Lipids in Health and Disease

RESEARCH

Open Access

Multiple machine learning algorithms identify 13 types of cell death-critical genes in large and multiple non-alcoholic steatohepatitis cohorts

Renao Jiang[1†], Longfei Dai[1†], Xinjian Xu[1] and Zhen Zhang[1*]

Abstract

Background Dysregulated programmed cell death pathways mechanistically contribute to hepatic inflammation and fibrogenesis in non-alcoholic steatohepatitis (NASH). Identification of cell death genes may offer insights into diagnostic and therapeutic strategies for NASH.

Methods Data from multiple NASH cohorts were integrated, and 12 machine learning algorithms were applied to identify key dysregulated cell death-related genes and develop a binary classification model for NASH. Spearman's rank correlation coefficients quantified associations between these genes and clinical markers, immune infiltration profiles, and signature genes encoding pro-inflammatory mediators, metabolic regulators, and fibrotic drivers. Gene set enrichment analysis (GSEA) was performed to delineate the mechanistic underpinnings of these key genes. Consensus clustering analysis was then used to stratify patients with NASH into distinct phenotypic subgroups based on expression levels of these genes.

Results A NASH prediction model, developed using the random forest (RF) algorithm, demonstrated high diagnostic accuracy across multiple cohorts. Four key genes, enriched in lipid metabolism and inflammation pathways, were identified. Their transcriptional levels were significantly correlated with the non-alcoholic fatty liver disease activity score (NAS), hepatic inflammatory infiltration, molecular signatures of metabolic dysregulation (lipid homeostasis regulators), and fibrosis progression. These genes also enabled accurate classification of patients with NASH into clusters reflecting varying disease severity.

Conclusions A binary classification model, developed using the RF algorithm, accurately identified patients with NASH. The four cell death genes, identified through 12 machine learning algorithms, represent potential biomarkers and therapeutic targets for NASH. These genes contribute to inflammation-related immune cell activation, lipid metabolism dysregulation, and liver fibrosis, highlighting the complex interplay between cell death and NASH progression.

Keywords NASH, Cell death, Machine learning, Prediction model

Introduction

Non-alcoholic fatty liver disease (NAFLD) represents a leading cause of chronic liver disease, affecting approximately one-quarter of the global population and posing a significant burden on global health [1–3]. Based on the

†Renao Jiang and Longfei Dai contributed equally to this work.

*Correspondence:
Zhen Zhang
zhz36@sina.com
[1] Department of General Surgery, The First Affiliated Hospital of Anhui Medical University, 218 Jixi Road, Hefei, Anhui Province 230022, China

non-alcoholic fatty liver disease activity score (NAS), NAFLD is classified into non-alcoholic fatty liver (NAFL) and non-alcoholic steatohepatitis (NASH) [4]. Specifically, an NAS <4 indicates NAFL, whereas an NAS ≥ 4 denotes NASH [5]. Although NAFL typically follows a benign course, approximately one-third of patients with NAFLD progress to NASH, which may result in irreversible liver fibrosis, cirrhosis, and hepatocellular carcinoma (HCC) [6]. The strong association between NASH, severe liver disease, and metabolic disorders underscores the urgent need for research in this field [7]. Liver biopsy remains the gold standard for diagnosing NASH; however, its clinical use is limited by invasiveness, sampling variability, and inconsistent inter-observer reliability [8]. Even under optimal conditions, biopsy carries risks of complications [9]. Consequently, many patients with potential NASH remain undiagnosed, highlighting the critical need for non-invasive biomarkers.

Dysregulated programmed cell death, mitochondrial dysfunction, and endoplasmic reticulum stress are mechanistically implicated in the pathogenesis and progression of NAFLD [6, 10]. In the context of NASH, it is well established that the accumulation of hepatic lipids and elevated levels of free fatty acids induce lipotoxicity in hepatocytes, leading to cell death and infiltration of numerous mononuclear cells into the liver, thereby initiating chronic inflammation [11]. Hepatocyte death and inflammation are widely recognised as pivotal in the progression from simple steatosis to steatohepatitis, promoting liver fibrosis and disease advancement [12, 13]. During hepatic lipid overload, lipid accumulation in the hepatocytes may cause steatosis or trigger cell death through diverse mechanisms [14]. Such cell death elicits stress signals from hepatocytes, activates inflammatory pathways, and results in chronic liver injury, abnormal wound healing responses, and fibrosis [15]. Beyond classical cell death modes, such as apoptosis and necrosis, novel modalities—including autophagic cell death, pyroptosis, and necroptosis—have been identified in metabolic liver injury [16]. Identification of regulatory genes governing cell death in NASH progression is paramount, as these genes may reveal therapeutic targets.

Apoptosis, autophagy, and ferroptosis were identified as pivotal cell death pathways in NASH progression. Furthermore, prior diagnostic models were predicated on key ferroptotic and mitochondrial genes implicated in NASH pathogenesis, a novel diagnostic model incorporating a 13-gene cell death signature demonstrated significantly higher accuracy in detecting NASH. [17, 18]. The novelty of this study lies in the development of a predictive model that accurately identifies key cell death-associated genes and classifies patients with NASH. This advancement supports non-invasive diagnosis and targeted therapy, with the ultimate aim of uncovering critical therapeutic targets for NASH.

Methods

Integration of NASH datasets from GEO

Eleven datasets (GSE115193, GSE126848, GSE130970, GSE147304, GSE164760, GSE48452, GSE61260, GSE63067, GSE66676, GSE89632, and GSE135251) including NASH samples were retrieved from the Gene Expression Omnibus (GEO) database. Samples were included according to the following criteria: (1) liver biopsy-confirmed cases with a NAFLD Activity Score (NAS) of ≥ 4 were classified as NASH; (2) baseline liver tissue samples without histopathological alterations were included in the control group. Each dataset was converted into a gene expression matrix using the corresponding platform file (Supplementary Table 1). Batch effects were removed from ten datasets using the ComBat method, and the merged data were randomly allocated to training and testing sets in a 7:3 ratio. The GSE135251 dataset was served as an external validation set. Additionally, datasets GSE145412 and GSE55645 were retrieved from the GEO database. Normal control samples from GSE145421 were combined with NASH samples from GSE55645 to create a harmonised NASH peripheral blood cohort after rigorous batch-effect correction using the ComBat algorithm. Peripheral blood validation confirmed diagnostic accuracy and supported the clinical feasibility of liquid biopsy.

Identification of Differentially Expressed Genes (DEGs) and pathway analysis in NASH

Differentially expressed genes (DEGs) between normal and NASH groups in the training set were filtered based on $|log\ FC| > 0.5$ and *adjusted P* values < 0.05. These genes were intersected with cell death-related genes to determine key genes. Details of the 13 cell death-related genes are provided in Supplementary Table 2.

The Metascape platform (https://metascape.org/gp/index.html#/main/step1), which integrates over 40 biological databases, was used for functional annotation, pathway enrichment analysis, and visualisation of gene/protein datasets. Biological pathways and disease types associated with the key genes were explored using the "Metascape" database [19].

Construction of binary classification models using machine learning algorithms

Twelve machine learning algorithms—Random Forest (RF), Least Absolute Shrinkage and Selection Operator (Lasso), Enet, Ridge, Stepwise Generalized Linear Model (Stepglm), Support Vector Machine (SVM), Generalized Linear Model Boosting (glmBoost), Linear Discriminant Analysis (LDA), Partial Least Squares

Regression Generalized Linear Model (plsRglm), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Naive Bayes—were applied to screen DEGs and develop a binary classification model for NASH. Model performance was assessed by calculating the area under the receiver operating characteristic curve (AUC) across all GEO cohorts, with results visualised in a heatmap.

The Concordance Index (C-index) was used to evaluate the diagnostic consistency of the model across datasets. A higher C-index indicates greater model accuracy and robustness, particularly for large-scale sample cohorts, supporting clinical applicability. The model with the highest C-index was selected for further analysis. Confusion matrices [20] validated the accuracy of the selected model.

Identification and clustering of immune cells in single-cell datasets

Human single-cell dataset GSE159977 and mouse single-cell datasets GSE129516 and GSE158241 were obtained from the GEO database. Quality control and preprocessing were conducted using the "Seurat" package, including filtering low-quality cells and genes and normalising data. Principal component analysis (PCA) [21] was applied for dimensionality reduction, followed by clustering based on cell type. Cell annotation was performed using the "Single R" package to identify specific immune cells. Data were then clustered into distinct immune cell types based on liver immune cell surface markers sourced from the "CellMarker 2.0" database [22]. Immune cell surface markers for the liver are listed in Supplementary Table 3.

Comprehensive analysis of key genes in NASH

Key genes identified through the intersection of results from all 12 machine-learning algorithms were analysed. Gene set enrichment analysis elucidated biological pathways associated with upregulation or downregulation of these genes in NASH. Spearman correlation analysis was employed to elucidate the relationships between the expression levels of these key genes and clinical markers such as body mass index (BMI), NAS, and total cholesterol (TCHO), as well as the abundance of various immune cells and the expression levels of signature genes encoding pro-inflammatory mediators, metabolic regulators, and fibrotic drivers. Immune cell infiltration was quantified using four established algorithms: single-sample gene set enrichment analysis (ssGSEA), CIBERSORT [23], microenvironment cell population counter (MCP-counter), and estimation of the proportion of immune and cancer cells (EPIC).

