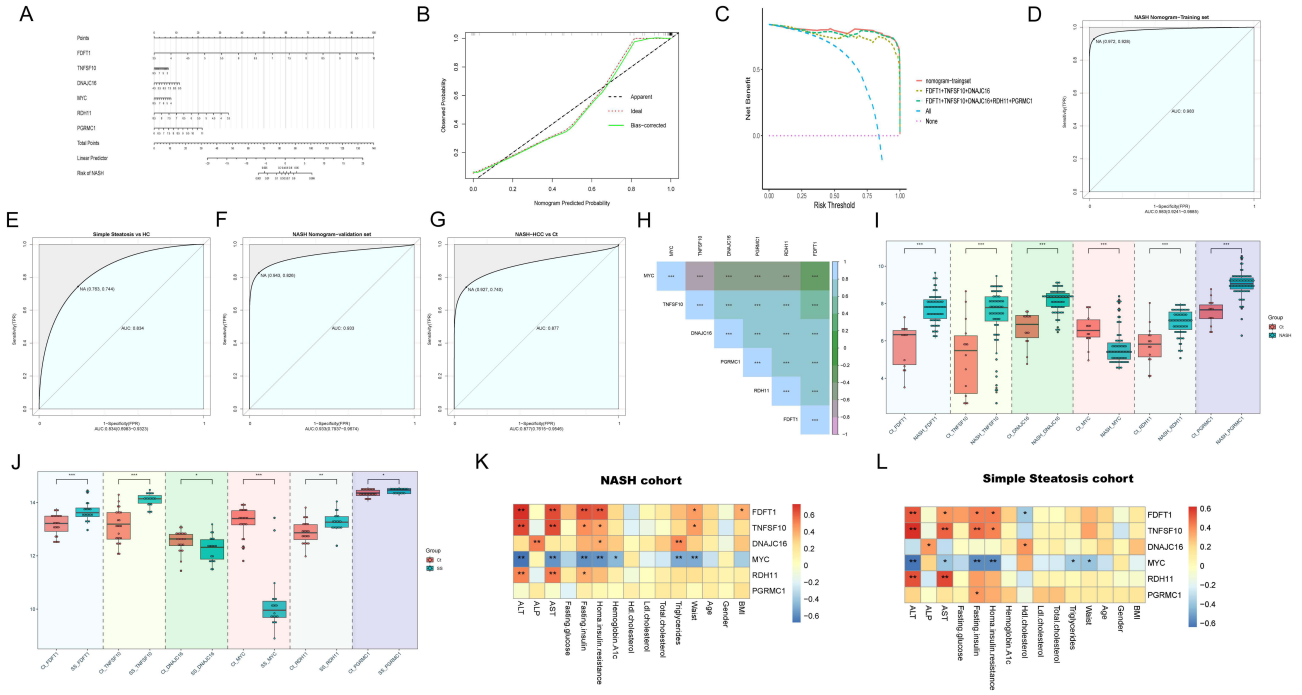


Figure 4 Construction of diagnostic model based on 6 feature genes, gene expression in the database, and correlation of six hub genes. (**A**) Nomograms are used to predict the incidence of NAFLD. (**B**) Calibration curves for assessing the predictive power of the nomogram model. (**C**) DCA curves for assessing the clinical value of the nomogram model. (**D**) ROC curves analysis of the feature genes in the training cohort. (**E-G**) ROC curves analysis of the feature genes in the three validation cohorts, including simple steatosis, NASH, and NASH-HCC. (**H**) Correlation analysis between the feature genes in NAFLD. (**I**) Feature gene expression in the NASH cohort. (**J**) Feature gene expression in the simple steatosis cohort. (**K** and **L**) Spearman correlation analysis of the feature genes and clinical variables in the simple steatosis and NASH cohorts. (Ns, P > 0.05; *P < 0.05; **P < 0.01; ***P < 0.001.).

https://doi.org/10.2147/JHC.S505752

(Va1) and 19 NASH-24Healthy Control in GSE89632 (Va2). The calibration curve (Figure 4B) and Decision Curve Analysis (DCA) (Figure 4C) showed the model is excellent in diagnostic accuracy and clinical utility. The predicted probability of the model is closer to the ideal curve with a slope of 1 (0–1) in the calibration curve. Decision Curve Analysis (DCA) showed that the model can bring more clinical benefit than a single part of six hub genes. It is more important that the nomogram model has a high diagnostic power with the area under the curve (AUC) of 0.983 (0.9241–0.98885) and the C-index is 0.9873 (Figure 4D). Subsequently, a similar ROC curve analysis was performed in two validation cohorts (Va1, Va2) with their AUC values are 0.834 (0.6983–0.9323) and 0.933 (0.7937–0.9674) (Figure 4E and F). Then, the diagnostic effects of the nomogram models and clinical data in Va1 and Va2 were compared, respectively (Supplementary Figure 2). Clinical data were grossly divided into four groups, including physical measurement index, glucose metabolism, lipid metabolism, and liver function indicator. Then, the diagnostic efficacy of the model was compared, respectively, in four clinical data groups, and all results showed that the model was optimal and had similar diagnostic performance with the liver function indicator, implying that the feature genes have high application value in clinical practice (Supplementary Figure 2).

Correlation and Biological Functions of Six Feature Genes

To investigate the correlation among the six hub genes further, correlation analysis was performed and visualized, which showed that these genes have a significantly high correlation (Figure 4H). The expression levels of six hub genes were compared, respectively, in NASH and Simple Steatosis samples with control samples. Except for MYC, the expression of the other five genes was higher in the training NASH group compared to the control group with all significance tests being less than 0.001 (Figure 4I). Furthermore, similar results were calculated except that the DANJC16 expression level was lower in the simple steatosis group compared to the control group (Figure 4J). Subsequently, Spearman's rank test processed the correlation between the six hub genes and clinical data in Va1 and Va2 cohorts (Figure 4K and L). Apart from PGRMC1, other genes are significantly correlated with multiple clinical indicators, such as ALT and AST, implying that the feature genes have a potent association with liver function in NAFLD.

Then, the biological functions of six feature genes (FDFT1, TNFSF10, DNAJC16, RDH11, PGRMC1, MYC) were analyzed by the GeneMANIA database (http://genemania.org/), and 20 related co-expression genes were identified (Figure S4A). KEGG and GO enrichment analysis based on the 26 genes were processed in the Metascape database (https://metascape.org/). The results showed that they were mainly enriched in sterol biosynthetic process and mevalonate arm of cholesterol biosynthesis pathway (Figure S4B). The sterol biosynthetic process is a conserved metabolic pathway responsible for synthesizing cholesterol and related sterols, essential for membrane integrity, hormone production, and bile acid synthesis. The mevalonate pathway serves as the foundational arm of cholesterol biosynthesis, initiating with acetyl-CoA condensation to form 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA).

To further investigate potential mechanisms in TRRUST database (https://www.grnpedia.org/), SP1 was identified as the most important transcription factor. SP1 is a ubiquitously expressed transcription factor that binds GC-rich promoter regions. Then, we searched upstream and downstream transcription factors of feature genes, respectively, and SP1 was found to be an upstream transcription factor of SREBF1, which synergizes with SREBF1 to enhance the transcription of genes involved in lipid synthesis. Furthermore, FDFT1 was identified as the downstream molecule of SREBF1, then SP1/SREBF1/FDFT1 transduction pathway was found in the cholesterol biosynthesis pathway.

