Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Research on the Cholesterol Metabolism-Related Genes in the Progression of Metabolic Fatty Liver Disease

Hao Chen

xingxingba1@yeah.net

Huadu District People's Hospital of Guangzhou

Liang Lin

Huadu District People's Hospital of Guangzhou

Xubin Qiu

Huadu District People's Hospital of Guangzhou

Laibin Zou

Huadu District People's Hospital of Guangzhou

Research Article

Keywords: Metabolic fatty liver disease, Cholesterol metabolism, Lipid deposition, INHBE, LGALS4, PPP1R3C

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11389621/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Objective

Cholesterol metabolic disturbance accompanies metabolic fatty liver disease (MAFLD) progression. This study aimed to identify cholesterol metabolism-related core genes and explore their roles in MAFLD.

Methods

Four liver microarray datasets were analyzed, with GSE135251 as the training set and GSE48452, GSE89632, and GSE126848 as validation sets. ssGSEA quantified cholesterol metabolic activity. Differential expression and WGCNA identified candidate genes, and LASSO, SVM-RFE, and XGBoost screened core biomarkers. Candidate drugs were predicted via DSigDB and validated by molecular docking. A high-fat diet (HFD) mouse model was used for in vivo verification.

Results

MAFLD samples showed elevated cholesterol metabolic activity ($P < 0.001$). WGCNA and differential analysis yielded 274 candidate genes enriched in lipid/steroid metabolism, inflammation, and detoxification. From 93 PPI hub genes, the three algorithms identified INHBE, LGALS4, and PPP1R3C, with AUCs of 0.940, 0.921, and 0.939 in the training set and good diagnostic performance in validation sets. Co-Dergocrine Mesilate, Etoposide, and Trichostatin A stably bound these targets. In HFD mice, hepatic expression of all three genes was upregulated, with elevated cholesterol, triglycerides, and marked lipid accumulation.

Conclusion

INHBE, LGALS4, and PPP1R3C are potential diagnostic biomarkers and therapeutic targets for MAFLD, whose progression is driven by cholesterol disturbance-mediated lipid deposition and inflammation.

1. Introduction

Metabolic fatty liver disease (MAFLD) is the most prevalent chronic metabolic liver disorder worldwide, with excessive and abnormal lipid deposition in hepatocytes as its core pathological feature. Its disease spectrum ranges from simple steatosis to steatohepatitis, liver fibrosis, and cirrhosis, and in severe cases it can progress to hepatocellular carcinoma, thereby posing a serious threat to human health [1–3]. Compared with the traditional concept of non-alcoholic fatty liver disease, MAFLD places greater emphasis on metabolic disturbance as the central pathogenic driver and is closely associated with metabolic abnormalities such as obesity, dyslipidemia, and insulin resistance [4, 5]. In recent years, driven

by high-fat dietary patterns and increasingly sedentary lifestyles, the global prevalence of MAFLD has continued to rise, exceeding 25% in the general population, with a noticeable trend toward younger onset [6, 7]. Clinical studies have confirmed that simple hepatic steatosis can progressively evolve into steatohepatitis in the absence of obvious symptoms, accompanied by hepatocyte inflammation, necrosis, and hepatic fibrosis, making it a major risk factor for the progression of chronic liver disease, liver failure, and hepatocarcinogenesis [8, 9].

The liver is the central organ responsible for cholesterol synthesis, metabolism, and transport. Maintaining cholesterol homeostasis is essential for preserving normal hepatic physiological function [10]. Cholesterol metabolism mainly comprises four core processes: endogenous synthesis, exogenous uptake, bile acid conversion, and lipid transport. Precise regulation of the genes involved in these processes is critical for maintaining lipid metabolic homeostasis [9]. Under normal physiological conditions, the liver avoids abnormal lipid accumulation by regulating rate-limiting enzymes of cholesterol synthesis, lipid-transport receptors, and metabolic pathway genes. When metabolic disturbances occur, however, excessive cholesterol synthesis, abnormal increases in uptake, and impaired metabolic clearance lead to intracellular cholesterol overload in hepatocytes and trigger lipotoxic injury [9, 10].

Accumulating evidence indicates that cholesterol metabolic dysfunction is among the initiating factors driving the occurrence and progression of MAFLD [11]. Accumulated free cholesterol disrupts hepatic mitochondrial structure and function, induces oxidative stress, promotes massive generation of reactive oxygen species, and subsequently activates inflammatory pathways such as NF-κB, leading to inflammatory infiltration, apoptosis, and necrosis of hepatocytes [12, 13]. At the same time, cholesterol imbalance exacerbates hepatic lipid deposition and creates a vicious cycle of “lipid metabolic disorder–lipid accumulation–inflammatory injury–fibrotic progression,” pushing simple steatosis toward severe fatty liver disease [14, 15]. Clinical epidemiological studies show that elevated serum total cholesterol and low-density lipoprotein cholesterol levels are significantly positively correlated with MAFLD risk, and that the prevalence of MAFLD is markedly higher in patients with hypercholesterolemia [16, 17]. Other studies have further shown that aberrant activation of hepatic cholesterol metabolic pathways markedly aggravates insulin resistance, worsens systemic metabolic disturbance, and accelerates disease progression [18, 19].

Existing metabolomic studies have established that cholesterol metabolic disturbance is one of the most core metabolic abnormalities in MAFLD and can serve as an important metabolic signature for disease diagnosis and severity assessment [20, 21]. However, current research largely focuses on differential analysis at the metabolite level, and systematic investigations into the expression patterns, regulatory networks, and clinical translational value of cholesterol metabolism-related regulatory genes during MAFLD progression remain lacking. Stable and reproducible core gene biomarkers for early MAFLD diagnosis are still scarce, and the discovery of candidate drugs targeting cholesterol metabolism-related genes remains insufficient.

To address this gap, the present study leveraged multiple human liver-tissue microarray datasets from GEO and integrated ssGSEA-based cholesterol metabolism scoring, WGCNA, and differential expression analysis to construct a candidate gene set related to cholesterol metabolism in MAFLD. LASSO, SVM-RFE, and XGBoost machine-learning algorithms were used to screen core biomarkers, whose diagnostic performance was validated in independent external datasets. Potential targeting drugs were predicted using the DSigDB database, and drug–target protein binding was evaluated by molecular docking. In addition, a high-fat diet-induced mouse MAFLD model was established to validate the in vivo expression changes of the core genes *INHBE*, *LGALS4*, and *PPP1R3C*, as well as hepatic lipid phenotypes. This study aims to elucidate the molecular mechanisms by which cholesterol metabolism-related genes regulate MAFLD progression, and to provide new theoretical support for early diagnosis, mechanistic research, and clinical targeted intervention in MAFLD

2. Methods

2.1 Data Sources and Preprocessing

Four MAFLD-related human liver tissue microarray datasets were downloaded from the Gene Expression Omnibus (GEO) database: GSE135251, GSE48452, GSE89632, and GSE126848. The GSE135251 dataset contained 206 MAFLD patient samples and 10 normal liver tissue control samples; GSE48452 comprised 18 MAFLD samples and 14 control samples; GSE89632 comprised 19 MAFLD samples and 24 control samples; and GSE126848 comprised 31 MAFLD samples and 14 control samples. GSE135251 was used as the training set, whereas GSE48452, GSE89632, and GSE126848 served as external validation sets. In addition, 50 cholesterol metabolism-related genes (CMRGs) were compiled from previous literature for subsequent metabolic score quantification and gene screening analyses [22].

2.2 Quantification of Cholesterol Metabolism Activity Scores

Single-sample gene set enrichment analysis (ssGSEA) was performed to quantify cholesterol metabolism activity in each sample. Based on the set of 50 CMRGs, the "ssgsea" function in the R package GSVA was used to calculate the cholesterol metabolism activity enrichment score for each sample in the training set, thereby precisely reflecting the activity level of cholesterol metabolism at the single-sample level. The Wilcoxon rank-sum test was used to compare metabolic scores between the MAFLD group and normal control group, so as to define the overall dysregulation characteristics of cholesterol metabolism under disease conditions.

2.3 Weighted Gene Co-Expression Network Analysis (WGCNA)

The R package WGCNA (version 1.72) was used to construct a gene co-expression network and to identify core gene modules highly associated with cholesterol metabolism activity and MAFLD disease phenotype. First, a soft-thresholding power was set to ensure that the network conformed to scale-free

topology (squared correlation coefficient threshold = 0.90). Genes were assigned to distinct co-expression modules using hierarchical clustering combined with the dynamic tree-cutting algorithm. Module eigengenes were calculated, and their correlations with the cholesterol metabolism activity score and disease phenotype were evaluated. Modules meeting $|r| > 0.3$ and $P < 0.05$ were selected as significant core modules, and all genes within these modules were extracted as candidate co-expressed genes.