Clustering NASH patients into distinct groups

The NASH patients were classified into two distinct clusters using the non-negative matrix factorization (NMF) [24] algorithm. Differences between clusters were evaluated based on key genes, pro-inflammatory genes, lipid metabolism genes, liver fibrosis genes, immune cell content, and biological pathways.

Experimental validation of key gene expression in NASH

Liver specimens were obtained from the Department of General Surgery of the First Affiliated Hospital of Anhui Medical University. The study protocol was approved by the Ethics Committee (Approval number: 2023497). Specimens from healthy individuals and patients with pathological obesity were fixed, dehydrated, embedded, and sectioned. Serial 4-μm liver sections underwent standardised H&E staining using Gill's haematoxylin (Biosharp, China) and eosin Y alcoholic solution (Biosharp, China), with differentiation in 1% acid alcohol for 30 s and bluing in 0.2% ammonium hydroxide for 1 min. RNA was extracted using the TRIzol reagent (Takara, China), followed by cDNA synthesis. Real-time quantitative PCR (qPCR) was conducted using *β-actin* as an internal reference gene to assess target gene expression. Primer sequences (Sangon Biotech, China) are listed in Supplementary Table 4. Proteins were extracted using ice-cold RIPA buffer with 1 mM PMSF (Beyotime, China), followed by electrophoretic separation, membrane transfer, blocking, incubation with primary and secondary antibodies. Immunoreactive bands were visualised and quantified using ImageJ software, with β-actin normalisation.

Statistical analysis

R version 4.3.2 and Strawberry Perl were used to execute the codes and scripts in this study. All statistical analysis was conducted with GraphPad Prism (version 9.0). Student's t-test was selectively employed for single comparisons when the data met parametric assumptions; otherwise, nonparametric tests were applied. $P < 0.05$ was considered statistically significant. Asterisks indicate the hierarchical thresholds: $P < 0.05$, $P < 0.01$, $P < 0.001$.

Flow chart

A comprehensive overview of the study design, experimental workflow, and analytical processes is illustrated in Fig. 1.

Results

Identification of twenty-one cell death genes associated with NASH

After batch-effect correction, samples from multiple cohorts (Fig. 2a) were integrated (Fig. 2b). The

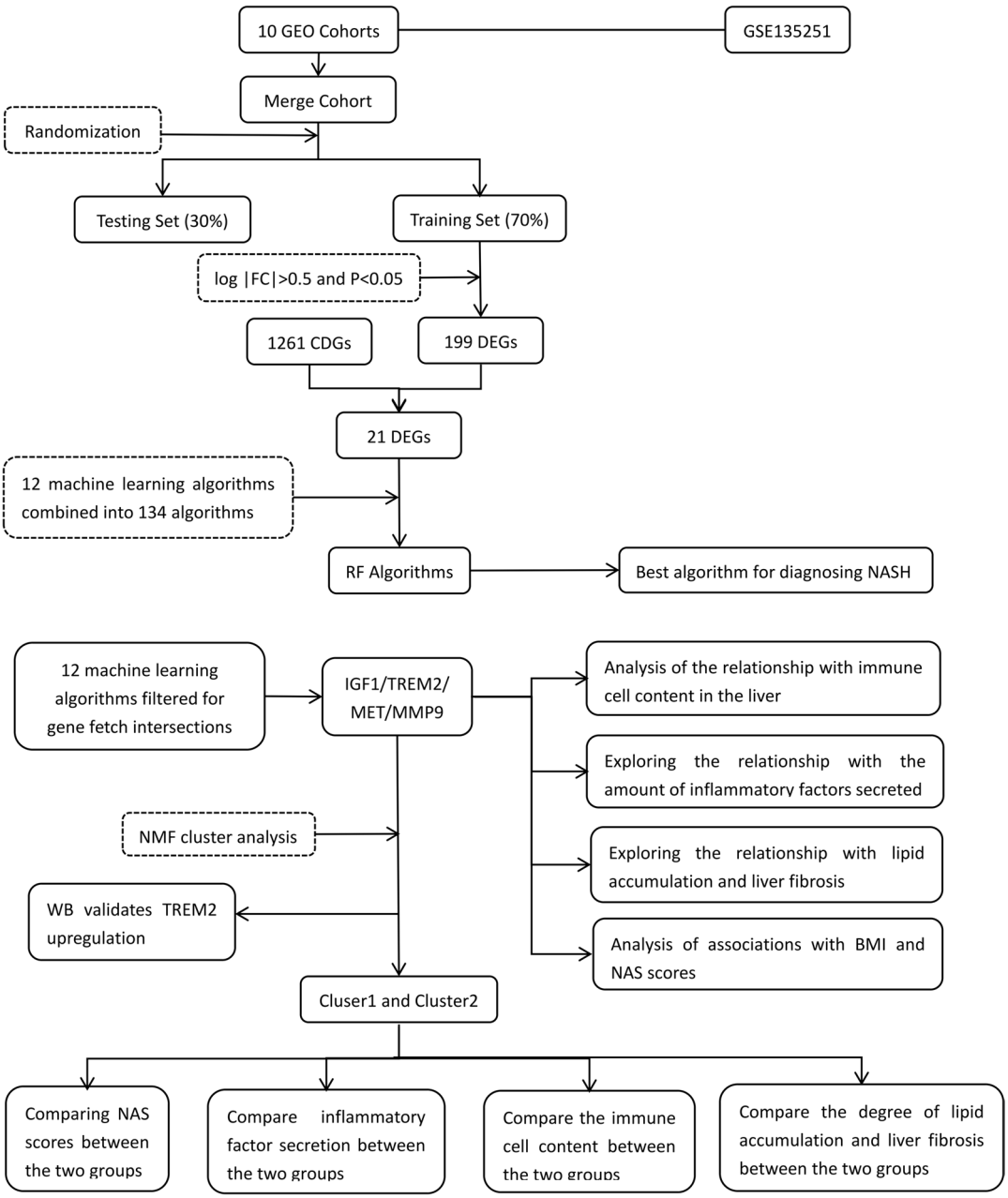

Fig. 1 The design thinking of this study

integrated dataset was randomly divided into training (70%) and testing (30%) sets (Fig. 2c). The number of normal and NASH patients included in the training, testing, and GSE135251 cohorts were 106/149, 43/52, and 10/155, respectively (Fig. 2d). Differential expression analysis revealed 67 downregulated and 132 upregulated genes in the NASH training cohort (Figs. 2e, f). These genes were intersected with 1,261 genes from 13 cell deaths to generate 21 common genes (Fig. 2g and h). Among these, seven genes, with IGF1 being the most prominent, were significantly downregulated in NASH, while 14 were significantly upregulated (Fig. 2i). These 21, considered potentially key regulators of the NASH pathogenic cascade, were further analysed to elucidate the underlying mechanisms. They were found to be involved in the "apoptosis signal pathway", "ferroptosis pathway", "lipid metabolism pathway" and