Experimental Verification of Six Feature Genes Expression

Subsequently, we try to validate the expression of feature genes in the NAFLD model based on the MCD diet-fed mouse and MCS-fed mouse as control. After 8 weeks, the NAFLD group showed obvious hepatic fat accumulation and ballooning degeneration. Also, the ELISA test showed the NAFLD group had a higher serum ALT and AST level (Figure 5A). Respective liver gross view (Figure 5B), H&E staining, Masson's staining, and Oil red O staining (Figure 5C) in the NAFLD and control groups were displayed. All these results proved that the NAFLD mouse model was convincing. Then, Real-time Quantitative Reverse Transcription Polymerase Chain Reaction (RT-qPCR) was performed after extracting total RNA from mouse livers to compare the difference in feature gene expression. And

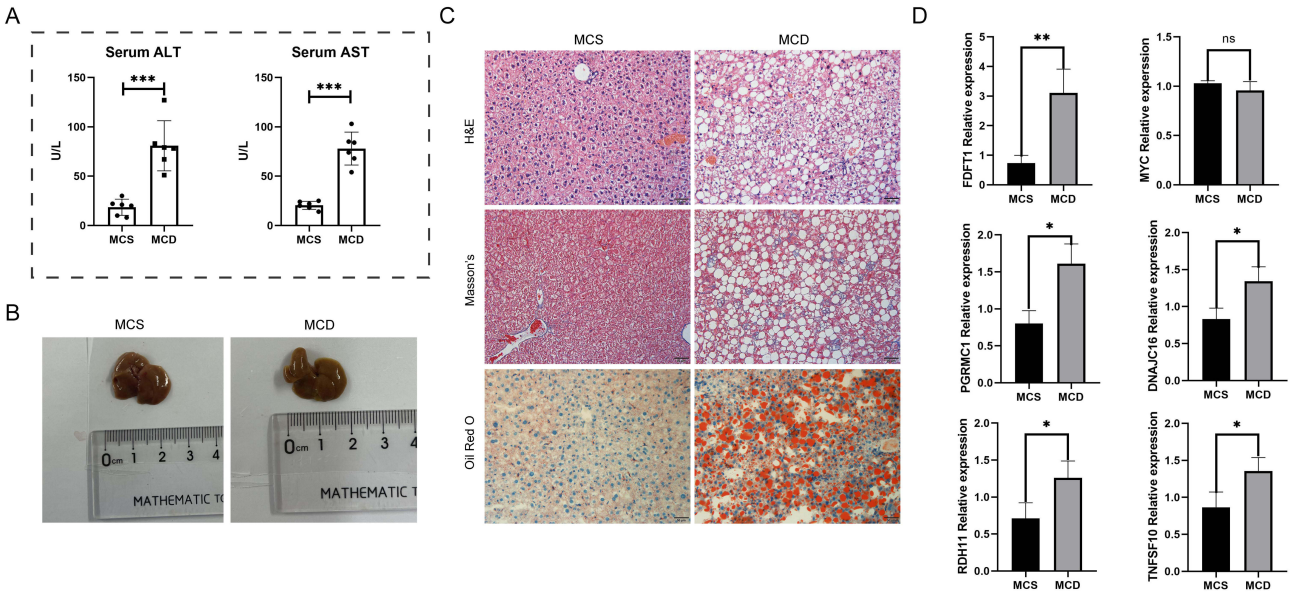


Figure 5 Experiments in NAFLD mouse model. (**A**) Serum ALT and AST were measured by Enzymatic techniques kits. (**B**) Liver gross comparison between MCS and MCD-fed mice. (**C**) H&E staining, Masson's staining, and Oil red O staining in the NAFLD and control groups. (MCS diet for control groups, original magnification, × 20); (MCD diet for NAFLD groups, original magnification, × 20) (**D**) Relative mRNA expression of the feature genes in two groups was compared by real-time quantitative fluorescence PCR (RT-qPCR). (N = 4-6) (Ns, P > 0.05; *P < 0.05; **P < 0.01; ***P < 0.001.).

except for MYC, the expression levels of the remaining five genes were significantly higher in the NAFLD group, and MYC had no difference in the two groups (Figure 5D).

Immune Infiltration Analysis Based on CIBERSORT

Immune infiltration analysis was demonstrated based on the "CIBERSORT" R package to evaluate the association between NAFLD and immunological mechanisms. Meanwhile, immune cells with great correlation to the six marker genes were explored. To compare the difference in the estimated proportion of 22 immune cells between NAFLD and control groups, a stacked histogram was painted (Figure 6A) and a heatmap of immune cells was plotted (Figure 6B). Subsequently, with $p < 0.05$ as a screening criterion, there are 7 immune cells screened as significantly different between the two groups (Figure 6C). The barplot was printed to illustrate that "γδ T cells", "Mast cell resting", "Macrophage M2", and "T cells CD4 memory resting" had higher abundance in the NAFLD group, while "Mast cells activated", "Neutrophils", "Monocytes" and "B cells naïve" had lower abundance in NASH group. Similar results were also obtained from the analysis of genes' correlation with immune cells, except for MYC, other five genes were highly positively correlated with "Macrophages.M2", "γδ T cells", "T.cells.CD4.memory. resting" and "Mast. cells. Resting", while negatively correlated with "Mast. cells. Activated" and "Neutrophils" (Figure 6D). Based on the function of "Macrophages.M2" and "γδ T cells", these results revealed that feather genes have intricate and inseparable connections with hepatic lobular remodeling, angiogenesis, and fibrosis, and are even involved in the regulation of innate immunity.

Validation of FDFT1 and PGRMC1 in NASH-HCC Cohort

To further validate the diagnostic efficacy of the diagnostic model in NASH-related HCC, the NASH-HCC cohort was also selected from GSE164760 in the GEO database from which 53 NASH-related HCCs, 14 controls, and 29 adjacent non-tumor NASH liver samples were contained. The nomogram constructed by 6 hub genes was used in the NASH-HCC cohort, and AUC is 0.877 (0.7615–0.9546) (Figure 4G). Then, single feature gene expression was compared in the NASH-HCC cohort based on "ANOVA" significance analysis, TNFSF10 and MYC have no significant expression difference, while expression of the other four genes showed an increasing trend in the contrast from normal to paracancerous tissue to cancerous tissue, especially in FDFT1 and PGRMC1 (Figure 6E). A new diagnostic model for NASH-HCC was constructed by FDFT1, DNAJC16, RDH11, and PGRMC1 and a nomogram was visualized