2.4 Screening of MAFLD-Associated Cholesterol Metabolism-Related Differential Genes

Differentially expressed genes (DEGs) between MAFLD and control samples in the training set were identified using the R package limma, with screening thresholds of $P < 0.05$ and $|\log2FC| > 1.0$. The intersection between WGCNA core module genes and DEGs was obtained by Venn analysis to yield MAFLD-related candidate cholesterol metabolism genes. Gene Ontology (GO) functional enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed using the clusterProfiler package (version 4.4.4) to clarify the functional mechanisms and signaling pathways involved. A protein–protein interaction (PPI) network of the candidate genes was constructed based on the STRING database (organism: *Homo sapiens*; confidence score > 0.4), and hub interacting genes within the network were extracted for subsequent analysis.

2.5 Machine Learning-Based Screening of Core Feature Genes

To identify optimal MAFLD cholesterol metabolism-related feature genes, this study integrated 12 classical machine learning algorithms to construct diagnostic prediction models. Feature selection and model training were completed through 113 algorithm combinations to obtain a subset of candidate genes. Subsequently, three machine learning algorithms were integrated to screen core feature biomarkers of MAFLD from these candidate genes. (i) LASSO regression: the R package glmnet was applied with 10-fold cross-validation and 1,000 iterations, and key feature genes were selected based on the minimum criterion of lambda.min, thereby eliminating gene collinearity and overfitting. (ii) SVM-RFE: through 5-fold cross-validation, low-weight features were progressively eliminated to identify the feature gene combination with the highest model accuracy and lowest error rate. (iii) XGBoost: the SHAPforxgboost package was used to calculate the contribution of each gene to model prediction, and core genes with high contribution were ranked and selected. Finally, the overlapping genes commonly identified by all three algorithms were regarded as potential core biomarkers for MAFLD.

2.6 Molecular Docking Validation of Drug–Target Binding

Approved drugs capable of targeting and regulating the core biomarkers were retrieved from the DSigDB database (https://dsigdb.tanlab.org/) to screen potential therapeutic agents. Molecular docking analysis was performed to validate the binding activity between candidate drugs and core targets. Standard accession numbers of the core target proteins were obtained from the UniProt database, and crystal

structures of the target proteins (PDB format) were downloaded from the RCSB PDB database (https://www.rcsb.org/); the 3D molecular structures of candidate drugs (SDF format) were downloaded from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/). The online molecular docking tool CB-Dock2 (http://clab.labshare.cn/cb-dock/) was used to perform blind docking simulations of the drugs against the target proteins, and binding sites, binding energies, and hydrogen-bond interactions were analyzed to verify drug–target binding affinity.

2.7 Animals and Experimental Procedures

All animal experiments were conducted in accordance with the ARRIVE guidelines and were approved by the Institutional Animal Care and Use Committee of Guangzhou Jennio Biotech Co., Ltd. (Approval No.: JENNIO-IACUC-2026-A056). Twelve male C57BL/6J mice (6 weeks old, weighing 18–22 g) were obtained from the Experimental Animal Center of Southern Medical University and housed in a temperature-controlled environment (20–24°C) under a 12 h:12 h light–dark cycle, with free access to food and water. The mice were randomly divided into twogroups (n = 6 per group): (1) Control group: fed a standard chow diet (Jiangsu Xietong Biotechnology Co., Ltd.) for 6 weeks; (2) Low-fat high-cholesterol group: fed a diet containing 15% fat, 1.25% cholesterol, and 0.5% cholic acid (Jiangsu Xietong Biotechnology Co., Ltd.) for 6 weeks.

2.8 Hematoxylin and Eosin (H&E) Staining

Liver specimens were fixed in 4% paraformaldehyde and embedded in paraffin. Paraffin blocks were cut into 5-μm-thick sections, which were then dewaxed and rehydrated. The sections were stained with H&E solution (Sigma-Aldrich, St. Louis, MO, USA) for 10 min. Pathological changes in the liver were evaluated by observing five different fields per section.

2.9 Oil Red O Staining

Liver specimens were collected and cut into frozen sections (5 μm), which were then stained with Oil Red O solution (Solarbio Science & Technology Co., Ltd., Beijing, China) according to the manufacturer's instructions. The percentage of steatotic hepatocytes was assessed based on relative lipid content, and quantitative analysis was performed using ImageJ software.

2.10 Statistical Analysis

All bioinformatics analyses were performed using R software (version 4.4.3). Comparisons between two groups were conducted using the unpaired *t*-test, and $P < 0.05$ was considered statistically significant.

3. Results

3.1 Cholesterol Metabolism Activity Is Significantly Dysregulated in MAFLD Patients

The GSE135251 dataset was used to construct the training set, and GSE48452, GSE89632, and GSE126848 served as validation sets. ssGSEA quantification showed that cholesterol metabolism activity scores were significantly higher in MAFLD samples than in normal controls ($P < 0.001$), confirming that MAFLD progression is characterized by marked overactivation of cholesterol metabolism and metabolic imbalance, thereby laying the foundation for subsequent gene screening and mechanistic investigation (Fig. 1A–D). Moreover, ROC analysis indicated that the overall cholesterol metabolism activity score had high diagnostic value for MAFLD (Fig. 2A–D).

3.2 Screening and Functional Enrichment of MAFLD-Associated Cholesterol Metabolism Candidate Genes

WGCNA showed that when the soft-thresholding power was set to 7, the network met the scale-free topology criterion ($r^2 = 0.90$), and 11 gene co-expression modules were ultimately identified (Fig. 3A–B). Among them, six modules were significantly correlated with the cholesterol metabolism activity score and MAFLD disease phenotype: the tan module ($r = -0.509$, $P < 0.001$), black module ($r = -0.490$, $P < 0.001$), salmon module ($r = -0.224$, $P < 0.001$), red module ($r = 0.658$, $P < 0.001$), brown module ($r = 0.761$, $P < 0.001$), and greenyellow module ($r = 0.654$, $P < 0.001$), collectively comprising 4,191 module genes (Fig. 3C). Differential expression analysis identified 471 MAFLD-related DEGs, including 247 upregulated and 224 downregulated genes (Fig. 3D). Venn analysis of the intersection yielded 274 cholesterol metabolism-related candidate dysregulated genes for MAFLD (Fig. 3E).

GO functional enrichment analysis showed that the candidate genes were significantly enriched across biological process (BP), cellular component (CC), and molecular function (MF) categories (Fig. 4A). In the BP category, the genes were predominantly enriched in heavy metal ion stress response and detoxification events, including cellular responses to cadmium, copper, and zinc ions, inorganic compound detoxification, and maintenance of intracellular zinc ion homeostasis; they were also significantly enriched in lipid and steroid metabolic processes, covering alcohol metabolism, cholesterol metabolism, steroid biosynthesis, and organic hydroxy compound synthesis, as well as in responses to toxic substances and interleukin-1-mediated inflammatory responses. CC terms mainly involved subcellular structures such as the RISC complex, RNAi effector complex, RNA polymerase II transcription regulator complex, late endosome, secretory vesicle lumen, and peroxisome. MF terms included phospholipid binding, glycolipid binding, cytokine activity, DNA-binding transcription activator activity, and amyloid-beta binding. KEGG pathway enrichment analysis indicated that the candidate genes were mainly involved in lipid and steroid metabolism, heavy metal ion toxicity response, xenobiotic detoxification, and cellular inflammatory response pathways. Collectively, these results suggest that the candidate genes may participate in the onset and progression of the studied phenotype by regulating metal ion detoxification, lipid metabolic reprogramming, and inflammatory responses (Fig. 4B). PPI network analysis showed that 93 of the 274 candidate genes exhibited significant protein–protein interactions and formed a core regulatory network; these genes were included in subsequent machine learning screening (Fig. 4C).

3.3 Machine Learning-Based Screening of MAFLD Core Biomarkers

Machine learning screening was performed based on the 93 core interacting genes. Using 12 machine learning algorithms, 113 prediction models were constructed, ultimately yielding 75 candidate genes (Fig. 4D). Further analysis showed that LASSO regression identified 14 feature genes, with an optimal lambda value of 0.0213 (Fig. 5A–B); the SVM-RFE model identified 7 core feature genes, at which point the model achieved the lowest error rate and highest accuracy (0.844) (Fig. 5B–C); and the XGBoost model ranked genes by contribution and selected the top 10 high-contribution feature genes (Fig. 5E). The overlapping genes identified by all three algorithms yielded three final candidate core genes: *INHBE*, *LGALS4*, and *PPP1R3C* (Fig. 5F).

3.4 *INHBE*, *LGALS4*, and *PPP1R3C* Are Identified as Potential MAFLD Biomarkers

ROC curve analysis showed that in the training set, the AUC values were 0.940 for *INHBE*, 0.921 for *LGALS4*, and 0.939 for *PPP1R3C* (Fig. 6A); in the validation sets, the mean AUC values were 0.894 for *INHBE*, 0.847 for *LGALS4*, and 0.815 for *PPP1R3C*, indicating that all three genes have good diagnostic and predictive performance for MAFLD (Fig. 6B–D). Expression analysis in the training set showed that *INHBE*, *LGALS4*, and *PPP1R3C* were all highly expressed in MAFLD samples (Fig. 6E). In the validation sets, *INHBE*, *LGALS4*, and *PPP1R3C* were significantly upregulated in the MAFLD groups of GSE48452, GSE89632, and GSE126848, indicating that the validation results were fully consistent with the training set and confirming the stable high expression of the three genes in MAFLD (Fig. 6F–H).