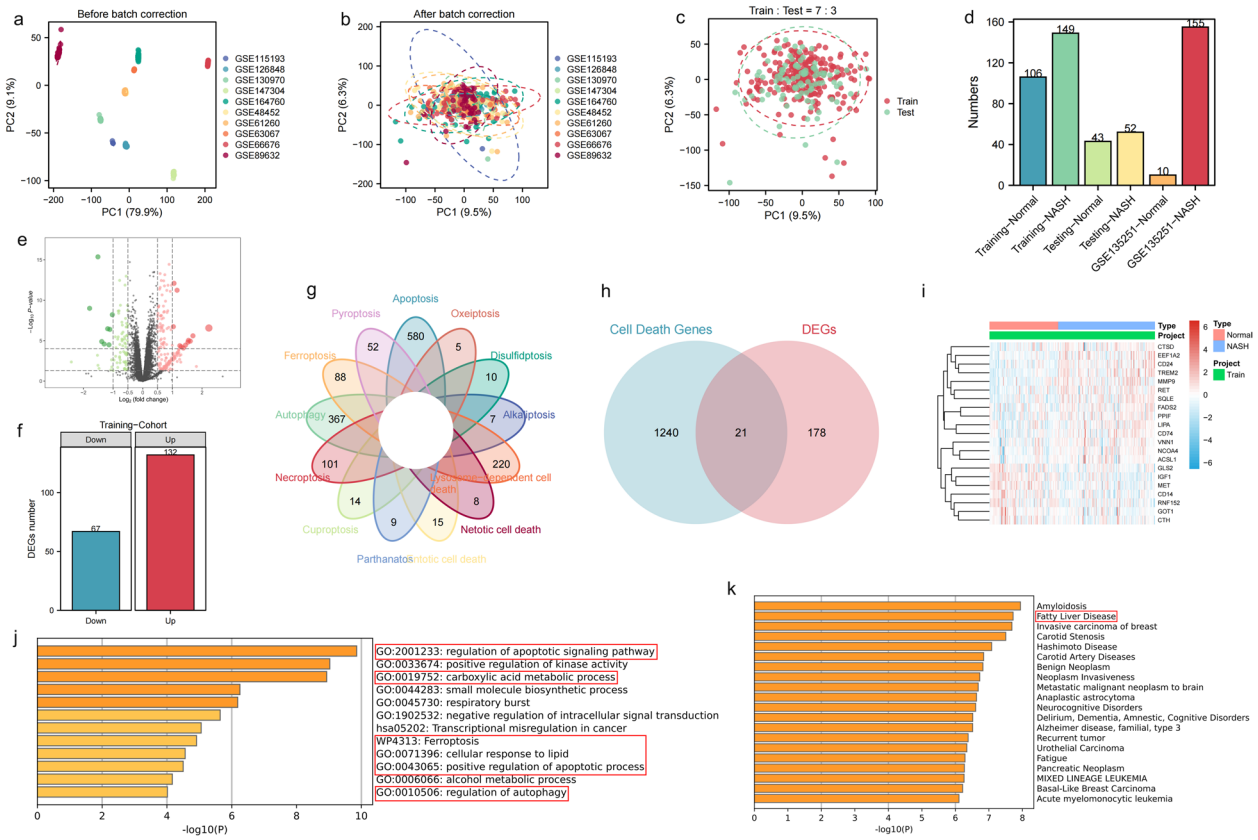

Fig. 2 Integration of GEO data and identification of cell death genes. **a** A PCA plot of multiple cohort distribution before removal of batch effects. **b** A PCA plot of the multiple cohort distribution after removing batch effects. **c** A PCA plot of samples distribution for the training and internal testing cohorts. **d** The number of normal samples and NASH samples. **e**, **f** Numbers of up-regulated and down-regulated differential genes in the training cohort. **g** Thirteen types of cell death and the number of genes each contains. **h** DEGs in NASH take intersection with cell death genes. **i** The up-regulated and down-regulated cell death genes in NASH. **j**, **k** Biological pathways and disease types involved in twenty-one genes

"autophagy pathway" (Fig. 2j), and showed strong associations with NAFLD (Fig. 2k).

Development of a binary classification model using the RF algorithm

The 21 genes were further refined using 134 iterations of 12 machine learning algorithms to construct a binary classification prediction model (Fig. 3a). The ROC values for predicting NASH patients across 16 GEO cohorts (Train, Test, Merge, GSE135251, GSE48452, GSE61260, GSE63067, GSE66676, GSE89632, GSE115193, GSE126848, GSE130970, GSE147304, GSE164760, GEO-Meta and blood cohort) were 1.000, 0.846, 0.960, 0.964, 0.992, 0.980, 0.857, 0.975, 1.000, 1.000, 0.996, 1.000, 0.960, 1.000, 0.925, and 0.744, respectively. The RF model identified five key genes, ranked by importance: *IGF1*, *TREM2*, *MET*, *NCOA4*, and *MMP9* (Fig. 3b). The RF model achieved the highest AUC values across all cohorts compared with individual gene diagnostic accuracy (Fig. 3c–q). Furthermore, the expression levels of these five genes showed consistent patterns across all cohorts: *IGF1* and *MET* were significantly downregulated in NASH, whereas the other three genes exhibited the opposite trend (Figs. 4a–o). In the peripheral blood cohort, *IGF1* was significantly downregulated, whereas the other four genes showed no statistically significant differential expression (Supplementary Fig. 7).

Enrichment of model genes in immune cells linked to inflammation and fibrosis

Human single-cell dataset GSE159977, mouse single-cell datasets 129,516, and GSE158241 were automatically annotated into four, thirteen, and nine cell types, respectively (Fig. 5a, d, g). Due to limitations in automatic annotation accuracy, datasets were manually re-annotated using liver immune cell surface markers from CellMarker 2.0, resulting in five, nine, and nine cell types, respectively (Fig. 5b, e, h). Enrichment of model genes in immune cells associated with NASH was analysed to understand their mechanistic roles. *NCOA4* was significantly enriched in

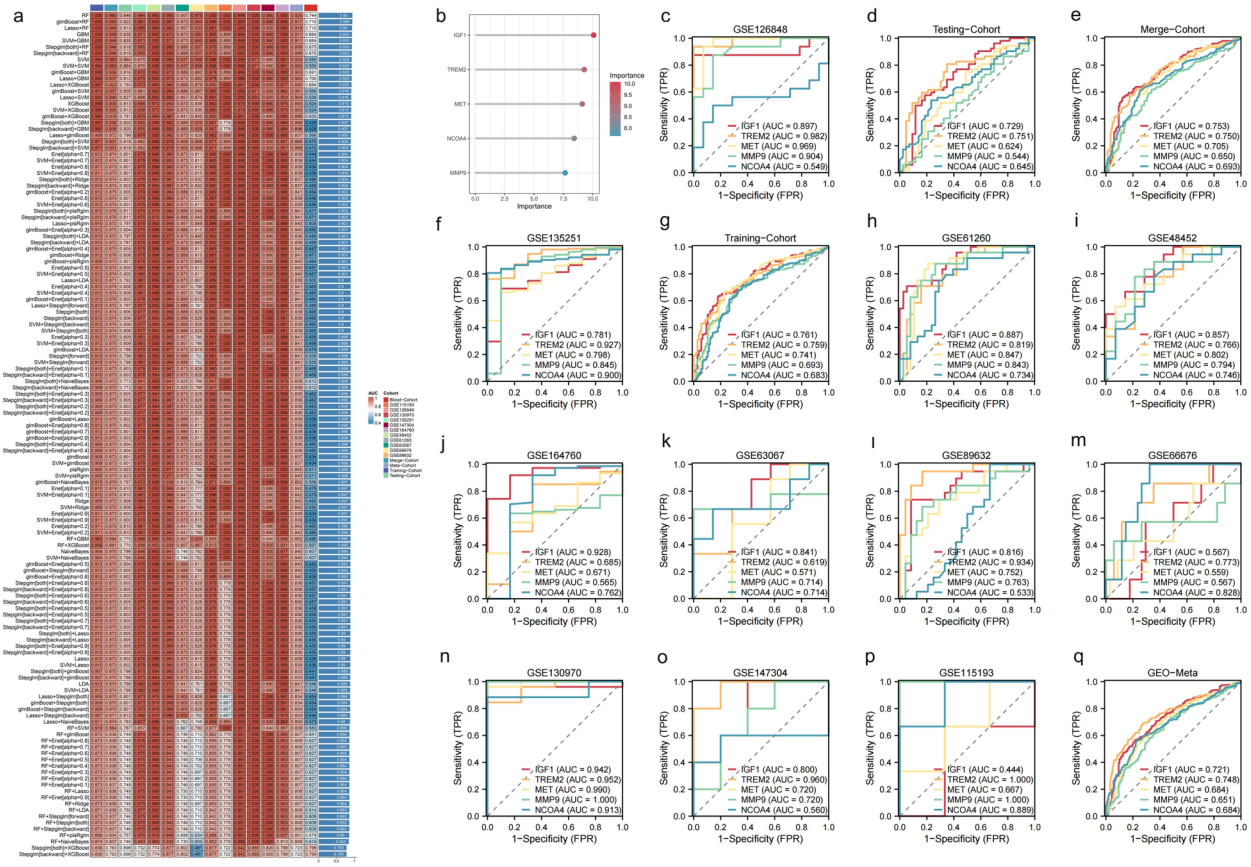

Fig. 3 The most effective diagnostic model for NASH was constructed with the RF algorithm. **a** The ROC values for the accuracy of diagnosing NASH patients across 15 GEO cohorts were derived from models built with 12 different machine learning algorithms. Among these, the Enet and Stepglm algorithms accepted input parameters. The Enet algorithm parameters were set to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9, while the Stepglm algorithm parameters were set as forward, backward, and both. **b** Five genes were selected by the RF algorithm and ranked according to their importance. **c**-**q** The ROC values ranking the accuracy of the models and the genes within the models in predicting NASH patients across 15 cohorts were presented

fibroblasts and hepatic stellate cells linked to liver fibrosis, *TREM2* in inflammatory macrophages, *MMP9* in pro-inflammatory neutrophils, and *IGF1* in hepatic stellate cells and macrophages (Fig. 5c, f, i).

Associations between four key genes and NAFLD clinical phenotypes

Further analyses focused on the most critical genes involved in cell death in NASH. Four common genes—*IGF1*, *TREM2*, *MET*, and *MMP9*—were identified by intersecting results from the 12 machine learning algorithms (Fig. 6a). Six cohorts with clinical information on NASH samples (including age, sex, BMI, and NAS) were analysed. No differences in expression levels of these genes were observed between male and female patients with NASH (Supplementary Fig. 1). However, the expression of *TREM2* demonstrated a strong positive association with NAS (Fig. 6b), whereas the expression of *MET* showed a significant inverse correlation with BMI and NAS (Fig. 6d, e). *IGF1* expression was positively correlated with age and negatively correlated with NAS (Fig. 6d, e, f). *MMP9* expression was positively correlated with NAS, hepatic steatosis, lobular inflammation, low-density lipoprotein (LDL), and total cholesterol (TCHO) (Fig. 6c, e, f, g).