https://doi.org/10.2147/JHC.S505752

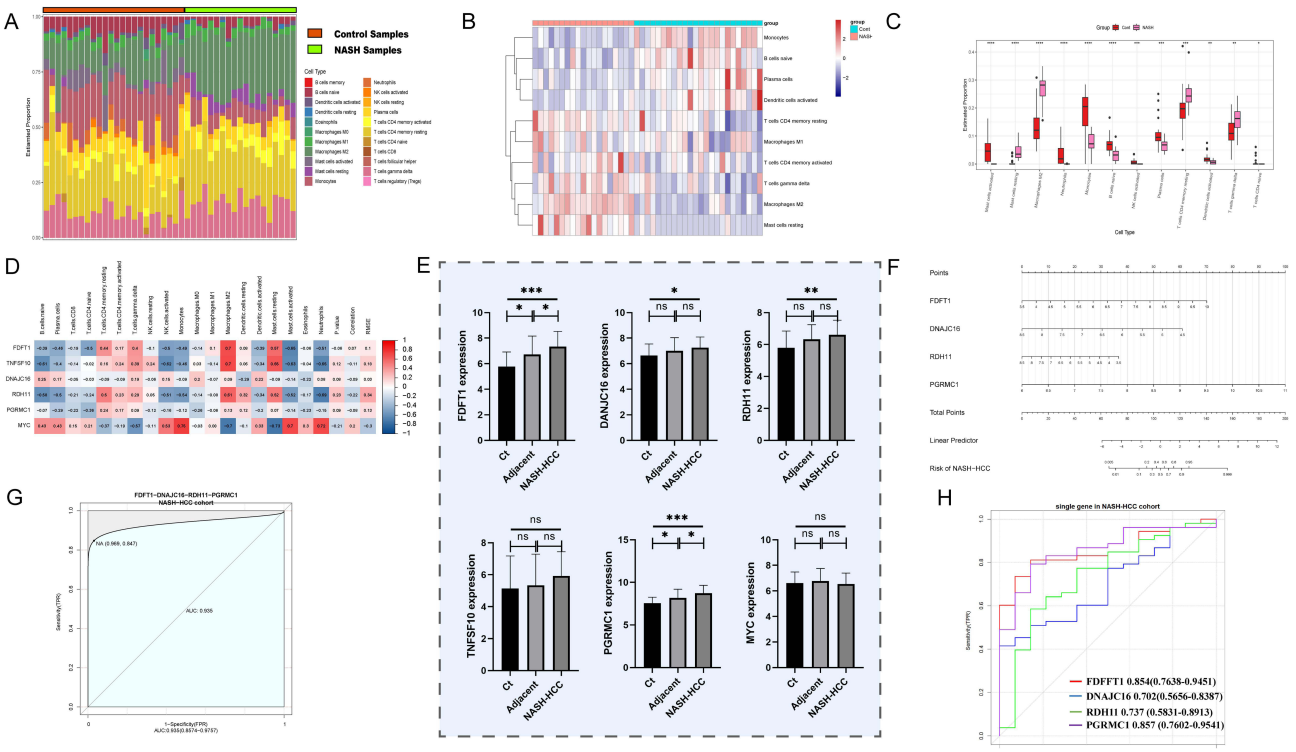


Figure 6 Immune infiltration analysis in NAFLD, validation of feature genes in NASH-HCC cohort. (**A**) The percentage of 22 types of immune cells in the NAFLD and control groups. (**B**) Heatmap of immune cells in NAFLD and control groups. (**C**) The difference in immune cells between the NAFLD and control groups. (**D**) Correlation analysis between feature genes and immune cells, each cell contains the corresponding correlation. (**E**) Expression of feature genes in healthy controls, tumor-adjacent tissues, and tumor tissues. (**F**) Nomogram in NASH-HCC diagnosis based on differential feature genes. (**G**) ROC curve of the new diagnostic model for NASH-HCC. (**H**) ROC curves of single feature genes for NASH-HCC. (Ns, P > 0.05; *P < 0.05; **P < 0.01; ***P < 0.001.).

(Figure 6F). The ROC curve was calculated with its AUC is 0.935 (0.8574−0.9757) and C-index is 0.938 (Figure 6G). To further compare the diagnostic efficacy of individual genes, ROC curves of single genes were painted (Figure 6H). FDFT1 (AUC: 0.854) and PGRMC1 (AUC:0.857) surpass DNAJC16 (AUC: 0.702) and RDH11 (AUC:0.737) with a better performance in diagnosing NASH-HCC.

Afterward, the correlation between the other genes and the two target genes in the NASH-HCC cohort was calculated individually, FDFT1 and PGRMC1 related significantly correlative genes were screened with $p < 0.05$ and correlation coefficient > 0.6 at least as a filtering condition. FDFT1 GO enrichment analysis (Figure 7A) manifested that FDFT1 relevant genes were mainly enriched in "cellular respiration", "oxidative phosphorylation" and ATP synthesis related biological process (BP term); "mitochondrial inner membrane" and "mitochondrial protein−containing complex" (CC term); "structural constituent of the ribosome" and "proton transmembrane transporter activity" (MF term). PGRMC1 GO enrichment analysis (Figure 7B) revealed that PGRMC1-relevant genes were mainly enriched in the "fatty acid metabolic process", "carboxylic acid catabolic process" and "cellular lipid catabolic process" (BP term); "peroxisome" and "mitochondrial matrix" (CC term); "enzyme inhibitor activity" and "electron transfer activity" (MF term). FDFT1 KEGG enrichment analysis (Figure 7C) manifested that FDFT1-relevant genes were mainly enriched in the "oxidative phosphorylation" pathway and other nervous system-related diseases. PGRMC1 KEGG enrichment analysis (Figure 7D) revealed that PGRMC1-relevant genes were mainly enriched in "peroxisome" and "amino acids and fatty acid degradation" pathways. All these enrichment analysis results demonstrated that FDFT1 has a critical relation with cellular respiration-related biological processes in the mitochondrial, such as "oxidative phosphorylation", and PGRMC1 was highly related to the catabolic process in lipid and amino acids, and also to the peroxisome pathway.

Single-Cell Sequencing Analysis of FDFT1 and PGRMC1

Subsequently, the distribution of FDFT1 and PGRMC1 in different cell types was calculated and visualized based on single-cell analysis in A Database of Human and Mouse Liver Transcriptome Landscapes at Single-cell Resolution