3.6 Subcellular Localization of Core Genes, Targeted Drug Prediction, and Molecular Docking Validation

Using the DSigDB database, drug prediction identified 15 potential drugs targeting INHBE, 10 targeting LGALS4, and 15 targeting PPP1R3C (Fig. 7A). Molecular docking showed that Co-Dergocrine Mesilate bound stably to the active pocket of the INHBE protein, with an optimal binding free energy of − 9.6 kcal/mol, and interacted with key amino acid residues of INHBE through hydrogen bonds, indicating excellent binding affinity (Fig. 7B–C). Etoposide bound to the optimal site of the LGALS4 protein with a binding free energy of − 8.2 kcal/mol, stably occupying the active region of LGALS4 and effectively modulating its protein activity (Fig. 7D–E). Trichostatin A bound to the optimal site of the PPP1R3C protein with a binding free energy of − 7.1 kcal/mol, stably occupying the active region of PPP1R3C and effectively modulating its protein activity (Fig. 7F–G). These findings confirm that the three drugs possess a molecular basis for targeted intervention in MAFLD.

3.7 In Vivo Validation in a Mouse Model

In the high-fat diet-induced mouse MAFLD model, gross anatomical examination revealed that the liver tissue of HFD-fed mice appeared pale and whitish, and abundant gallstones were observed in the gallbladder, whereas none of these phenomena were present in the control group (Fig. 8A). HE staining showed obvious fatty vacuoles and lipid accumulation in the liver tissue of model mice (Fig. 8B), and Oil Red O staining likewise confirmed lipid accumulation, consistent with the pathological features of MAFLD (Fig. 8C). IHC detection confirmed that the gene and protein expression levels of *INHBE*, *LGALS4*, and *PPP1R3C* in the liver tissue of the model group were significantly higher than those in the normal control group, fully consistent with the bioinformatics analysis results (Fig. 8D).

4. Discussion

The core pathogenesis of MAFLD lies in the imbalance of hepatic metabolic homeostasis, among which cholesterol metabolic disturbance serves as a key initiating factor driving disease progression from simple lipid deposition to inflammation and fibrosis [9, 23]. Previous studies have mostly focused on the lipidomic and metabolomic differences of MAFLD and have clarified abnormal changes in cholesterol metabolites [24]; however, systematic multi-dataset studies on the mechanistic roles and clinical translational value of core cholesterol-metabolism-regulating genes during MAFLD progression remain limited. In this study, by integrating multiple GEO datasets and combining ssGSEA-based metabolic scoring, WGCNA, differential gene analysis, multi-machine-learning screening, molecular docking, and in vivo validation in high-fat-diet mice, we systematically investigated the regulatory roles of cholesterol metabolism-related genes in MAFLD and identified *INHBE*, *LGALS4*, and *PPP1R3C* as potential core biomarkers and therapeutic targets for MAFLD.

Our ssGSEA analysis confirmed that cholesterol metabolism activity scores were significantly elevated in the liver tissues of MAFLD patients, suggesting abnormal activation of the overall hepatic cholesterol metabolic pathway and imbalance of lipid homeostasis, ultimately leading to hepatocellular lipid accumulation. Combined WGCNA and differential expression screening yielded 274 candidate genes with dysregulated cholesterol metabolism. Functional enrichment showed that, in addition to participating in classical MAFLD pathological processes such as lipid deposition, steroid biosynthesis, and inflammatory response, these candidate genes were also significantly enriched in biological events including heavy metal ion stress and xenobiotic detoxification. This indicates that cholesterol metabolic disturbance is not an isolated event; rather, it can act synergistically with cellular detoxification and oxidative stress pathways to jointly promote MAFLD progression. Excessive cholesterol accumulation damages hepatocyte mitochondria, promotes ROS release, activates downstream inflammatory signaling pathways, and amplifies hepatic inflammatory injury, thereby forming a vicious cycle of “cholesterol metabolic disturbance–lipotoxic injury–inflammatory activation–disease progression” [14, 25].

Based on the PPI network, 93 core genes with protein–protein interactions were obtained. Further joint screening with three machine learning algorithms—LASSO, SVM-RFE, and XGBoost—followed by intersection analysis, identified three core genes: *INHBE*, *LGALS4*, and *PPP1R3C*. Multi-dataset ROC

validation showed that all three genes exhibited good diagnostic performance for MAFLD in both the training set and three independent external validation sets, and were stably upregulated in MAFLD samples, suggesting that they may serve as potential biomarkers for the early diagnosis of MAFLD.

INHBE, an inhibin βE subunit, participates in the regulation of metabolic homeostasis, cell proliferation, and inflammatory responses; existing studies suggest that it is involved in hepatic metabolic disturbance and steatosis [26]. *LGALS4*, a member of the galectin family, mediates cell adhesion, inflammatory signaling, and lipid-related signal regulation; it is upregulated in chronic liver diseases and contributes to shaping the hepatic inflammatory microenvironment [27, 28]. *PPP1R3C*, as a regulatory subunit of protein phosphatase 1, is involved in glycogen and lipid metabolism regulation, and its increased expression can promote lipid accumulation in hepatocytes [29, 30]. Our in vivo IHC results confirmed that *INHBE*, *LGALS4*, and *PPP1R3C* protein expression was significantly upregulated in the liver tissue of high-fat-diet-induced MAFLD mice, accompanied by elevated hepatic cholesterol and triglycerides and obvious steatosis, further supporting the roles of these three genes in MAFLD pathogenesis.

To explore potential clinical translational value, candidate drugs targeting these three core genes were predicted using the DSigDB database, and molecular docking was performed. The docking results showed that Co-Dergocrine Mesilate, Etoposide, and Trichostatin A could stably bind the proteins *INHBE*, *LGALS4*, and *PPP1R3C*, respectively, with relatively low binding free energies, indicating molecular potential for targeted regulation of these targets and providing candidate molecular clues for MAFLD targeted drug development.

This study still has several limitations. First, it is fundamentally a bioinformatics study combined with an animal model; the diagnostic value of *INHBE*, *LGALS4*, and *PPP1R3C* in human MAFLD cohorts, as well as the in vivo therapeutic effects of the candidate drugs, still require validation by large-sample clinical cohorts and additional animal intervention experiments. Second, we only verified the upregulation of these three core genes in MAFLD liver tissue; the specific upstream and downstream molecular mechanisms by which *INHBE*, *LGALS4*, and *PPP1R3C* regulate hepatic lipid deposition and inflammatory injury still need to be elucidated by in vitro cellular functional experiments. Finally, the mutual regulatory network between the cholesterol metabolic pathway and these three core genes, as well as the mechanism by which multiple genes cooperatively promote MAFLD progression, warrant further investigation.

In summary, this study systematically demonstrates that dysregulation of cholesterol metabolism-related genes is an important driver of MAFLD progression. *INHBE*, *LGALS4*, and *PPP1R3C* can serve as novel biomarkers for early MAFLD diagnosis and as potential targets for precision therapy. Cholesterol metabolic disturbance promotes MAFLD progression by regulating lipid deposition and inflammatory pathways. The findings of this study provide a theoretical and experimental basis for mechanistic dissection, early screening, and development of targeted intervention strategies in metabolic fatty liver disease.

Abbreviations

AUC
area under the (ROC) curve
BP
biological process
CB-Dock2
cavity-detection guided blind docking, version 2
CC
cellular component
CMRGs
cholesterol metabolism-related genes
CV
cross-validation
DEGs
differentially expressed genes
FC
fold change
GEO
Gene Expression Omnibus
GO
Gene Ontology
GSEA
gene set enrichment analysis
GSVA
gene set variation analysis
H&E
hematoxylin and eosin
HFD
high-fat diet
IHC
immunohistochemistry
KEGG
Kyoto Encyclopedia of Genes and Genomes
LASSO
least absolute shrinkage and selection operator
log2FC
$\log_2$ (base-2) fold change
MAFLD
metabolic dysfunction-associated fatty liver disease

PPI
protein–protein interaction
PDB
Protein Data Bank
ROC
receiver operating characteristic
ROS
reactive oxygen species
SHAP
SHapley Additive exPlanations
ssGSEA
single-sample gene set enrichment analysis
STRING
Search Tool for the Retrieval of Interacting Genes/Proteins
SVM-RFE
support vector machine–recursive feature elimination
WGCNA
weighted gene co-expression network analysis
XGBoost
extreme gradient boosting
INHBE
inhibin subunit beta E
LGALS4
galectin 4
PPP1R3C
protein phosphatase 1 regulatory subunit 3C

Declarations

Ethics approval:

All animal experiments were conducted in accordance with the ARRIVE guidelines and were approved by the Institutional Animal Care and Use Committee of Guangzhou Jennio Biotech Co., Ltd. (Approval No.: JENNIO-IACUC-2026-A056).

Consent for publication:

Not applicable, this manuscript does not include any details, images, or videos relating to an individual person.

Competing interests:

The authors declare that they have no competing interests.

Funding:

Huadu District Basic and Applied Basic Research Joint Funding Project (District-Hospital Collaboration), Guangzhou (Project No.: 25HDQYLH02).