Expression levels of key genes influence immunity, inflammation, metabolism, and fibrosis

In NASH, decreased *IGF1* expression was associated with significant downregulation of nicotinate and nicotinamide metabolism, linoleic acid metabolism, and glycine, serine, and threonine metabolism (Fig. 7a). Upregulation of *TREM2* was linked to significant overexpression of valine, leucine, and isoleucine biosynthesis and chemokine signalling pathway (Fig. 7b). Conversely, when *MET* was downregulated, nicotinate and nicotinamide metabolism was significantly underexpressed (Fig. 7c). Upregulation of *MMP9* was correlated with

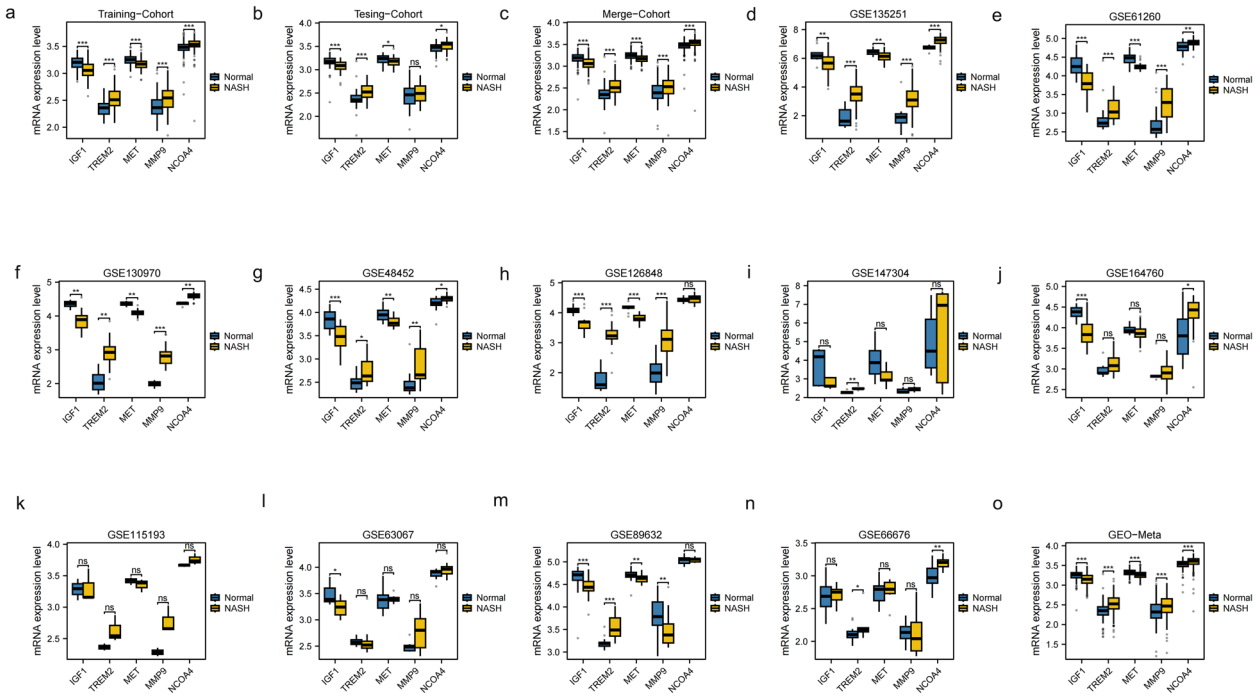

Fig. 4 The genes within the model exhibited differential expression between normal liver and NASH. **a**–**o** IGF1 and MET were either down-regulated or showed a trend towards down-regulation in NASH, whereas TREM2, MMP9, and NCOA4 displayed the opposite pattern

significant overexpression of pathways related to type II diabetes mellitus, MAPK signalling, and chemokine signalling (Fig. 7d). Four immune infiltration algorithms consistently revealed significant differences in immune cell infiltration levels between normal and NASH groups (Fig. 7e), with increased fibroblasts and macrophages content in the NASH group (Supplementary Fig. 2). The influence of key gene expression on immune cell content in NASH was further examined (Fig. 7f). *IGF1* and *MET* expression levels were negatively correlated with fibroblast content and positively correlated with M2 macrophage content. *MMP9* expression was positively correlated with neutrophil and Th1 cell content, while *TREM2* expression was positively correlated with fibroblast and M1 macrophage contents.

The expression of pro-inflammatory genes (*CCL2, IL1 A, IL1B, IL6*, and *TNFα*) was analysed. *IGF1* expression was negatively correlated with *TNFα*, whereas *TREM2* showed a positive correlation. *MMP9* expression was positively correlated with the levels of *CCL2, IL1B, and IL6* (Fig. 7g). *SCD, SREBF1, CEBPB,* and *FAS* are genes involved in fatty acids and cholesterol synthesis, while *ACOX1, PPARA*, and *CPT1 A* are associated with β-fatty acid oxidation. *LCN2* is related to both inflammation and lipid metabolism, and *PDGFA, COL1 A1,* and *COL3 A1* are implicated in fibrosis. *IGF1* expression was negatively correlated with *COL1 A1, COL3 A1, PDGFA,* and *SCD* and positively correlated with *PPARA*. *MET* expression was negatively correlated with *COL1 A1, LCN2*, and *SREBF1,* and positively correlated with *ACOX1, CPT1 A*, and *PPARA*. *MMP9* expression was negatively correlated with *ACOX1* and positively correlated with *CEBPB*. *TREM2* expression was positively correlated with *COL1 A1, COL3 A1, FAS, LCN2,* and *PDGFA* (Fig. 7h).

Non negative matrix factorization analysis of four key genes in NASH

NASH specimens were stratified into two molecular subtypes using non-negative matrix factorisation (NMF) clustering, based on the hepatic expression profiles of four key genes (Fig. 8a). Subsequent correlation analyses revealed significant differences in clinical phenotypes between the subtypes. The distribution of samples across the two clusters was clearly distinct (Fig. 8b). The expression levels of *IGF1* and *MET* were significantly lower in cluster 1, whereas *MMP9* and *TREM2* were higher (Fig. 8c). Elevated NAS distinguished cluster 1 from cluster 2 in both the training and the validation cohort GSE135251 (Fig. 8d, e). Additionally, cluster 1 demonstrated significant upregulation of pro-inflammatory effectors (*CCL2, IL1B, TNFα*) (Fig. 8f). Fibrosis-related genes (*PDGFA, COL1 A1, COL3 A1*), the lipid transport gene *LCN2*, and the lipid synthesis gene *CEBPB* were significantly upregulated in cluster 1, whereas β-fatty