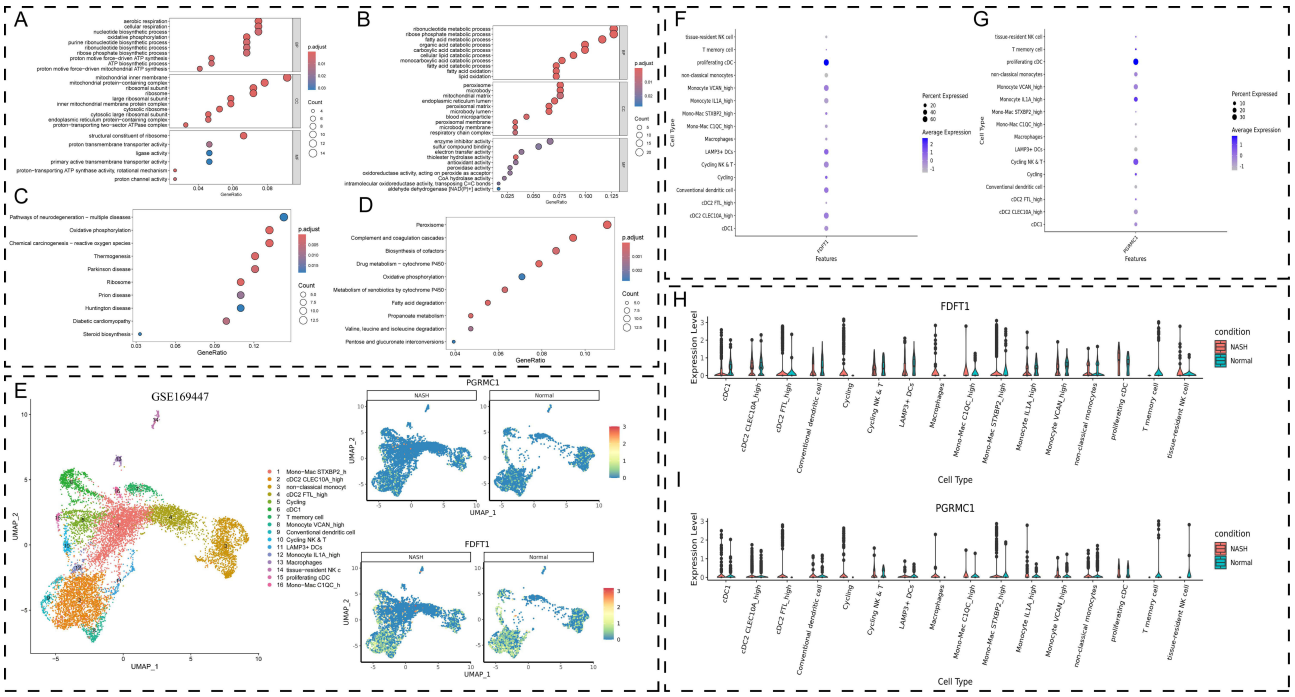


Figure 7 Single gene enrichment analyses and single-cell sequencing analyses of PGRMC1 and FDFT1. (**A-D**) Single gene enrichment analyses of PGRMC1 and FDFT1 in NAFLD, including GO analyses of FDFT1 (**A**) and PGRMC1 (**B**), and KEGG analyses of FDFT1 (**C**) and PGRMC1 (**D**). (**E**) UMAP was plotted after cluster analysis in GSE169447, each color represents a cell population identified after clustering. (**F** and **G**) Average expression and the percentage of cells expressing in different cell types were plotted. (**H-I**) Expression of FDFT1 and PGRMC1 in different cell types was compared between NAFLD and control groups.

(scLiverDB, http://bioinfo.life.hust.edu.cn/liverdb). First, the Non-Alcoholic Steatohepatitis data set (GSE169447) (Figure 7E) from the GEO database was chosen and target genes' average expression and the percentage of cells expressing in different cell types was painted (Figure 7F and G), which illustrated that FDFT1 was abundantly expressed in proliferating cDC, monocyte, cycling NK&T cell and conventional dendritic cell, while PGRMC1 was also abundantly expressed in proliferating cDC, monocyte and cycling NK&T cell. Furthermore, the expression of FDFT1 and PGRMC1 in different cell types was compared between NAFLD and control groups (Figure 7H and I). Consistent with the results of the immune infiltration analysis, compared with the control group, the expression of FDFT1 and PGRMC1 in macrophages was much higher in the NASH group. This provides further evidence that FDFT1 and PGRMC1 may regulate NAFLD progression by influencing the monocyte-macrophage system and innate immunity. Also, FDFT1 and PGRMC1 may be inextricably linked to the antigen presentation function of dendritic cells in the pathogenesis of NAFLD.

FDFT1 and PGRMC1 in HCC

To further demonstrate the important biological expression of FDFT1 and PGRMC1 in hepatic diseases, the TCGA-LIHC dataset (50 controls, 371 tumor samples) in the TCGA database (https://portal.gdc.cancer.gov/) was chosen. The expression of FDFT1 and PGRMC1 in the HCC cohort was analyzed in The University of Alabama at Birmingham Cancer Data Analysis Portal (UALCAN, https://ualcan.path.uab.edu/). FDFT1 has a significantly high expression in HCC samples with a p-value of 2.92700308435201E-11 (Figure 8A) and the high expression group has a shorter survival (Figure 8B and Supplementary Figure 3A), while PGRMC1 has no difference in expression level of tumor and control groups (Supplementary Figure 3B). Furthermore, an overall survival analysis was performed and results showed that the higher expression of FDFT1 was accompanied by a shorter overall survival (OS) (Figure 8C) but not with recurrence-free survival (RFS) and Progression-Free-Survival (PFS) (Supplementary Figure 3C). The pan-cancer analysis in FDFT1 was executed in Tumor IMmune Estimation Resource database (TIMER, https://cistrome.shinyapps.io/timer/) (Figure 8D), which showed that FDFT1 has significant expression difference in colon adenocarcinoma, stomach adenocarcinoma,