Author contributions:

(1) Investigation, data curation, formal analysis, validation, writing-original draft, methodology and supervision: Hao Chen. (2) Software and visualization: Xubin Qiu and Xubin Qiu. (3) Conceptualization, project administration, resources and writing-review & editing: Hao Chen and Laibin Zou. All authors approved the final version of manuscript.

Data availability:

All data generated or analyzed during this study are included in this published

article and its supplementary information flies.

References

1. .: 1, Younossi ZM, Golabi P, Paik JM et al (2023) The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review. Hepatology, 77(4): 1335–1347. https://doi.10.1097/hep.0000000000000004. PMC10026948
2. Quek 2, Chan JKE, Wong ZY et al (2023) 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, 8(1): 20–30. https://doi.10.1016/s2468-1253(22)00317-x
3. He 3, Wu JY, Lin Z et al (2025) The evolution of artificial intelligence technology in non-alcoholic fatty liver disease. Front Radiol, 5: 1634165. https://doi.10.3389/fradi.2025.1634165. PMC12480972
4. .: 4, Eslam M, Newsome PN, Sarin SK et al (2020) A new definition for metabolic dysfunction-associated fatty liver disease: An international expert consensus statement. J Hepatol, 73(1): 202–209. https://doi.10.1016/j.jhep.2020.03.039
5. .: 5, Eslam M, Sanyal AJ, George J (2020) MAFLD: A Consensus-Driven Proposed Nomenclature for Metabolic Associated Fatty Liver Disease. Gastroenterology, 158(7): 1999–2014.e1. https://doi.10.1053/j.gastro.2019.11.312

6. Crane 6, Eslick HGD, Gofton C et al (2024) Global prevalence of metabolic dysfunction-associated fatty liver disease-related hepatocellular carcinoma: A systematic review and meta-analysis. Clin Mol Hepatol, 30(3): 436–448. https://doi.10.3350/cmh.2024.0109. PMC11261220
7. : Goyal NP, Xanthakos S, Schwimmer JB (2025) Metabolic dysfunction-associated steatotic liver disease in children. Gut, 74(4): 669–677. https://doi.10.1136/gutjnl-2023-331090
8. .: 8, Rinella ME, Lazarus JV, Ratziu V et al (2023) A multisociety Delphi consensus statement on new fatty liver disease nomenclature. J Hepatol, 79(6): 1542–1556. https://doi.10.1016/j.jhep.2023.06.003
9. Feng 9, Zhang XR, Yang Z et al (2024) Mechanism of Metabolic Dysfunction-associated Steatotic Liver Disease: Important role of lipid metabolism. J Clin Transl Hepatol, 12(9): 815–826. https://doi.10.14218/jcth.2024.00019. PMC11393839
10. : Li XM Li. Unlocking Cholesterol Metabolism in Metabolic-Associated Steatotic Liver Disease: Molecular Targets and Natural Product Interventions. Pharmaceuticals (Basel), (2024) 17(8). https://doi.10.3390/ph17081073. PMC11358954
11. .: 11, Syed-Abdul MM (2023) Lipid Metabolism in Metabolic-Associated Steatotic Liver Disease (MASLD). Metabolites, 14(1). https://doi.10.3390/metabo14010012. PMC10818604
12. Xu 12, Pang XY, Fan X (2025) Mitochondria in oxidative stress, inflammation and aging: from mechanisms to therapeutic advances. Signal Transduct Target Ther, 10(1): 190. https://doi.10.1038/s41392-025-02253-4. PMC12159213
13. Wang 13, Tanaka XN, Hu X et al (2019) A high-cholesterol diet promotes steatohepatitis and liver tumorigenesis in HCV core gene transgenic mice. Arch Toxicol, 93(6): 1713–1725. https://doi.10.1007/s00204-019-02440-7
14. .: 14, Horn CL, Morales AL, Savard C et al (2022) Role of Cholesterol-Associated Steatohepatitis in the Development of NASH. Hepatol Commun, 6(1): 12–35. https://doi.10.1002/hep4.1801. PMC8710790
15. .: 15, Ho CM, Ho SL, Jeng YM et al (2019) Accumulation of free cholesterol and oxidized low-density lipoprotein is associated with portal inflammation and fibrosis in nonalcoholic fatty liver disease. J Inflamm (Lond), 16: 7. https://doi.10.1186/s12950-019-0211-5. PMC6444889
16. Huang 16, Guo HY, Liu Z et al (2022) Remnant Cholesterol Predicts Long-term Mortality of Patients With Metabolic Dysfunction-associated Fatty Liver Disease. J Clin Endocrinol Metab, 107(8): e3295-e3303. https://doi.10.1210/clinem/dgac283
17. .: 17, Hirano T, Satoh N, Ito Y (2024) Specific Increase in Small Dense Low-Density Lipoprotein-Cholesterol Levels beyond Triglycerides in Patients with Diabetes: Implications for Cardiovascular Risk of MAFLD. J Atheroscler Thromb, 31(1): 36–47. https://doi.10.5551/jat.64271. PMC10776337
18. .: 18, Cao X, Wang N, Yang M et al (2025) Lipid Accumulation and Insulin Resistance: Bridging Metabolic Dysfunction-Associated Fatty Liver Disease and Chronic Kidney Disease. Int J Mol Sci,

26(14). https://doi.10.3390/ijms26146962.
PMC12295539

19. Kang 19, Song MJ, Kang ES et al (2025) Pathophysiology, development, and mortality of major non-communicable diseases in metabolic dysfunction-associated steatotic liver disease: A comprehensive review. Int J Biol Sci, 21(13): 5691–5703. https://doi.10.7150/ijbs.117211. PMC12509686
20. Popov 20, Idle JR (2024) Metabolomic Hallmarks of Obesity and Metabolic Dysfunction-Associated Steatotic Liver Disease. Int J Mol Sci, 25(23). https://doi.10.3390/ijms252312809. PMC11640922
21. : McGlinchey AJ, Govaere O, Geng D et al (2022) Metabolic signatures across the full spectrum of non-alcoholic fatty liver disease. JHEP Rep, 4(5): 100477. https://doi.10.1016/j.jhepr.2022.100477. PMC9006858
22. Liu 22, Liu WL, Kuang T et al (2025) Cholesterol metabolism-related genes predict immune infiltration and prognosis in gastric cancer patients. J Cancer, 16(7): 2087–2102. https://doi.10.7150/jca.104389.
PMC12036097
23. .: 23, Steinberg GR, Valvano CM, De Nardo W et al (2025) Integrative metabolism in MASLD and MASH: Pathophysiology and emerging mechanisms. J Hepatol, 83(2): 584–595. https://doi.10.1016/j.jhep.2025.02.033
24. Li 24, Piao BS, Fu Y et al (2026) Lipid metabolism-MAFLD crosstalk: mechanisms and therapy. Front Endocrinol (Lausanne), 17: 1785178. https://doi.10.3389/fendo.2026.1785178. PMC13038599
25. .: 25, Arguello G, Balboa E, Arrese M et al (2015) Recent insights on the role of cholesterol in non-alcoholic fatty liver disease. Biochim Biophys Acta, 1852(9): 1765-78. https://doi.10.1016/j.bbadis.2015.05.015
26. .: 26, Deaton AM, Dubey A, Ward LD et al (2022) Rare loss of function variants in the hepatokine gene INHBE protect from abdominal obesity. Nat Commun, 13(1): 4319. https://doi.10.1038/s41467-022-31757-8.
PMC9329324
27. : Cao ZQXL Guo. The role of galectin-4 in physiology and diseases. Protein Cell (2016) 7(5): 314 – 24. https://doi.10.1007/s13238-016-0262-9.
PMC4853315
28. Xu 28, Zhao WY, Weng J et al (2025) Galectin-4 drives anti-PD-L1/BVZ resistance by regulating metabolic adaptation and tumour-associated neutrophils in hepatocellular carcinoma. Gut, https://doi.10.1136/gutjnl-2025-336374
29. Dang 29, Hao YS, Zhou W et al (2019) The traditional Chinese formulae Ling-gui-zhu-gan decoction alleviated non-alcoholic fatty liver disease via inhibiting PPP1R3C mediated molecules. BMC Complement Altern Med, 19(1): 8. https://doi.10.1186/s12906-018-2424-1.
PMC6323852

30. López-Soldado 30, Guinovart IJJ (2023) J Duran. Active Glycogen Synthase in the Liver Prevents High-Fat Diet-Induced Glucose Intolerance, Decreases Food Intake, and Lowers Body Weight. Int J Mol Sci, 24(3). https://doi.10.3390/ijms24032574. PMC9917303