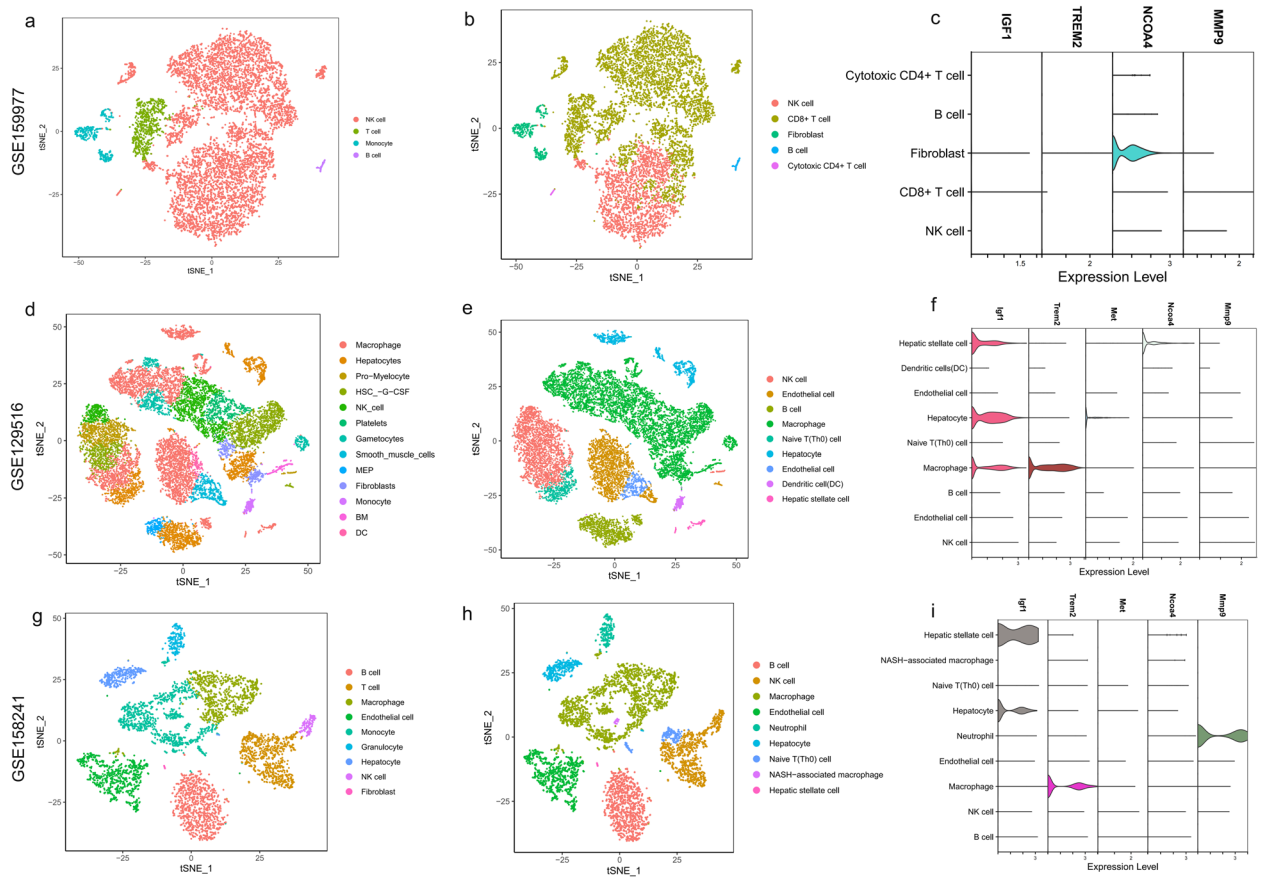

Fig. 5 The model genes were enriched in various immune cells. **a**, **d**, **g** Single-cell data were automatically annotated into different cell types. **b**, **e**, **h** Single-cell data were manually annotated into different cell types. **c**, **f**, **i** Cell types significantly enriched with model genes were identified

acid oxidation genes (*PPARA, ACOX1*) were significantly downregulated (Fig. 8g). Pathways related to MAPK signalling and type II diabetes mellitus were significantly enriched in cluster 1 (Fig. 8h). Immune cell content differed significantly between clusters (Fig. 8i), with higher fibroblast and macrophage content in cluster 1 and greater M2 macrophage content in cluster 2 (Supplementary Fig. 3).

Experimental validation of key genes expression

Histological examination with haematoxylin and eosin (H&E) staining confirmed normal liver morphology in one group and NASH pathology in another (Fig. 9a). At the mRNA level, *IGF1* and *MET* expression were significantly downregulated in NASH, whereas *MMP9* and *TREM2* were significantly upregulated (Fig. 9b) (Supplementary Table 5). Due to limited prior research on *TREM2* in NASH, its expression was validated at the protein level. Western blot analysis confirmed significant upregulation of *TREM2* in NASH (Fig. 9c, d), with the original blot image provided in Supplementary Fig. 4.

Discussion

The global rise in obesity has contributed to an increasing incidence of NAFLD [25, 26]. NASH, a severe clinical manifestation of NAFLD, is characterised by steatosis, inflammation, cellular injury, and apoptosis [27]. If untreated, NASH may progress to cirrhosis, liver fibrosis, and potentially to liver cancer [28]. Despite its high prevalence and significant impact on global health, no approved treatments for NASH currently exist due to incomplete understanding of its pathogenesis [29, 30]. Identification of key cell death-related genes in NASH is therefore critical for improving diagnosis and developing targeted therapies.

In this study, a predictive model based on five genes (*IGF1, TREM2, MET, MMP9, NCOA4*) developed using the RF algorithm, demonstrated high diagnostic accuracy

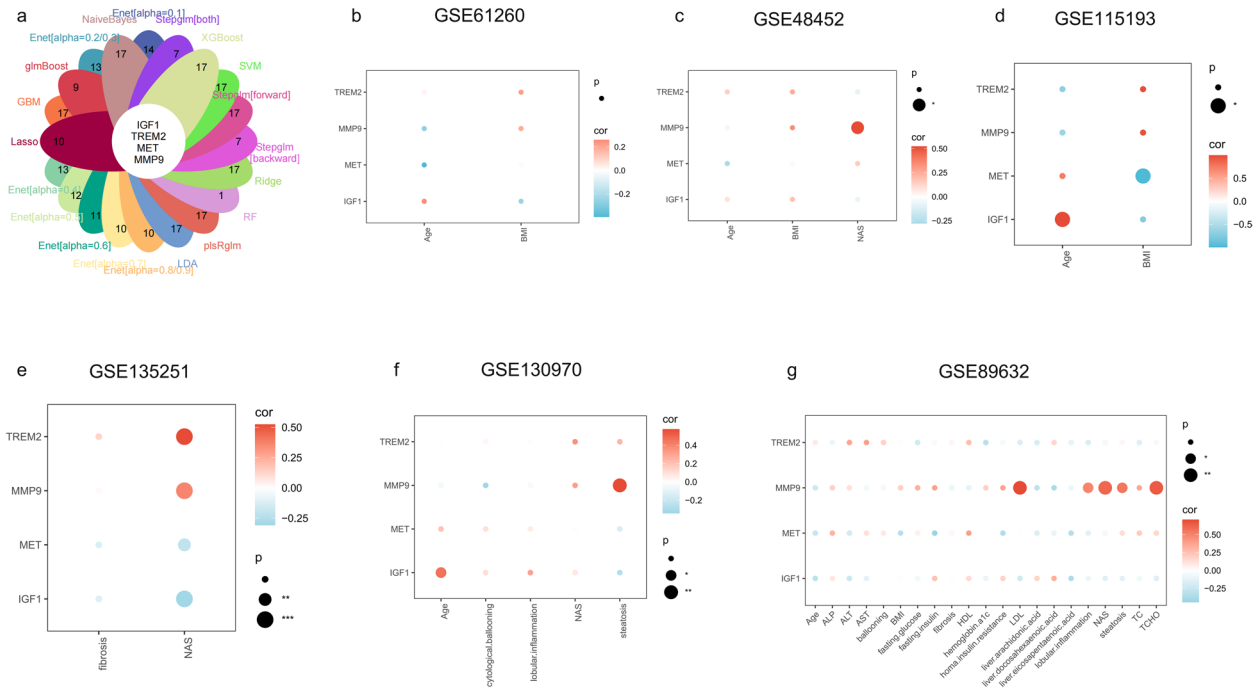

Fig. 6 The expression levels of key genes influence clinical metrics. **a** Twelve machine learning algorithms identified the same four genes. **b**-**g** The expression levels of these four key genes were associated with various clinical markers related to NASH in six GEO cohorts containing clinical information

for NASH across multiple cohorts. Furthermore, compared with liver biopsy, this model's non-invasive nature supports its potential for widespread clinical adoption.

Four cell death-related genes—*IGF1, MET, MMP9*, and *TREM2*—were identified as key regulators in NASH through 12 machine learning algorithms. Expression levels of these genes enabled stratification of patients with NASH into two distinct subtypes with differing disease severity, offering valuable insights for clinical assessment.

IGF1 was found to be downregulated in NASH and enriched in stellate cells. Dysregulation of *IGF1* may trigger stellate cell transformation into myofibroblasts, leading to the secretion of large amounts of extracellular matrix (ECM) that promote liver fibrosis [31]. This is further supported by its negative correlation with fibroblast content and with the expression levels of fibrosis-related genes (*PTGFA, COL1 A1*, and *COL3 A1*). Furthermore, *IGF1* is significantly enriched in liver macrophages, with its expression showing a positive correlation with the content of anti-inflammatory M2 macrophages [32] and a negative correlation with the pro-inflammatory factor TNFα. These findings indicates that *IGF1* is involved in inflammatory cell infiltration and promotes inflammation in NASH. Additionally, *IGF1* was negatively correlated with the expression of fatty acid synthesis gene *SCD* and positively correlated with the expression of β-fatty acid oxidation gene *PPARA*, indicating a role in lipid metabolism and accumulation in NASH. Downregulation of *IGF1* was also associated with the suppression of linoleic acid metabolism and amino acid metabolic pathways in the liver, thereby exacerbating NASH progression. Notably, the NAS, which reflects the histological diagnosis of NASH [33], was reduced when *IGF1* was downregulated. Therefore, the downregulation of *IGF1* in NASH appears to promote lipid accumulation, inflammatory infiltration, and fibrosis progression in the liver.

TREM2 was significantly upregulated in NASH and enriched in hepatic macrophages. Its expression was positively correlated with fibroblast and M1 macrophage content, as well as *TNFα, COL1 A1, COL3 A1, PDGFA, FAS,* and *LCN2* expression. Enrichment of *TREM2* was observed in chemokine signalling pathways, and its expression was positively correlated with NAS. These findings suggest that *TREM2* promotes inflammatory infiltration by activating M1 macrophages, stimulates lipid synthesis and contributes to hepatic fibrosis, potentially serving as a clinical biomarker for NASH severity [34], with higher expression levels indicating a more severe disease.