https://doi.org/10.2147/JHC.S505752

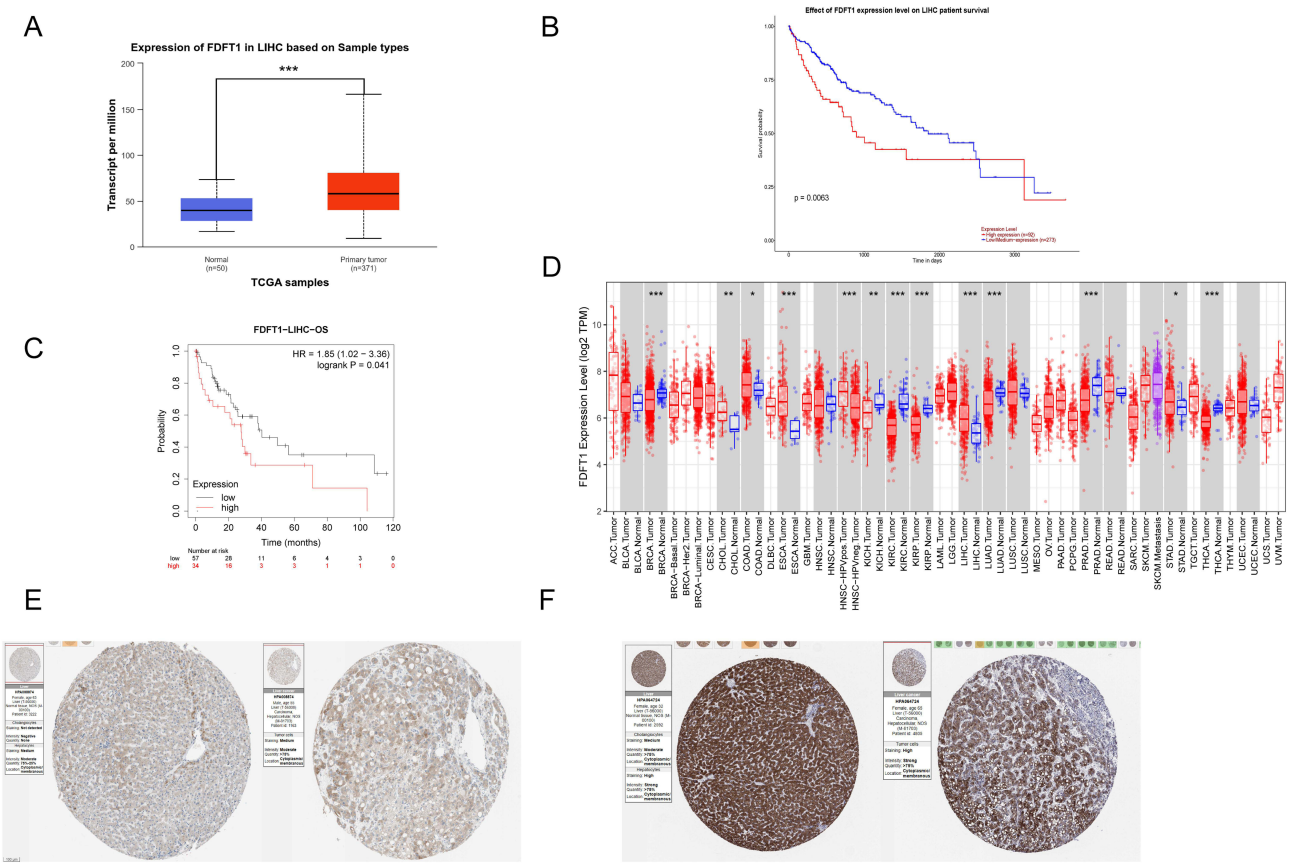


Figure 8 FDFT1 and PGRMC1 in HCC. (**A**) Expression of FDFT1 in TCGA liver cancer database. (**B**) Survival analysis of FDFT1 is compared in high and low-expression HCC groups. (**C**) Overall Survival analysis of FDFT1 expression level in HCC. (**D**) Pan-cancer analysis of FDFT1 in TIMER database. (**E** and **F**) The protein expression of FDFT1 and PGRMC1 in normal and HCC samples in the HPA database. (Ns, P > 0.05; *P < 0.05; **P < 0.01; ***P < 0.001.).

esophageal carcinoma, kidney carcinoma, thyroid carcinoma, and breast carcinoma, similar results were also obtained in the pan-cancer analysis of PGRMC1 (Supplementary Figure 3D). According to the results, FDFT1 and PGRMC1 have significant expression differences in multiple digestive cancers, implying that they may be remarkable marker genes in malignant tumors of the digestive system. Ultimately, the protein expression was searched in the human protein profile (HAP) database (https://www.proteinatlas.org/). Then, the protein expression of PGRMC1 and FDFT1 in normal liver tissue and hepatocellular carcinoma tissue was compared (Figure 8E and F).

Discussion

NAFLD is more and more becoming the most universal chronic liver disease, with accumulation of lipids due to disorders of lipid metabolism, and NASH is becoming an increasingly important cause of liver cancer development. Moreover, as a multi-systemic metabolic disease, NAFLD often co-exists with T2DM and metabolic syndrome, exacerbating the disease process, its global incidence is increasing year by year, and the pathogenesis is still not fully understood, constituting a major challenge in liver metabolic diseases.[18] Whereas, highly efficient and safe therapeutic modalities are not yet available worldwide, and current diagnostic methods still have many limitations. So, it is critical to explore new biomarkers to shed new light on the pathogenesis, early diagnosis, treatment, and prevention of NAFLD, especially for NASH-related HCC.

In this research, we selected a portion of GSE164760 as the NAFLD training set, and 151 DEGs-related key genes (DKGs) were screened after difference and WGCNA analysis. Then, GO analysis, KEGG analysis, and GSEA analysis were performed, and these genes were mainly enriched in lipid metabolism-related pathways and energy metabolism-related pathways, indicating cholesterol biosynthesis pathway. To further screen feature genes, deep learning analysis in

151 DEGs-related key genes (DKGs) was processed, and the top 20 significant genes in the neural network model were acquired for their higher SHAP value. Considering the co-morbidity with metabolic syndrome, 1815 key MS genes were identified to take an intersection with top-20 NAFLD genes, and three common genes in both MS and NAFLD were obtained. PPI analysis screened MYC as a central gene. SHapley Additive exPlanations (SHAP) showed that FDFT1, TNFSF10, and DNAJC16 have the highest values, indicating that the sensitivity and specificity of the three genes are far superior to other genes. So, FDFT1, TNFSF10, DNAJC16, RDH11, PGRMC1, and MYC were recognized as feature genes in NAFLD.

Farnesyl-diphosphate farnesyltransferase 1 (FDFT1) encodes squalene synthase, which is a key enzyme participating in coordinating the condensation of two farnesyl pyrophosphate (FPP) molecules, a key step in sterol biosynthesis. The expression of FDFT1 is strongly associated with NAFLD activity scores and has a possible causal link to cholesterol metabolism, energy use, and fibrosis progression in the liver.[19,20] Tumor necrosis factor (TNF) superfamily member 10 (TNFSF10) is a member of the TNF family and is predominantly expressed on the surface of immune cells, such as cytotoxic T cells and natural killer (NK) cells. TNFSF10 can induce apoptosis in tumor or infected cells and may also trigger non-apoptotic signaling pathways through activation of pro-inflammatory pathways. Some studies showed human TNF-related apoptosis-inducing ligands that could induce activated hepatic stellate cell apoptosis might be feasible for antihepatofibrotic therapy.[21] DNAJC16 is a member of the DNAJ family and is a co-chaperone that together with HSP70s controls protein homeostasis.[22] The upregulation of DNAJC16 expression can enhance cell apoptosis, and previous studies have shown autophagy is enhanced in the early stage of NAFLD and may slow down the progression of NAFLD by inhibiting the "second strike" that causes NAFLD. Retinol dehydrogenase 11 (RDH11) is a microsomal short-chain dehydrogenase that recognizes retinoids as substrates and prefers NADPH as a cofactor.[23] A study has validated that RDH11 has crosstalk with cholesterol; on the one hand, cholesterol levels regulate RDH11 transcription through sterol-regulatory element-binding protein (SREBP2), and on the other hand, RDH11 deficiency can lead to elevated cholesterol levels.[24] Progesterone receptor membrane component 1 (PGRMC1) belongs to the membrane-associated progesterone receptor (MAPR) gene family and regulates cholesterol synthesis by activating the lanosterol demethylase.[12] It has been shown that PGRMC1 expression is inversely associated with hepatocellular carcinoma survival and that PGRMC1 influences the development of HCC by regulating EGFR-mediated pro-inflammatory responses.[25] MYC encodes transcriptional regulators that cause activation or repression of key genes involved in multiple biological processes at multiple stages of NAFLD, such as hepatocarcinogenesis, metabolic reprogramming, and transformation of the immunosuppressive microenvironment.[26] MYC enhances cholesterol biosynthesis and supports cell proliferation; the higher MYC mRNA level was accompanied by higher HDL cholesterol and unsaturated fatty acid proportions, as well as lower fat mass, glucose, and transaminase.[27,28]

While prior studies primarily focused on transcriptomic profiling in NAFLD or mechanisms of lipid metabolism,[13,29] our study uniquely integrates RNA sequencing in NAFLD with data from Metabolic Syndrome patient samples. Deep learning and comprehensive bioinformatics allowed us to identify feature genes as novel biomarkers by screening the disease diagnostic model. Earlier works established the association between feature genes and steatosis through knock-out mouse models. Our study advances this by demonstrating through patient sample data, and we show the interrelationships between the feature genes. Compared to other biomarkers screened by published articles,[30,31] we have innovated screening methods and validated them at various stages in the NAFLD disease spectrum.