Figures

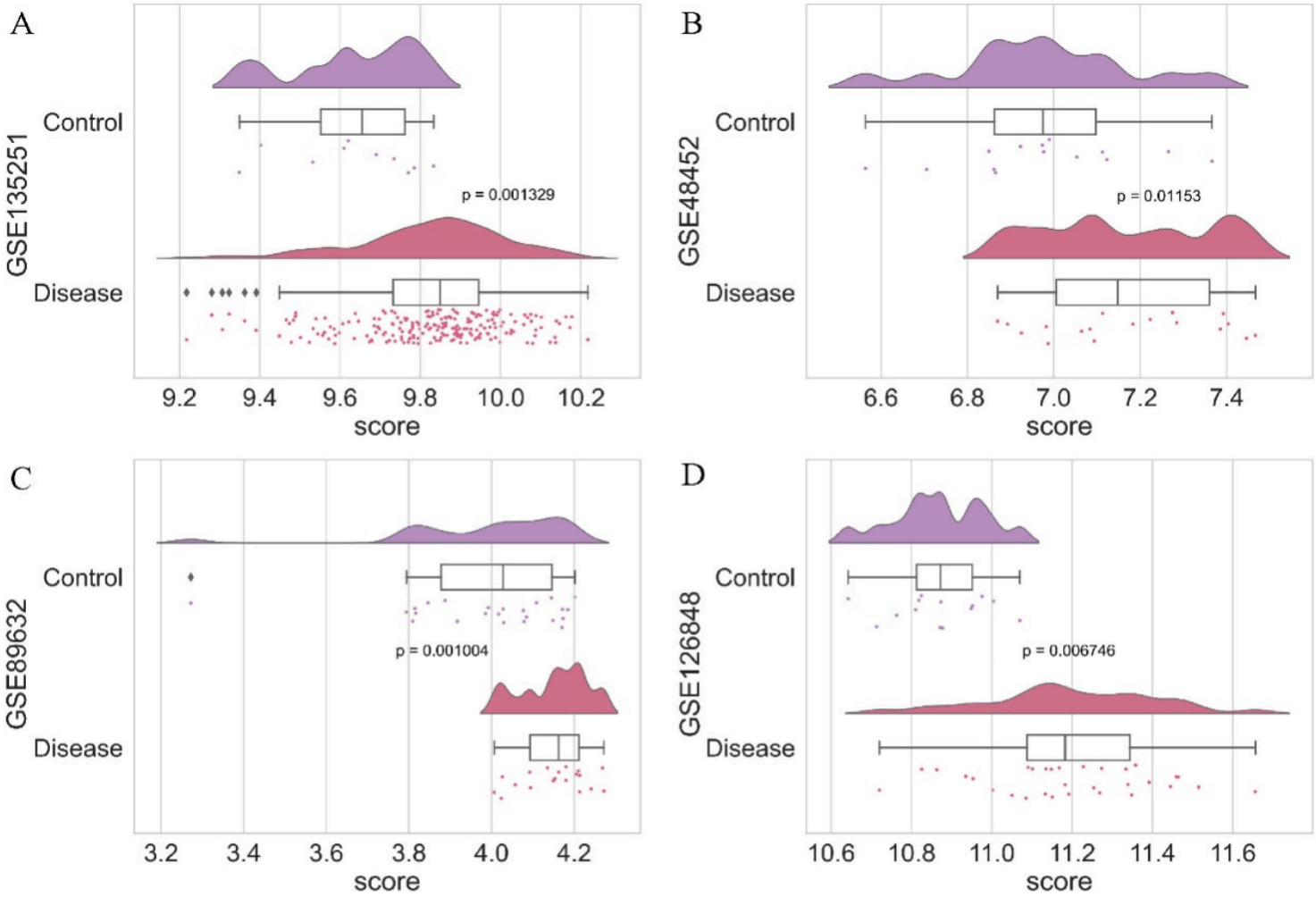


Figure 1

Cholesterol metabolism activity is significantly elevated in MAFLD liver tissue. (A–D) ssGSEA-derived cholesterol metabolism activity scores, calculated from a 50-gene cholesterol metabolism-related gene set (CMRGs), compared between MAFLD and normal control samples in the training set GSE135251 (A) and in the external validation sets GSE48452 (B), GSE89632 (C), and GSE126848 (D).

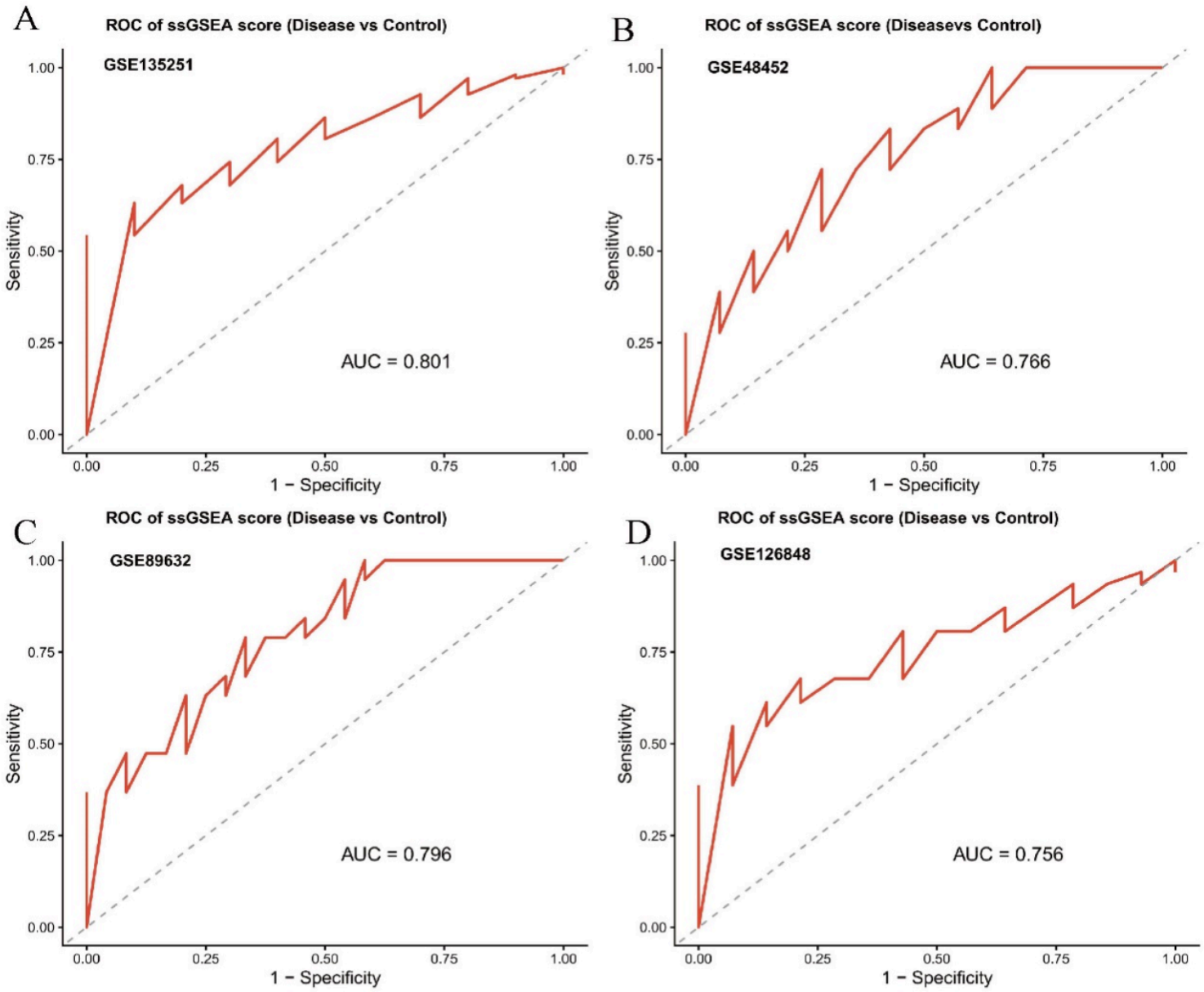


Figure 2

The cholesterol metabolism activity score has diagnostic value for MAFLD. (A–D) Receiver operating characteristic (ROC) curves evaluating the ability of the overall ssGSEA cholesterol metabolism activity score to discriminate MAFLD from normal controls in GSE135251 (A), GSE48452 (B), GSE89632 (C), and GSE126848 (D). The area under the curve (AUC) with 95% confidence interval is indicated in each panel.