MMP9 was upregulated in NASH and significantly enriched in neutrophils. Its expression was positively correlated with neutrophils, macrophages, and Th1 cell

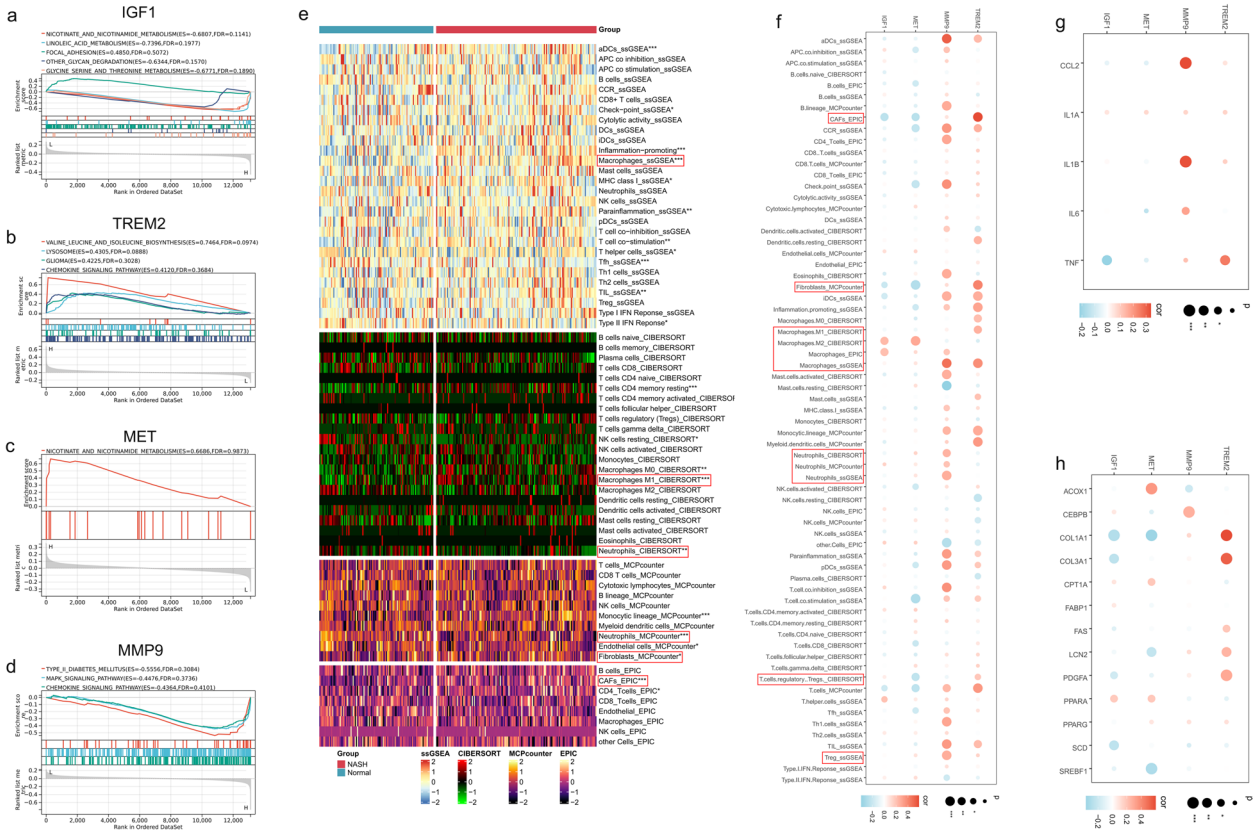

Fig. 7 Key genes are associated with inflammation, lipid metabolism, and fibrosis. **a**-**d** Biological pathways significantly enriched when these four genes are up-regulated or down-regulated in NASH. Higher or lower enrichment scores (ES) indicate more pronounced enrichment of biological pathways in the samples. A lower false discovery rate (FDR) indicates higher confidence in the results. **e** Significant differences in the content of certain immune cells were observed between the normal group and the NASH group. **f** The expression levels of key genes were significantly correlated with immune cell populations. **g** The expression levels of key genes were significantly associated with inflammatory genes. **h** The expression levels of key genes were significantly associated with genes related to lipid metabolism and fibrosis

content [35, 36], as well as *CCL2, IL1B, IL6,* and *CEBPB* expression, but negatively correlated with *ACOX1*. Enrichment of *MMP9* in MAPK, chemokine signalling, and type 2 diabetes mellitus pathways was noted. Positive correlations with clinical markers, including NAS, low-density lipoprotein (LDL), and total cholesterol (TCHO), suggest that MMP9 enhances inflammatory infiltration via MAPK pathway activation and neutrophil-derived factor release, promotes fatty acid synthesis, and inhibits fatty acid oxidation.

Downregulation of *MET* in NASH was negatively correlated with BMI and NAS. *MET* expression also showed a negative correlation with fibroblast content but a positive correlation with M2 macrophage content. Its associations with lipid synthesis, fatty acid oxidation, and hepatic fibrosis suggest that *MET* downregulation contributes to inflammatory infiltration, lipid accumulation, and hepatic fibrosis in NASH.

In this study, four cell death-related genes—*TREM2, IGF1, MET,* and *MMP9*—were identified as critical to NASH pathogenesis. Insulin-like growth factor 1 (*IGF1*), primarily synthesised in the liver in response to growth hormone (GH) stimulation, plays a critical role in metabolic regulation. Dysregulation of *IGF1* or GH signalling exacerbates insulin resistance and promotes hepatocyte injury through lipotoxicity [37]. Notably, studies have demonstrated that *IGF1* overexpression attenuates hepatic stellate cell activation in mice with CCl4-induced liver injury and may inhibit fibrosis by promoting stellate cell senescence [38, 39]. Triggering receptor expressed on myeloid cells 2 (*TREM2*), an innate immune receptor predominantly expressed on myeloid cells, is expressed in hepatic Kupffer cells. Elevated *TREM2* expression has been observed in multiple murine models of liver injury and hepatocytes from high-fat (HF) diet-fed mice [40–42]. Mechanistically, *TREM2*

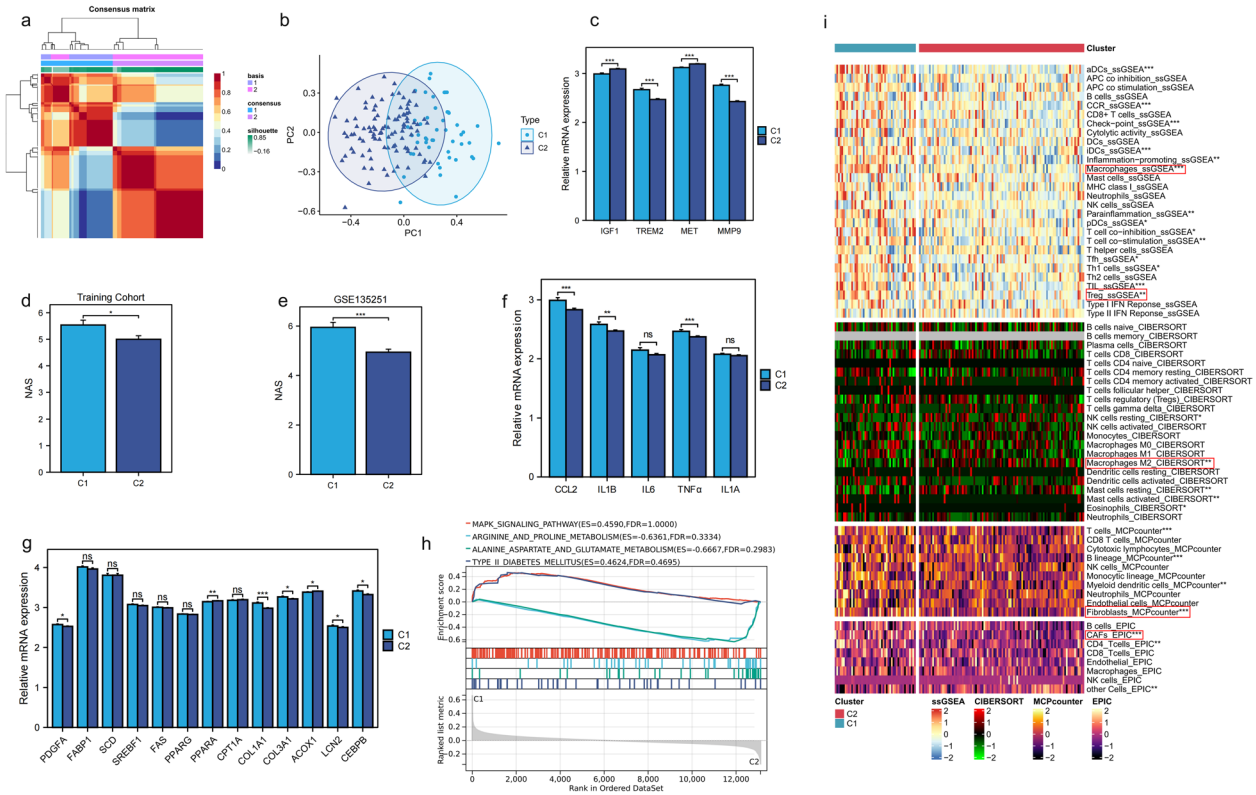

Fig. 8 Patients were classified into clusters 1 and clusters 2 of varying disease severity. **a** NMF clustering analysis. Patients with NASH were divided into two clusters. **b** Samples in the two clusters were distinctly separated. **c** The expression of key genes differed significantly between the two clusters. **d** The NAS score was higher in cluster 1 of the training cohort. **e** The NAS score was higher in cluster 1 of the GSE13251 cohort. **f** Inflammatory genes were expressed at higher levels in cluster 1. **g** Metabolic genes differed significantly between the two clusters. **h** Pathways significantly enriched between the two clusters. **i** The content of immune cells showed significant differences between the two clusters

suppresses Toll-like receptor (TLR) signalling in macrophages, thereby inhibiting pro-inflammatory cytokine production and chemokine release [43]. Depletion of *TREM2* exacerbates hepatic damage by amplifying reactive oxygen species (ROS) DNA damage, disrupting cellular redox homeostasis, and promoting apoptosis [43]. Matrix metalloproteinase 9 (*MMP9*), a gelatinase belonging to the matrix metalloproteinase (MMP) family, facilitates ECM degradation and remodelling and plays a critical role in suppressing hepatic fibrosis. In early-stage fibrosis, *MMP9* transcription is upregulated by *ETV4*, which promotes collagenolysis and ECM remodelling. However, in advanced fibrosis, *MMP9* expression is suppressed due to increased matrix stiffness, which perpetuates HSC activation and collagen deposition [44].