Then, based on these six genes, we constructed a new disease diagnostic model and evaluated various aspects of the model's efficacy using ROC curves, C indices, calibration curves, and DCA analysis. The model was validated in different datasets, such as the simple steatosis group and the NASH group. All these results prove that the diagnostic efficacy of this model is excellent. The diagnostic efficacy of the model is still superior to clinical indicators because of its high AUC value, such as compared to physical measurement index, glucose metabolism, lipid metabolism, and liver function indicators. In the simple steatosis group, the model remains effective, revealing that this model may also help in the early diagnosis of NAFLD. With the increasing prevalence of NAFLD, the universality of liver biopsy as the gold standard has been limited, and the development of non-invasive diagnostic methods and new biomarkers have become mainstream.[32,33] Studies such as multi-omics and non-coding RNA have provided new markers for providing non-invasive diagnostic modalities.[34] During the last few years, indexes such as FLI (fatty liver index)[35] and HSI (hepatic

steatosis index)[36] did not achieve the desired results. Our research contributes to the development of new markers and can drive translational applications through the development of simplified assay panels, such as the development of serum test kits related to feature genes. If conditions permit, large-scale prospective randomized controlled trials may be considered to validate the application of new markers.

Compared to the control group, RT-qPCR experiments based on RNA extracted from mice fed high-fat diets demonstrated that the expression of all genes was significantly higher in the disease group, except for the expression of the MYC, which was not significantly different. Subsequently, the differential distribution of immune cells and the relationship of feature genes to multiple immune cells were also analyzed. Consistent with the results of immune infiltration analyses, single-cell sequencing analyses also suggested that FDFT1 and PGRMC1 may regulate NAFLD progression by influencing the monocyte-macrophage system and innate immunity. All these results suggest that signature genes are inextricably linked to Macrophages.M2 and γδ T cells, and it is hypothesized that signature genes may accelerate hepatic fibrosis, pseudofollicular formation, and even cancer in NAFLD by modulating the immune microenvironment.[37] Hepatic macrophages play a central role in the pathogenesis of chronic liver injury and are considered a potential target in the fight against liver fibrosis. Macrophages are divided into two subpopulations, M1-type and M2-type macrophages, based on their function.[38] Functionally, M2 macrophages participate in the removal of cellular debris, apoptotic cells, and tissue repair and have characteristics that result in vascular lesions, scarring, and tissue fibrosis.[39] The γδ T cells are one of the important components of innate immune cells, and the liver is one of the organs where they are most abundant. Several studies have shown that hepatic γδ T cells play a role in recruiting neutrophils in liver injury by secreting IL-17A. Interestingly, macrophages and γδ T cells have crosstalk in neutrophil accumulation and cellular inflammation.[40] In conclusion, the relationship between feature genes and immune cells still needs to be further explored, and our findings may provide new directions for the study of immune mechanisms in NAFLD.

To further validate the diagnostic efficacy of the model and the sensitivity of the feature genes in NASH-HCC, a cohort in GSE164760 was selected, the model also showed excellent diagnostic efficacy in the NASH-HCC group and single-gene ROC analysis showed that FDFT1 and PGRMC1 are the most sensitive feature genes. At the same time, in the NASH-HCC group, expression of feature genes was compared in control, tumor-adjacent tissue, and tumor tissue, which showed expressions of FDFT1 and PGRMC1 have a significantly increasing tendency. Emerging evidence from limited studies has begun to investigate the mechanistic roles of FDFT1 and PGRMC1 in hepatocellular carcinoma (HCC) pathogenesis. A study showed that targeting FDFT1 impedes HCC progression through metabolic remodeling, where depletion of cholesterol and bile acid pools mediated by the HNF4A-ALDOB-AKT1 cascade restricts oncogenic growth.[41] Mechanistically, suppressing FDFT1 expression reduces HCC cell proliferation, metastatic potential, and tumorigenesis in both experimental models and living systems, while excessive FDFT1 expression enhances malignant progression by accelerating HCC cell growth and dissemination. Some studies have verified that FDFT1 has a relation with ferroptosis of liver cancer cells, reactive oxygen species (ROS) and glutathione peroxidase (GPX4) are involved in the regulation.[42] As for PGRMC1 in liver cancer, PGRMC1 has been reported to have interactions with epidermal growth factor receptor (EGFR), which is proven to mediate hepatocellular carcinoma development.[43] Bioinformational analysis based on GEO data has demonstrated that the mRNA expression level of PGRMC1 in hepatocellular carcinoma patients is inversely correlated with survival, and PGRMC1 knockdown plays an anti-tumor role via suppression of pro-inflammatory immune responses.[44] Meanwhile, in another article, PGRMC1 significantly blocks c-Myc-induced orthotopic HCC formation by inhibiting c-Myc protein translation via the PERK/p-eIF2α signaling pathway.[45] These findings highlight the critical role of FDFT1 and PGRMC1 in linking metabolic disturbances to cancer development in HCC, suggesting innovative therapeutic strategies and diagnostic biomarkers in NASH-HCC.

To explore the key signaling pathways associated with PGRMC1 and FDFT1, single-gene enrichment analyses showed that FDFT1 and PGRMC1 are primarily associated with energy metabolic pathways and lipid metabolic pathways in cells, implying potential crosstalk between cholesterogenic genes and glucose metabolic pathway. The Akt/mTOR pathway, a classical glucose metabolism and cell proliferation signaling pathway, has been demonstrated to have an interaction with FDFT1 by several studies.[29,46] Several studies have shown that knockdown or inhibition of PGRMC1 can lead to an increase in fasting glucose and HbA1c, causing a range of glucose metabolism disorders.[47] Lipid metabolism and glucose metabolism are inextricably intertwined and have mutual transformation.

Whether there is an interaction between FDFT1 and PGRCM1 is an interesting question, and it was proved in the STRING database (STRING: functional protein association networks (string-db.org)) that they have co-expression and interaction. Although no study has yet involved the interaction of FDFT1 with PGRMC1 in the liver, a study about molecular mechanisms of breast cancer progression has identified FDFT1 as one of the downstream target molecules of PGRMC1 by mass spectrometry analysis.[13] In addition, Chromatin immunoprecipitation (ChIP)-chip combined with transcriptional profiling on HepG2 human hepatoma cells has demonstrated that FDFT1 is a target of SREBP.[48] Also, PGRMC1 can bind to SREBP as a key factor regulating hepatic de novo lipogenesis (DNL).[25] These results suggest that FDFT1 interacts with PGRMC1 during cholesterol synthesis and that PGRMC1 is an upstream molecule of FDFT1.

Furthermore, to investigate the importance of PGRMC1 and FDFT1 in hepatocellular carcinoma, their expression was analyzed based on the TCGA database, and the results showed that only FDFT1 was highly expressed in cancer tissues, and the OS was less in the high-expression group than in the low-expression group. Finally, the protein expression of FDFT1 and PGRMC1 was searched and FDFT1 in tumor cells presents a deeper staining with a higher expression.

The main innovations of this article are that there are currently no studies in the field of NAFLD-related research that combine bioinformatics analysis with deep learning and few studies that variously validate the new markers identified in a NASH-HCC cohort. Of course, our study still has limitations; we have only performed simple experiments to verify the expression of feature genes in a NAFLD mouse model, and more and more detailed mechanism-related experiments are needed. Although previous studies have also screened for signature genes in NASH-HCC, the diagnostic efficacy of key molecules in the cholesterol synthesis pathway in NASH-HCC was identified for the first time, further supporting the perspective that the cholesterol-synthesis pathway plays a key role in the pathogenesis as well as the disease progression of NASH-HCC. We hope this study will bring some new valuable ideas for NASH-HCC research.