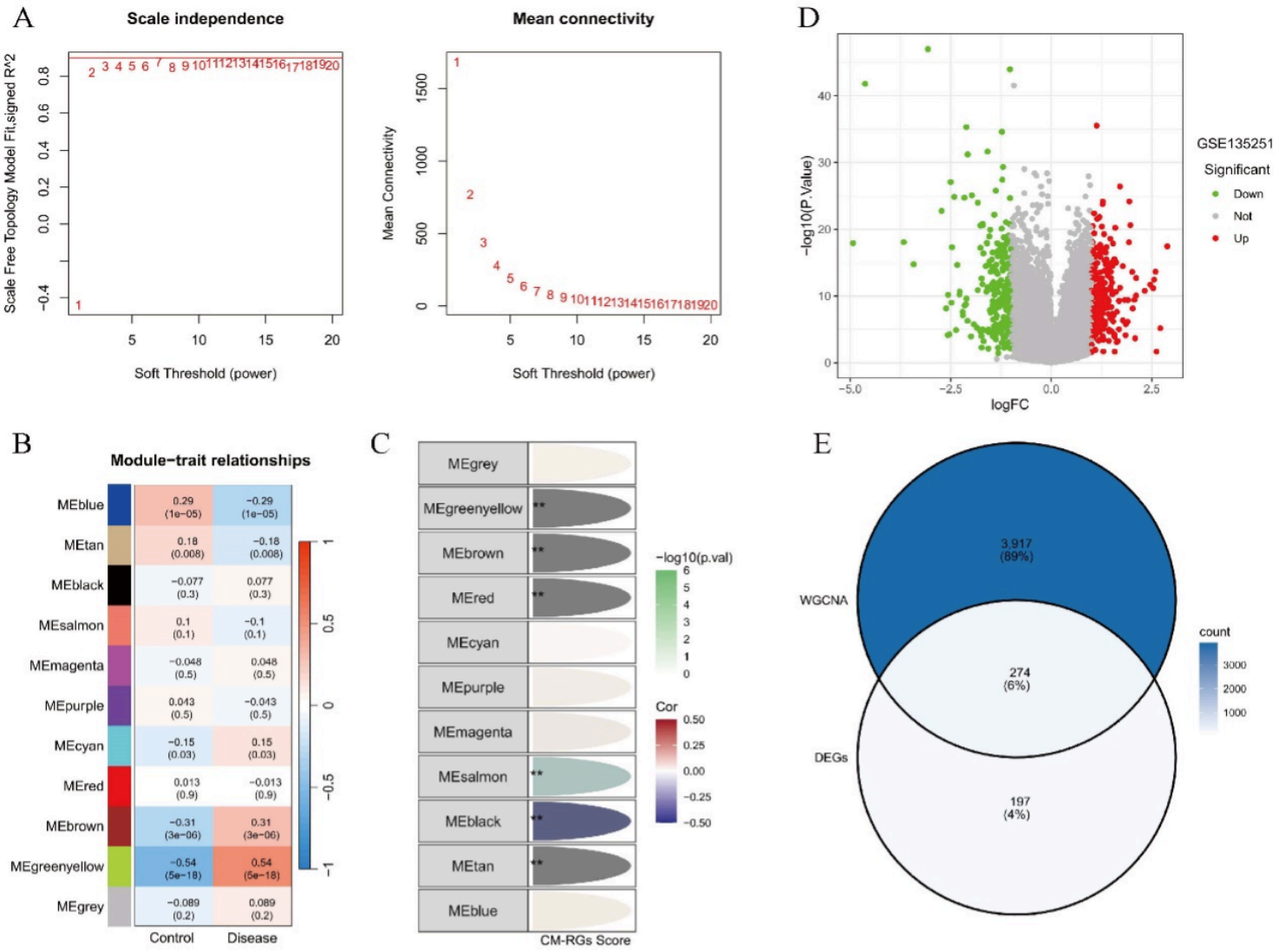


Figure 3

WGCNA and identification of MAFLD-related cholesterol metabolism candidate genes. (A) Scale-free topology fit index (left) and mean connectivity (right) as a function of the soft-thresholding power (β). A power of 7 was selected, at which the network satisfied the scale-free topology criterion (r^2 = 0.90). (B) Gene clustering dendrogram with module assignment by dynamic tree cutting; 11 co-expression modules were identified and are displayed in distinct colors beneath the dendrogram. (C) Heatmap of module–trait relationships. Each row corresponds to a module eigengene and each column to a trait (cholesterol metabolism activity score and MAFLD status). Each cell shows the correlation coefficient (upper) and the corresponding *P* value (lower). Six modules met the significance threshold ($|r| > 0.3$, $P < 0.05$): tan (r= −0.509), black (r = −0.490), salmon (r = −0.224), red (r = 0.658), brown (r = 0.761), and greenyellow (r = 0.654), collectively containing 4,191 module genes. (D) Volcano plot of DEGs between MAFLD and control samples in the training set (limma; thresholds $P < 0.05$ and $|\log_2 FC| > 1.0$). Red and blue dots denote significantly upregulated (n = 247) and downregulated (n = 224) genes, respectively; grey dots denote genes with no significant change. (E) Venn diagram showing the overlap between

WGCNA module genes and DEGs, which yielded 274 MAFLD-related cholesterol metabolism candidate dysregulated genes.

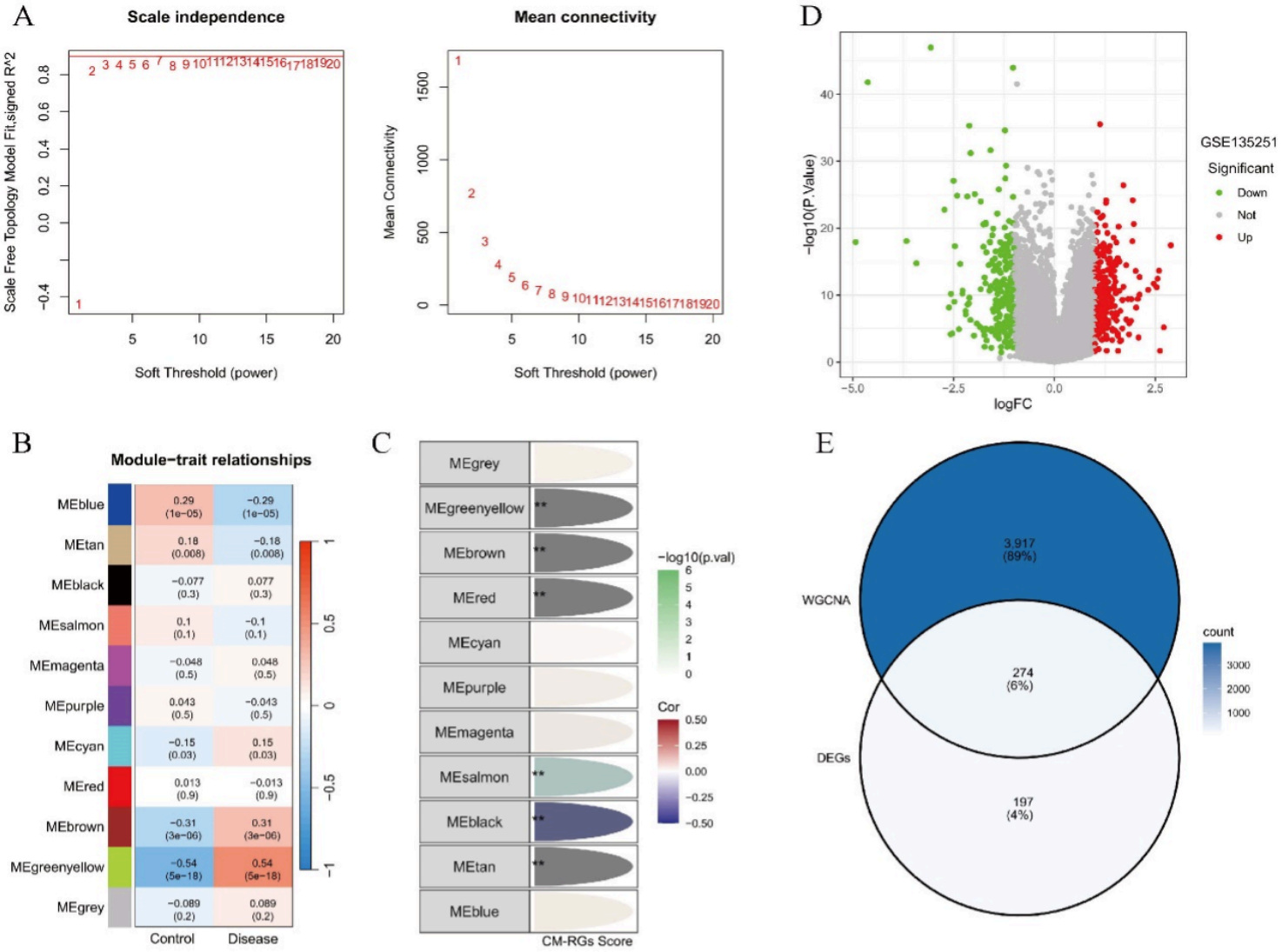


Figure 4

Functional enrichment, protein–protein interaction network, and machine learning-based model construction. (A) GO functional enrichment of the 274 candidate genes across the BP, CC, and MF categories. The top enriched terms are shown; the color scale indicates the adjusted *P* value and the dot size indicates the number of enriched genes. (B) KEGG pathway enrichment of the candidate genes. (C) Protein–protein interaction (PPI) network constructed using the STRING database (organism: *Homo sapiens*; confidence score > 0.4). Of the 274 candidate genes, 93 showed significant interactions and formed a core regulatory network; node size and color intensity are proportional to the node degree. (D) Performance of 113 prediction model combinations derived from 12 classical machine learning algorithms. Heatmap shows the AUC of each combination in the training and validation sets; the top-ranked combination is indicated. Feature selection across these models yielded 75 candidate genes for subsequent analysis.