Study strengths and limitations

This study identified *IGF1, TREM2, MET, NCOA4,* and *MMP9* as potential non-invasive biomarkers for NASH through comprehensive bioinformatic analysis and machine learning, particularly using the RF algorithm. Diagnostic performance was further validated using serum samples from patients with NASH, highlighting the model's clinical utility and addressing key limitations associated with invasive liver biopsy. These findings propose a novel, clinically applicable approach for early detection and disease monitoring in NAFLD. However, the precise molecular mechanisms by which these cell death-related genes contribute to NASH pathogenesis remain unclear. Future studies should explore the dynamic interactions between these genes and hepatic lipid metabolism, inflammatory pathways, and fibrotic signalling to identify actionable therapeutic targets.

Conclusion

In this study, multiple machine learning algorithms were employed to develop non-invasive diagnostic models for NASH using data from large, multicohort

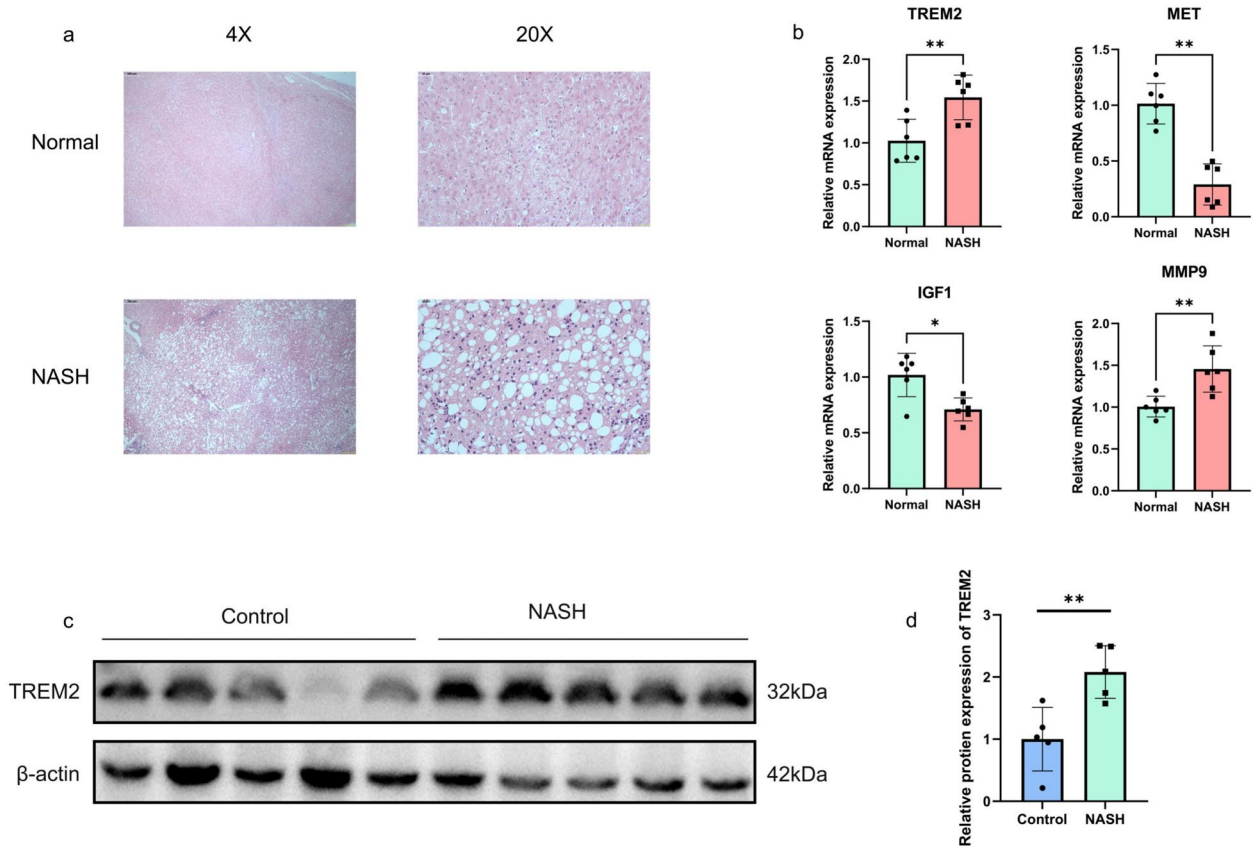

Fig. 9 The expression levels of four key genes were found to differ between the two groups. **a** Pathological pictures of normal liver and NASH. **b** q-PCR experiments. **c** WB experiments. **d** WB results were visualized using Image J software

samples. Among these, the RF algorithm demonstrated the highest diagnostic accuracy. The intersection of genes identified across algorithms revealed four key genes—*TREM2, IGF1, MET,* and *MMP9*— that are significantly associated with NASH pathogenesis. These genes are involved in inflammatory infiltration, lipid synthesis, fatty acid oxidation, and hepatic fibrosis. Importantly, their expression profiles enable the stratification of patients with NASH according to disease severity. Collectively, these findings provide a foundation for the development of non-invasive diagnostic tools and contribute to the identification of molecular targets for therapeutic intervention in NASH.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12944-025-02588-5.

Additional file 1: Supplementary Fig. 1. Comparison of IGF1, TREM2, MET, MMP9 expression levels between males and females. (a) GSE48452 cohort. (b) GSE61260 cohort. (c) GSE89632 cohort. (d) GSE130970 cohort

Additional file 2: Supplementary Fig. 2. The differences in the content of various immune cells between the control and NASH. (a)EPIC algorithm. (b) ssGSEA algorithm. (c) MCPcounter algorithm. (d) CIBERSORT algorithm

Additional file 3: Supplementary Fig. 3. The differences in the content of various immune cells between cluster 1 and cluster 2. (a)EPIC algorithm. (b) ssGSEA algorithm. (c) MCPcounter algorithm. (d) CIBERSORT algorithm

Additional file 4: Supplementary Fig. 4. The original WB figures. (a) PVDF membrane before being cut. (b) After the PVDF membrane was cropped. (c) Chemiluminescence figure of β-actin containing marker. (d) Chemiluminescence figure of TREM2 containing marker

Additional file 5: Supplementary Fig. 5. Ethics in Biology Chinese Edition

Additional file 6: Supplementary Fig. 6. Principal Component Analysis (PCA) plots are used to show the distribution of each dataset before and after the removal of the batch effect

Additional file 7: Supplementary Fig. 7. Expression of IGF1, TREM2, MET, MMP9 and NCOA4 in peripheral blood cohorts

Additional file 8: Supplementary Table 1. The specific information of the 11 GEO cohorts included in this study. Supplementary Table 2. The number and names of genes comprising the 13 cell death types. Supplementary Table 3. Immune cell surface markers in human and mouse liver from the CellMarker 2.0 database. Supplementary Table 4. Primer information for the five genes involved in this study. Supplementary Table 5. Statistical information of PCR results of four key genes

Additional file 9: Supplementary Table 6. Strengths and limitations of the 12 machine learning algorithms

Acknowledgements
Acknowledgments to GEO database (https://www.ncbi.nlm.nih.gov/geo/), Metascape database (https://metascape.org/gp/index.html#/main/step1), CellMarker2.0 database (http://bio-bigdata.hrbmu.edu.cn/CellMarker/index.html).

Statements and declarations
The authors declare that they have no competing interests.

Authors' contributions
Study conceptualization and methodology were established by R.A.J. NASH data was acquired, processed, and analyzed by RAJ. Experiments were carried out by RAJ, LFD, XJX. Critical revision for intellectual content carried out by ZZ. All authors critically reviewed and approved the final manuscript, and take full responsibility for the integrity of the work.

Funding
This work was funded by the Anhui Provincial Department of Education (Project Number: 2023 AH053301).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
The ethical aspect of the study was aptly addressed, with the research protocol receiving due approval from the Ethics Committee of The First Affiliated Hospital of Anhui Medical University (Approval number: 2023497) (*Supplementary Fig. 5*). To uphold ethical standards, all participants provided their informed consent prior to their inclusion in the study.