Conclusion

In this article, we combine bioinformatic analysis with deep learning to leverage the high sensitivity of neural networks and finally identify six feature genes. Experiments in the NAFLD mouse model were performed to validate the expression of feature genes. Diagnostic models based on feature genes were constructed and validated at various disease stages of NAFLD, including simple steatosis, NASH, and NASH-HCC. Furthermore, FDFT1 and PGRMC1, vital genes in the cholesterol synthesis pathway, were further screened because of their outstanding value in NASH-HCC. Our study can facilitate the discovery of new diagnostic markers, improve the diagnostic efficacy of NASH-related HCC, and provide new meaningful findings for the prognosis and treatment of NASH-HCC.

Abbreviations

NASH-HCC, Nonalcoholic Steatohepatitis (NASH)-Related Hepatocellular Carcinoma; DEGs, Differentially Expressed Genes; DKGs, DEGs-related Key Genes; MCD, Methionine-Choline Deficient Diet; MCS, Methionine-Choline Sufficient Diet; SHAP Value, SHapley Additive exPlanations (SHAP) Value; AUC, Area Under Curve; ROC, Receiver Operating Characteristic curve; DCA, Decision Curve Analysis.

Data Sharing Statement

This article and its supplementary information files include all crucial data generated or analyzed during this study. Our datasets used during the current study are available and can be downloaded from the GEO and TCGA databases.

Ethics Approval and Consent to Participate

All the animals were acclimated under standard laboratory conditions (ventilated room, 25 ± 1°C, 60 ± 5% humidity, 12 h light/dark cycle) and had free access to standard water and food. Eight-week-old male C57BL/6 J mice were purchased from Zhejiang Vital River Laboratory Animal Technology Co., Ltd. (Zhejiang, China). All procedures were conducted in accordance with the "Guiding Principles in the Care and Use of Animals" (China) and were approved by the Animal Experimental Ethics Committee of Shandong Provincial Hospital (No. 2023-159).

The human data involved in this article was from TCGA and GEO databases. TCGA and GEO belong to public databases. The patients involved in the database have obtained ethical approval. Users can download relevant data for

free for research and publishing relevant articles. The study was approved by the ethical committee of Shandong Provincial Hospital (NO.2023-033). We certify that the study was performed in accordance with the 1964 declaration of HELSINKI and later amendments. Our study is based on open-source data and does not contain any data that can identify individuals, so written informed consent was waived by the institutional review board.

Acknowledgments

We gratefully acknowledge the efforts of all authors who contributed to this article and the GEO database and TCGA database for supporting this study. Also, we acknowledge the funding support from the National Natural Science Foundation of China, the Natural Science Foundation of Shandong Province, and Xinjiang Uygur Autonomous Region.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis, and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was supported by the National Natural Science Foundation of China (NO. 82203658; 82160124; 82100641), Science and Technology Department of Xinjiang Uygur Autonomous Region (NO. 2022E02044), Science and Technology Bureau of Xinjiang Production and Construction Corps (NO. 2022AB024), the Natural Science Foundation of Shandong Province (NO. ZR2021QH276).

Disclosure

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Guo X, Yin X, Liu Z, Wang J. Non-Alcoholic Fatty Liver Disease (NAFLD) pathogenesis and natural products for prevention and treatment. *Int J mol Sci*. 2022;23.
2. Younossi Z, Anstee QM, Marietti M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. *Nat Rev Gastroenterol Hepatol*. 2018;15:11–20. doi:10.1038/nrgastro.2017.109
3. Quek J, Chan KE, Wong ZY, et al. Global prevalence of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in the overweight and obese population: a systematic review and meta-analysis. *Lancet Gastroenterol Hepatol*. 2023;8:20–30. doi:10.1016/S2468-1253(22)00317-X
4. Huang DQ, El-Serag HB, Loomba R. Global epidemiology of NAFLD-related HCC: trends, predictions, risk factors and prevention. *Nat Rev Gastroenterol Hepatol*. 2021;18:223–238. doi:10.1038/s41575-020-00381-6
5. Llovet JM, Willoughby CE, Singal AG, et al. Nonalcoholic steatohepatitis-related hepatocellular carcinoma: pathogenesis and treatment. *Nat Rev Gastroenterol Hepatol*. 2023;20:487–503. doi:10.1038/s41575-023-00754-7
6. Buzzetti E, Pinzani M, Tsochatzis EA. The multiple-hit pathogenesis of non-alcoholic fatty liver disease (NAFLD). *Metabolism*. 2016;65:1038–1048. doi:10.1016/j.metabol.2015.12.012
7. Singal AG, Kanwal F, Llovet JM. Global trends in hepatocellular carcinoma epidemiology: implications for screening, prevention and therapy. *Nat Rev Clin Oncol*. 2023;20:864–884. doi:10.1038/s41571-023-00825-3
8. Chalasani N, Younossi Z, Lavine JE, et al. The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American association for the study of liver diseases. *Hepatology*. 2018;67:328–357. doi:10.1002/hep.29367
9. Caussy C, Reeder SB, Sirlin CB, Loomba R. Noninvasive, quantitative assessment of liver fat by MRI-PDFF as an endpoint in NASH trials. *Hepatology*. 2018;68:763–772. doi:10.1002/hep.29797
10. Li X, Zhao Y, Zhang D, et al. Development of an interpretable machine learning model associated with heavy metals' exposure to identify coronary heart disease among US adults via SHAP: findings of the US NHANES from 2003 to 2018. *Chemosphere*. 2023;311(137039):137039. doi:10.1016/j.chemosphere.2022.137039
11. Cabré N, Luciano-Mateo F, Chapski DJ, et al. Glutaminolysis-induced mTORC1 activation drives non-alcoholic steatohepatitis progression. *J Hepatol*. 2021. doi:10.1016/j.jhep.2021.04.037
12. Rohe HJ, Ahmed IS, Twist KE, Craven RJ. PGRMC1 (progesterone receptor membrane component 1): a targetable protein with multiple functions in steroid signaling, P450 activation and drug binding. *Pharmacol Ther*. 2009;121:14–19. doi:10.1016/j.pharmthera.2008.09.006
13. Asperger H, Stamm N, Gierke B, et al. Progesterone receptor membrane component 1 regulates lipid homeostasis and drives oncogenic signaling resulting in breast cancer progression. *Breast Cancer Res*. 2020;22:75. doi:10.1186/s13058-020-01312-8