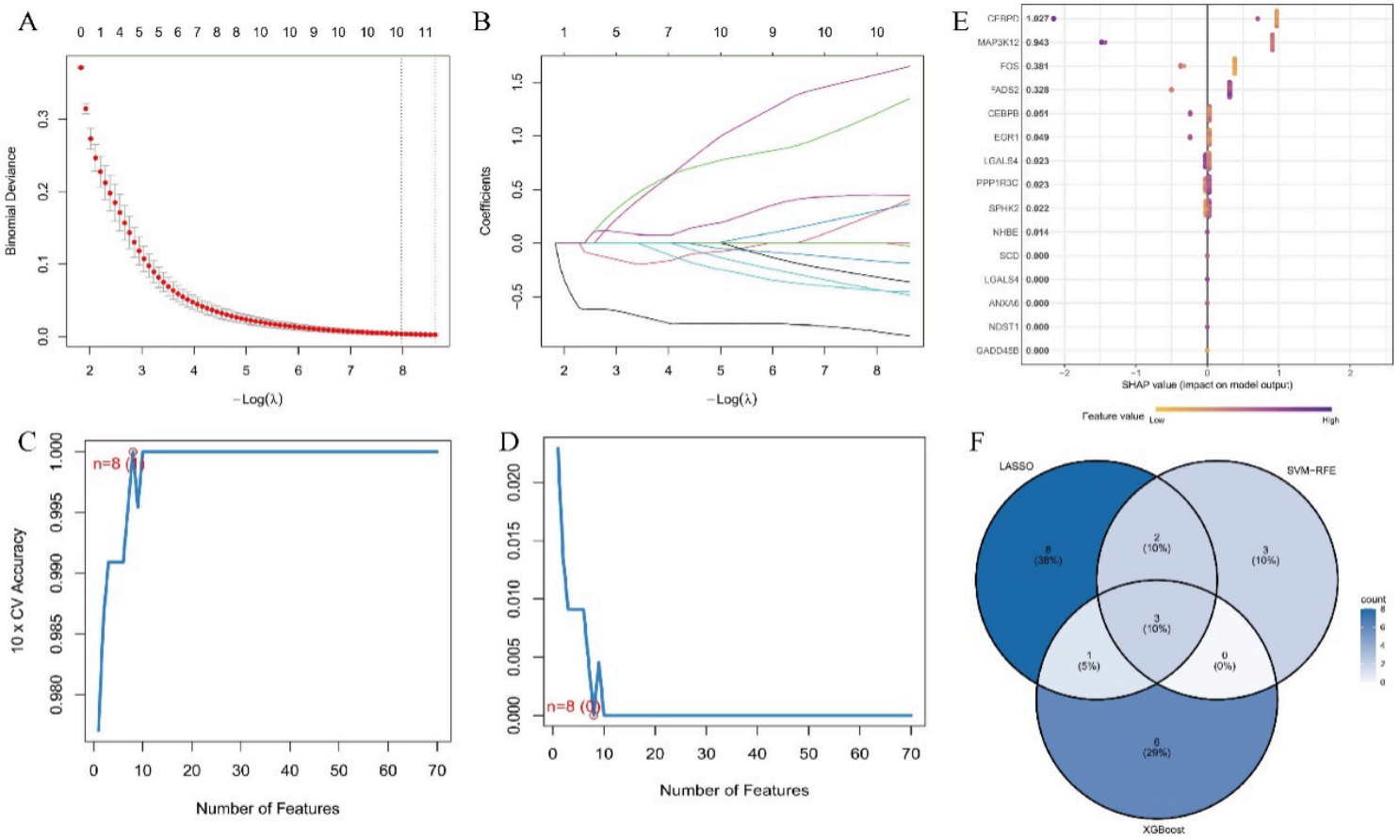


Figure 5

Machine learning-based screening of core MAFLD biomarkers. (A) LASSO coefficient profiles of the 93 core interacting genes plotted against log(λ). (B) Ten-fold cross-validation for LASSO: binomial deviance (y-axis) versus log(λ) (x-axis). The vertical dashed line indicates λ.min = 0.0213, at which 14 feature genes were retained. (C) SVM-RFE: cross-validation error rate as a function of the number of retained features; the minimum error rate was achieved with 7 feature genes. (D) SVM-RFE: cross-validation accuracy as a function of the number of retained features; the maximum accuracy was 0.844 at 7 feature genes. (E) XGBoost: ranking of the top 10 feature genes according to their mean SHAP (SHapley Additive exPlanations) values, which quantify each gene's contribution to model prediction. (F) Venn diagram showing the overlap among the gene sets identified by LASSO (n = 14), SVM-RFE (n = 7), and XGBoost (n = 10). Three genes—*INHBE*, *LGALS4*, and *PPP1R3C*—were commonly identified by all three algorithms and were defined as the core MAFLD biomarkers.

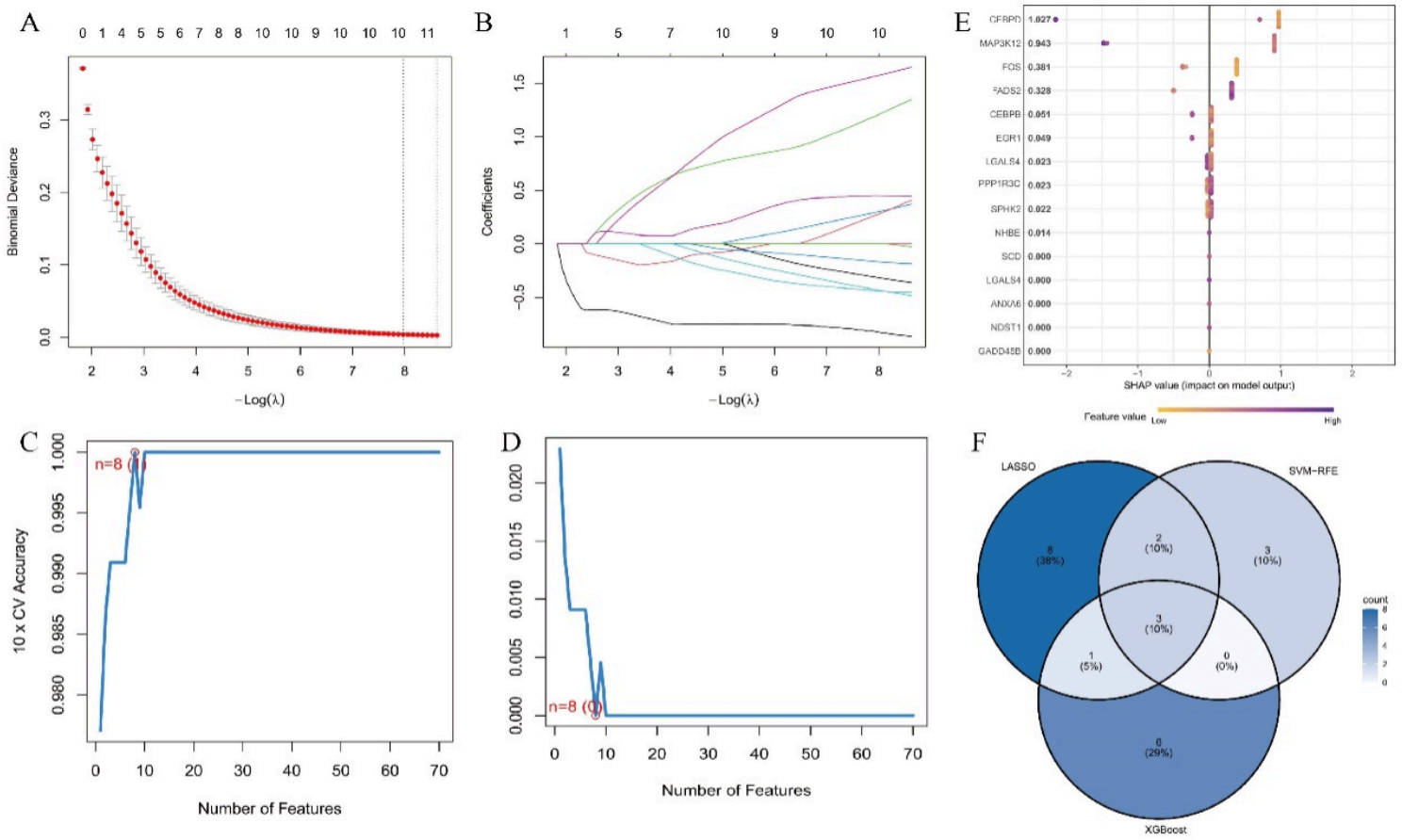


Figure 6

Diagnostic performance and expression validation of *INHBE*, *LGALS4*, and *PPP1R3C*. (A) ROC curves for the three core genes in the training set GSE135251. AUC = 0.940 for *INHBE*, 0.921 for *LGALS4*, and 0.939 for *PPP1R3C*. (B–D) ROC curves for the three genes in the external validation sets GSE48452 (B), GSE89632 (C), and GSE126848 (D). Mean AUC values across the validation sets were 0.894 for *INHBE*, 0.847 for *LGALS4*, and 0.815 for *PPP1R3C*. (E) Expression of *INHBE*, *LGALS4*, and *PPP1R3C* in MAFLD versus control samples in the training set GSE135251. (F–H) Expression of the three genes in the MAFLD and control groups of GSE48452 (F), GSE89632 (G), and GSE126848 (H). Data are presented as box plots; *P* values were determined by the Wilcoxon rank-sum test. $P < 0.05$, $P < 0.01$, $^{*}P < 0.001$.