Competing interests
The authors declare no competing interests.

Received: 9 January 2025 Accepted: 29 April 2025
Published online: 08 May 2025

References

1. Xu X, Poulsen KL, Wu L, Liu S, Miyata T, Song Q, Wei Q, Zhao C, Lin C, Yang J. Targeted therapeutics and novel signaling pathways in non-alcohol-associated fatty liver/steatohepatitis (NAFL/NASH). Signal Transduct Target Ther. 2022;7:287.
2. Saiman Y, Duarte-Rojo A, Rinella ME. Fatty Liver Disease: Diagnosis and Stratification. Annu Rev Med. 2022;73:529–44.
3. Younossi Z, Anstee QM, Marietti M, Hardy T, Henry L, Eslam M, George J, Bugianesi E. Global burden of NAFLD and NASH. trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol. 2018;15:11–20.
4. Brunt EM, Kleiner DE, Wilson LA, Belt P, Neuschwander-Tetri BA. Nonalcoholic fatty liver disease (NAFLD) activity score and the histopathologic diagnosis in NAFLD. distinct clinicopathologic meanings. Hepatology. 2011;53:810–20. https://doi.org/10.1002/hep.24127.
5. Tincopa MA, Loomba R. Non-invasive diagnosis and monitoring of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis. Lancet Gastroenterol Hepatol. 2023;8:660–70.
6. Friedman SL, Neuschwander-Tetri BA, Rinella M, Sanyal AJ. Mechanisms of NAFLD development and therapeutic strategies. Nat Med. 2018;24:908–22.
7. Kanwal F, Shubrook JH, Younossi Z, Natarajan Y, Bugianesi E, Rinella ME, Harrison SA, Mantzoros C, Pfotenhauer K, Klein S, et al. Preparing for the NASH Epidemic: A Call to Action. Gastroenterology. 2021;161:1030-1042.e1038.
8. Leow WQ, Chan AW, Mendoza PGL, Lo R, Yap K, Kim H. Non-alcoholic fatty liver disease: the pathologist's perspective. Clin Mol Hepatol. 2023;29:S302-s318.
9. Piccinino F, Sagnelli E, Pasquale G, Giusti G. Complications following percutaneous liver biopsy. A multicentre retrospective study on 68,276 biopsies. J Hepatol. 1986;2:165–73.
10. Fromenty B, Roden M. Mitochondrial alterations in fatty liver diseases. J Hepatol. 2023;78:415–29.
11. Machado MV, Diehl AM. Pathogenesis of Nonalcoholic Steatohepatitis. Gastroenterology. 2016;150:1769–77.
12. Schuster S, Cabrera D, Arrese M, Feldstein AE. Triggering and resolution of inflammation in NASH. Nat Rev Gastroenterol Hepatol. 2018;15:349–64.
13. Lee KC, Wu PS, Lin HC. Pathogenesis and treatment of non-alcoholic steatohepatitis and its fibrosis. Clin Mol Hepatol. 2023;29:77–98.
14. Marra F, Svegliati-Baroni G. Lipotoxicity and the gut-liver axis in NASH pathogenesis. J Hepatol. 2018;68:280–95.
15. Schwabe RF, Tabas I, Pajvani UB. Mechanisms of Fibrosis Development in Nonalcoholic Steatohepatitis. Gastroenterology. 2020;158:1913–28.
16. Park W, Wei S, Kim BS, Kim B, Bae SJ, Chae YC, Ryu D, Ha KT. Diversity and complexity of cell death: a historical review. Exp Mol Med. 2023;55:1573–94.
17. Dai L, Jiang R, Zhan Z, Zhang L, Qian Y, Xu X, Yang W, Zhang Z. Machine learning-based algorithm identifies key mitochondria-related genes in non-alcoholic steatohepatitis. Lipids Health Dis. 2024;23:137.
18. Dai L, Yuan W, Jiang R, Zhan Z, Zhang L, Xu X, Qian Y, Yang W, Zhang Z. Machine learning-based integration identifies the ferroptosis hub genes in nonalcoholic steatohepatitis. Lipids Health Dis. 2024;23:23.
19. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10:1523.
20. Cabot JH, Ross EG. Evaluating prediction model performance. Surgery. 2023;174:723–6.
21. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26:303–4.
22. Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, Chen J, Jiang W, Yang K, Ou Q, et al. Cell Marker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. 2023;51:D870-d876.
23. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods Mol Biol. 2018;1711:243–59.
24. Liefeld T, Huang E, Wenzel AT, Yoshimoto K, Sharma AK, Sicklick JK, Mesirov JP, Reich M. NMF Clustering: Accessible NMF-based Clustering Utilizing GPU Acceleration. J Bioinform Syst Biol. 2023;6:379–83.
25. Fan JG, Kim SU, Wong VW. New trends on obesity and NAFLD in Asia. J Hepatol. 2017;67:862–73.
26. Polyzos SA, Kountouras J, Mantzoros CS. Obesity and nonalcoholic fatty liver disease: From pathophysiology to therapeutics. Metabolism. 2019;92:82–97.
27. Sheka AC, Adeyi O, Thompson J, Hameed B, Crawford PA, Ikramuddin S. Nonalcoholic Steatohepatitis: A Review. Jama. 2020;323:1175–83.
28. Younossi ZM. Non-alcoholic fatty liver disease - A global public health perspective. J Hepatol. 2019;70:531–44.

29. Tokushige K, Ikejima K, Ono M, Eguchi Y, Kamada Y, Itoh Y, Akuta N, Yoneda M, Iwasa M, Yoneda M, et al. Evidence-based clinical practice guidelines for nonalcoholic fatty liver disease/nonalcoholic steatohepatitis 2020. J Gastroenterol. 2021;56:951–63.
30. Sumida Y, Yoneda M. Current and future pharmacological therapies for NAFLD/NASH. J Gastroenterol. 2018;53:362–76.
31. Wang S, Li K, Pickholz E, Dobie R, Matchett KP, Henderson NC, Carrico C, Driver I, Borch Jensen M, Chen L, et al. An autocrine signaling circuit in hepatic stellate cells underlies advanced fibrosis in nonalcoholic steatohepatitis. Sci Transl Med. 2023;15:eadd3949.
32. Kazankov K, Jørgensen SMD, Thomsen KL, Møller HJ, Vilstrup H, George J, Schuppan D, Grønbæk H. The role of macrophages in nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. Nat Rev Gastroenterol Hepatol. 2019;16:145–59.
33. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, Ferrell LD, Liu YC, Torbenson MS, Unalp-Arida A, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology. 2005;41:1313–21.
34. Indira Chandran V, Wernberg CW, Lauridsen MM, Skytthe MK, Bendixen SM, Larsen FT, Hansen CD, Grønkjær LL, Siersbæk MS, Caterino TD, et al. Circulating TREM2 as a noninvasive diagnostic biomarker for NASH in patients with elevated liver stiffness. Hepatology. 2023;77:558–72.
35. Romagnani S. Biology of human TH1 and TH2 cells. J Clin Immunol. 1995;15:121–9.
36. Romagnani S. Th1/Th2 cells. Inflamm Bowel Dis. 1999;5:285–94.
37. Jensen EA, Young JA, Mathes SC, List EO, Carroll RK, Kuhn J, et al. Cross-talk between the growth hormone/insulin-like growth factor-1 axis and the gut microbiome: A new frontier for microbial endocrinology. Growth Hormon IGF Res. 2020;53–54:101333.
38. Sanz S. Expression of insulin-like growth factor I by activated hepatic stellate cells reduces fibrogenesis and enhances regeneration after liver injury. Gut. 2005;54:134–41.
39. Nishizawa H, Iguchi G, Fukuoka H, Takahashi M, Suda K, Bando H, et al. IGF-I induces senescence of hepatic stellate cells and limits fibrosis in a p53-dependent manner. Sci Rep. 2016;6:34605.
40. Coelho I, Duarte N, Barros A, Macedo MP, Penha-Gonçalves C. Trem-2 Promotes Emergence of Restorative Macrophages and Endothelial Cells During Recovery From Hepatic Tissue Damage. Front Immunol. 2021;11:616044.
41. Ray K. A protective role for TREM2 in liver injury. Nat Rev Gastroenterol Hepatol. 2018;15:130–1.
42. Hendrikx T, Porsch F, Kiss MG, Rajcic D, Papac-Miličević N, Hoebinger C, et al. Soluble TREM2 levels reflect the recruitment and expansion of TREM2+ macrophages that localize to fibrotic areas and limit NASH. J Hepatol. 2022;77:1373–85.
43. Perugorria MJ, Esparza-Baquer A, Oakley F, Labiano I, Korosec A, Jais A, et al. Non-parenchymal TREM-2 protects the liver from immune-mediated hepatocellular damage. Gut. 2019;68:533–46.
44 Goyale A, Jain A, Smith C, Papatheodoridi M, Misas MG, Roccarina D, et al. Assessment of non-alcoholic fatty liver disease (NAFLD) severity with novel serum-based markers: A pilot study. Strnad P, editor. PLoS ONE. 2021;16:e0260313.

Publisher's Note