14. Lugo H, González-Avella JC, San Miguel M. Local connectivity effects in learning and coordination dynamics in a two-layer network. *Chaos*. 2020;30(083125). doi:10.1063/5.0006908
15. Xu Y, Zhang H. Convergence of deep convolutional neural networks. *Neural Networks*. 2022;153:553–563. doi:10.1016/j.neunet.2022.06.031
16. Bukhari ST, Mohy-Ud-Din H. A systematic evaluation of learning rate policies in training CNNs for brain tumor segmentation. *Phys Med Biol*. 2021;66:105004. doi:10.1088/1361-6560/abe3d3
17. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2:56–67. doi:10.1038/s42256-019-0138-9
18. Byrne CD, Targher G. NAFLD: a multisystem disease. *J Hepatol*. 2015;62:S47–64. doi:10.1016/j.jhep.2014.12.012
19. Chalasani N, Guo X, Loomba R, et al. Genome-wide association study identifies variants associated with histologic features of nonalcoholic fatty liver disease. *Gastroenterology*. 2010;139(1567–1576):1576.e1561–1566. doi:10.1053/j.gastro.2010.07.057
20. Stättermayer AF, Rutter K, Beinhardt S, et al. Role of FDFT1 polymorphism for fibrosis progression in patients with chronic hepatitis C. *Liver Int*. 2014;34:388–395. doi:10.1111/liv.12269
21. Li R, Li Z, Feng Y, et al. PDGFRβ-targeted TRAIL specifically induces apoptosis of activated hepatic stellate cells and ameliorates liver fibrosis. *Apoptosis*. 2020;25:105–119. doi:10.1007/s10495-019-01583-3
22. Pulido P, Leister D. Novel DNAJ-related proteins in Arabidopsis thaliana. *New Phytol*. 2018;217:480–490. doi:10.1111/nph.14827
23. Belyaeva OV, Wu L, Shmarakov I, Nelson PS, Kedishvili NY. Retinol dehydrogenase 11 is essential for the maintenance of retinol homeostasis in liver and testis in mice. *J Biol Chem*. 2018;293:6996–7007. doi:10.1074/jbc.RA117.001646
24. Votava JA, John SV, Li Z, Chen S, Fan J, Parks BW. Mining cholesterol genes from thousands of mouse livers identifies aldolase C as a regulator of cholesterol biosynthesis. *J Lipid Res*. 2024;65(100525):100525. doi:10.1016/j.jlr.2024.100525
25. Lee SR, Kwon SW, Kaya P, et al. Loss of progesterone receptor membrane component 1 promotes hepatic steatosis via the induced de novo lipogenesis. *Sci Rep*. 2018;8:15711. doi:10.1038/s41598-018-34148-6
26. Liu F, Liao Z, Zhang Z. MYC in liver cancer: mechanisms and targeted therapy opportunities. *Oncogene*. 2023;42:3303–3318. doi:10.1038/s41388-023-02861-w
27. Yang F, Kou J, Liu Z, Li W, Du W. MYC enhances cholesterol biosynthesis and supports cell proliferation through SQLE. *Front Cell Dev Biol*. 2021;9:655889.
28. Cheng W, Li M, Zhang L, et al. New roles of N6-methyladenosine methylation system regulating the occurrence of non-alcoholic fatty liver disease with N6-methyladenosine-modified MYC. *Front Pharmacol*. 2022;13(973116). doi:10.3389/fphar.2022.973116
29. Dong X, Zhu Y, Wang S, et al. Bavachinin inhibits cholesterol synthesis enzyme FDFT1 expression via AKT/mTOR/SREBP-2 pathway. *Int Immunopharmacol*. 2020;88(106865):106865. doi:10.1016/j.intimp.2020.106865
30. Ouyang G, Wu Z, Liu Z, et al. Identification and validation of potential diagnostic signature and immune cell infiltration for NAFLD based on cuproptosis-related genes by bioinformatics analysis and machine learning. *Front Immunol*. 2023;14:1251750.
31. Zhang Z, Wang S, Zhu Z, Nie B. Identification of potential feature genes in non-alcoholic fatty liver disease using bioinformatics analysis and machine learning strategies. *Comput Biol Med*. 2023;157:106724.
32. Sanyal AJ, Castera L, Wong VW. Noninvasive Assessment of Liver Fibrosis in NAFLD. *Clin Gastroenterol Hepatol*. 2023;21:2026–2039. doi:10.1016/j.cgh.2023.03.042
33. Masoodi M, Gastaldelli A, Hyötyläinen T, et al. Metabolomics and lipidomics in NAFLD: biomarkers and non-invasive diagnostic tests. *Nat Rev Gastroenterol Hepatol*. 2021;18:835–856. doi:10.1038/s41575-021-00502-9
34. Di Mauro S, Scamporrino A, Filippello A, et al. Clinical and molecular biomarkers for diagnosis and staging of NAFLD. *Int J mol Sci*. 2021;22:11905. doi:10.3390/ijms222111905
35. Bedogni G, Bellentani S, Miglioli L, et al. The fatty liver index: a simple and accurate predictor of hepatic steatosis in the general population. *BMC Gastroenterol*. 2006;6(33). doi:10.1186/1471-230X-6-33
36. Lee JH, Kim D, Kim HJ, et al. Hepatic steatosis index: a simple screening tool reflecting nonalcoholic fatty liver disease. *Dig Liver Dis*. 2010;42:503–508. doi:10.1016/j.dld.2009.08.002
37. Skuratovskaia D, Vulf M, Khaziakhmatova O, et al. Tissue-specific role of macrophages in noninfectious inflammatory disorders. *Biomedicines*. 2020;8:400. doi:10.3390/biomedicines8100400
38. Tacke F, Zimmermann HW. Macrophage heterogeneity in liver injury and fibrosis. *J Hepatol*. 2014;60:1090–1096. doi:10.1016/j.jhep.2013.12.025
39. Wang Y, Huang S, Kong W, et al. Corilagin alleviates liver fibrosis in zebrafish and mice by repressing IDO1-mediated M2 macrophage repolarization. *Phytomedicine*. 2023;119(155016):155016. doi:10.1016/j.phymed.2023.155016
40. Wang X, Sun R, Wei H, Tian Z. High-mobility group box 1 (HMGB1)-Toll-like receptor (TLR)4-interleukin (IL)-23-IL-17A axis in drug-induced damage-associated lethal hepatitis: interaction of γδ T cells with macrophages. *Hepatology*. 2013;57:373–384. doi:10.1002/hep.25982
41. Cai D, Zhong GC, Dai X, et al. Targeting FDFT1 reduces cholesterol and bile acid production and delays hepatocellular carcinoma progression through the HNF4A/ALDOB/AKT1 axis. *Adv Sci*. 2025;e2411719.
42. Xia P, Wen GM, Zheng XH, Zhao ZY. Differences of ferroptosis-related genes between white and Asian patients with liver cancer. *Am J Cancer Res*. 2023;13:3659–3667.
43. Ito Y, Takeda T, Sakon M, et al. Expression and clinical significance of erb-B receptor family in hepatocellular carcinoma. *Br J Cancer*. 2001;84:1377–1383. doi:10.1054/bjoc.2000.1580
44. Lee SR, Lee JG, Heo JH, et al. Loss of PGRMC1 delays the progression of hepatocellular carcinoma via suppression of pro-inflammatory immune responses. *Cancers*. 2021;13.
45. Ji F, Zhang J, Mao L, et al. Liver-specific gene PGRMC1 blocks c-Myc-induced hepatocarcinogenesis through ER stress-independent PERK activation. *Nat Commun*. 2025;16:50. doi:10.1038/s41467-024-55745-2
46. Weng ML, Chen WK, Chen XY, et al. Fasting inhibits aerobic glycolysis and proliferation in colorectal cancer via the Fdft1-mediated AKT/mTOR/HIF1α pathway suppression. *Nat Commun*. 2020;11:1869. doi:10.1038/s41467-020-15795-8
47. Cao T, Chen Q, Zhang B, et al. Clozapine induced disturbances in hepatic glucose metabolism: the potential role of PGRMC1 signaling. *Front Endocrinol*. 2021;12:727371. doi:10.3389/fendo.2021.727371
48. van der Meer DL, Degenhardt T, Väisänen S, et al. Profiling of promoter occupancy by PPARalpha in human hepatoma cells via ChIP-chip analysis. *Nucleic Acids Res*. 2010;38:2839–2850. doi:10.1093/nar/gkq012