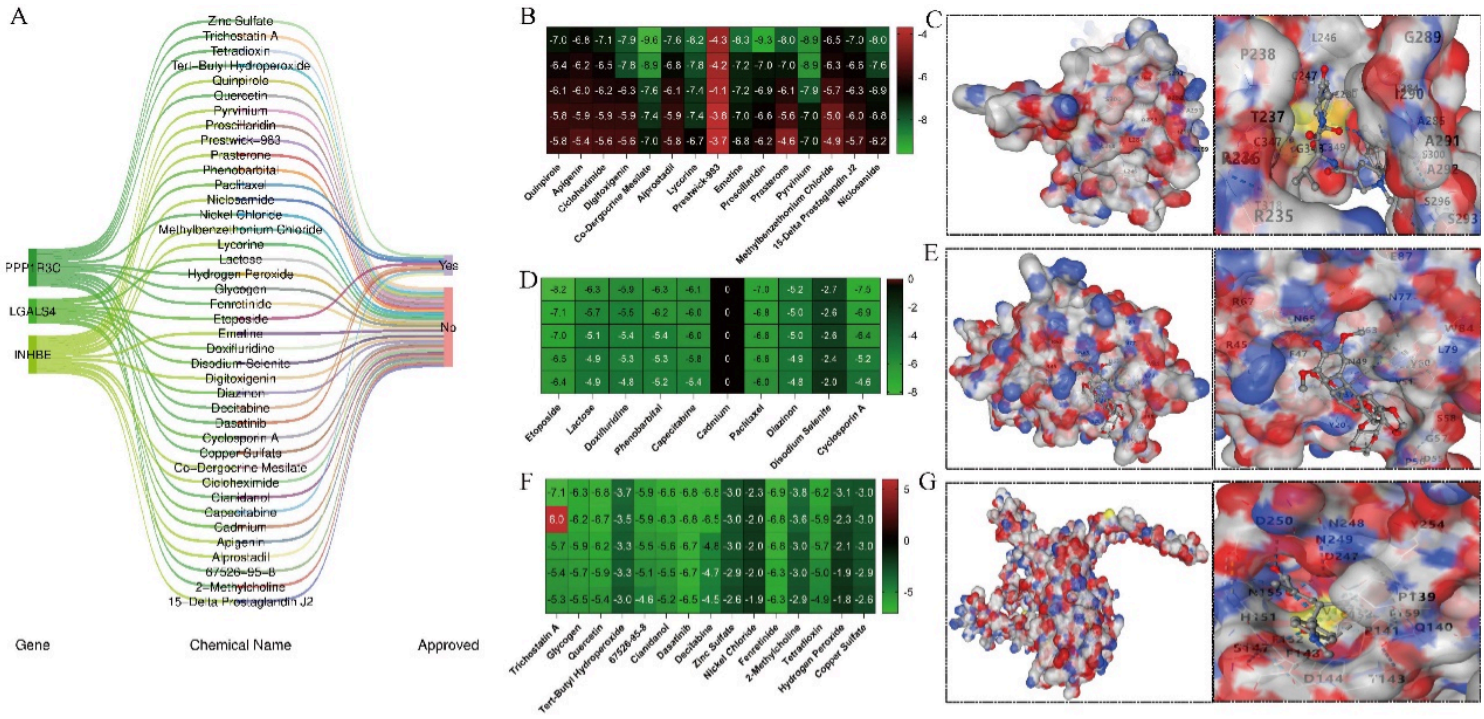


Figure 7

Targeted drug prediction and molecular docking validation. (A) Drug–gene interaction network predicted using the DSigDB database. A total of 15, 10, and 15 potential compounds were predicted to target INHBE, LGALS4, and PPP1R3C, respectively. (B, C) Binding mode of Co-Dergocrine Mesilate to the INHBE protein: (B) three-dimensional cartoon representation of the ligand within the predicted active pocket, and (C) two-dimensional schematic of the intermolecular interactions. The optimal binding free energy was −9.6 kcal/mol, and hydrogen bonds with key amino acid residues of INHBE are indicated by dashed lines. (D, E) Binding mode of Etoposide to the LGALS4 protein: (D) three-dimensional representation and (E) two-dimensional interaction schematic; optimal binding free energy = −8.2 kcal/mol. (F, G) Binding mode of Trichostatin A to the PPP1R3C protein: (F) three-dimensional representation and (G) two-dimensional interaction schematic; optimal binding free energy = −7.1 kcal/mol.

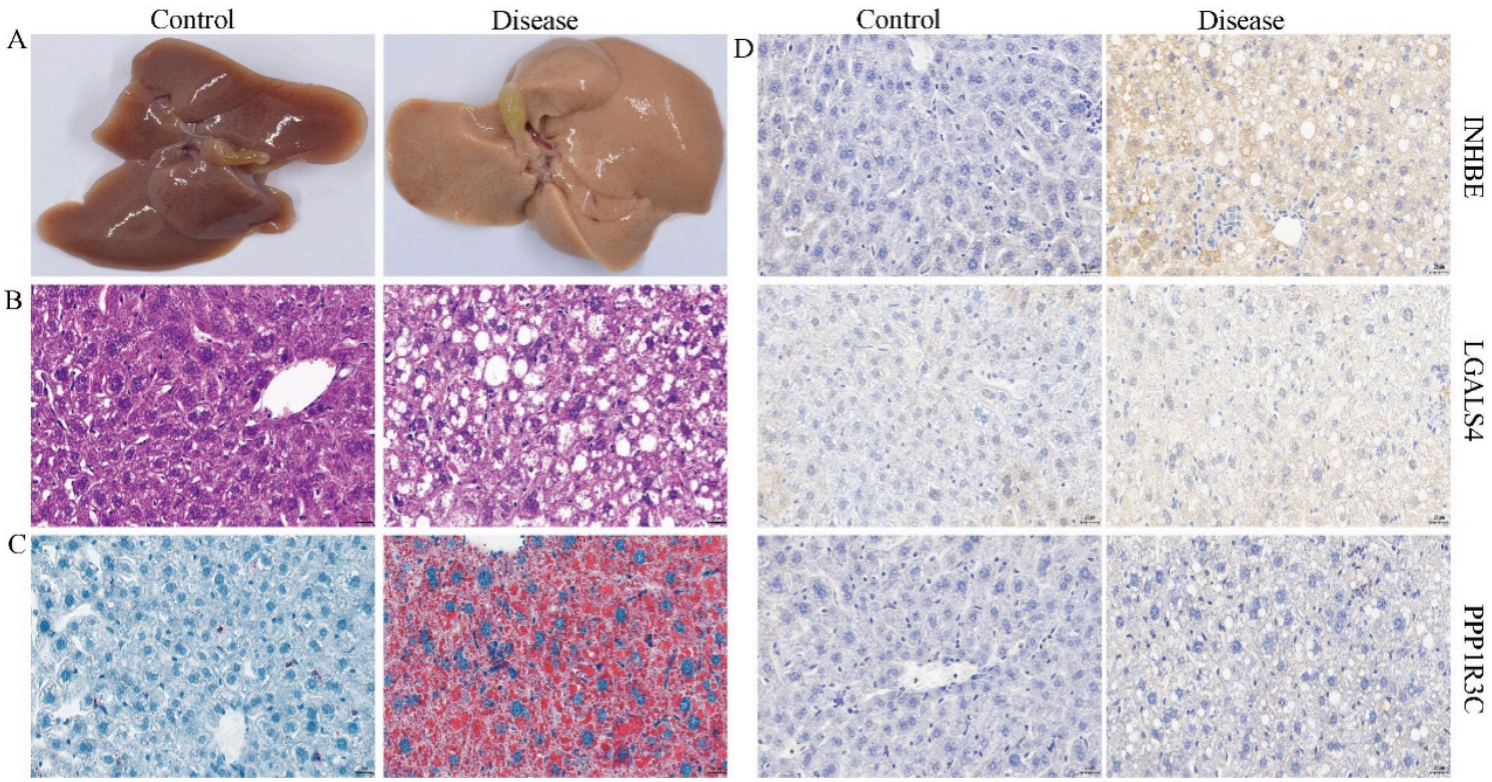


Figure 8

In vivo validation in a HFD-induced mouse MAFLD model. (A) Representative gross anatomical images of livers and gallbladders from control and HFD-fed mice. Livers from HFD-fed mice appeared pale and whitish, and abundant gallstones were observed in the gallbladder; no such changes were present in controls. (B) Representative H&E stained liver sections showing marked fatty vacuoles and lipid accumulation in HFD-fed mice. (C) Representative Oil Red O stained liver sections confirming hepatic lipid droplet accumulation in HFD-fed mice. (D) Representative IHC staining of INHBE, LGALS4, and PPP1R3C protein in liver sections from control and HFD-fed mice, with quantitative analysis of the percentage of positive area. Protein expression of all three genes was significantly higher in the HFD group than in the control group. Scale bars = 5 μm, n = 6 mice per group.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GraphicalAbstract.jpg