An Integrated Multi-omics Mendelian Randomization Identifies Predictive Transcription Gene Signature of Liver Fibrosis

Xiaoyan Wang

cathywxy@sjtu.edu.cn

Shanghai Jiao Tong University  https://orcid.org/0000-0001-5568-2131

Lin Zhang
Jiangjin District Central Hospital

Yuhaow Chang
Shanghai Jiao Tong University

Yuhuai Guo
Shanghai Jiao Tong University Medical Library: Shanghai Jiao Tong University School of Medicine

Guangze Yang
The University of Adelaide

Wenjun Xie
Shanghai Jiao Tong University

Min Zhu
Shanghai Sixth People's Hospital: Shanghai 6th Peoples Hospital Affiliated to Shanghai Jiao Tong University

Jisi Teng
Shanghai Jiao Tong University

Jessie Shen
Aurora Health Care Geneva Family Practice SC: Aurora Health Care

Wei Jia
Shanghai Sixth People's Hospital: Shanghai 6th Peoples Hospital Affiliated to Shanghai Jiao Tong University

Shaoqiu Chen
University of Hawai‘i at Manoa

Tianlu Chen
Shanghai Sixth People's Hospital: Shanghai 6th Peoples Hospital Affiliated to Shanghai Jiao Tong University

Youping Deng
Abstract

Background

Liver fibrosis is a critical deteriorating onset stage in NASH (Nonalcoholic steatohepatitis) progression towards cirrhosis and even liver cancer. Currently, there is still a lack of non-invasive diagnostic markers for hepatic fibrosis. We conduct multiple public databases associated with Pathway, Network and Mendelian randomization (MR) analysis to identify transcribed genes potentially involved in liver fibrosis and assess their diagnostic efficiency applicable to multiple races.

Methods

We first leveraged the advanced capabilities of the MetaIntegrator package in R. Four discovery cohorts and four validation cohorts were searched for expression profiling that biopsy diagnosed NASH patients and then the results were validated in plasma samples of two Chinese cohorts. The resulting gene signature was then conducted by GO enrichment analysis and DisGeNET enrichment analysis. Network analysis were employed using MetaboAnalyst 5.0. We then conducted MR analysis using data from IEU Open GWAS project (average N = 23,818), and GWAS Catalog (N = 8,299), the UK Biobank (N = 3,108) and FinnGen (average N = 373,007).

Results

Through the primary analysis of the eight cohorts and subsequent validation in Chinese cohorts, we identified a 25-gene signature that can predict NASH and liver fibrosis with a high accuracy (ROC ≥ 0.87). Pathway, network and MR analysis revealed 21 metabolites and 12 genes have causal associations with NASH/liver fibrosis. And eventually a 12-gene signature predictive (ROC ≥ 0.75) were validated as a valuable tool for distinguishing Chinese patients with liver fibrosis from those with normal NAFLD or NASH.

Conclusions

This study developed a 12-gene signature for predicting liver fibrosis, demonstrating the utility of an integrated genome-metabolome-Mendelian Randomization approach for predicting disease progression across various databases.

Introduction

Nonalcoholic fatty liver disease (NAFLD) has a global incidence of 32.4% and shows an increasing trend in developing countries [1]. While simple steatosis is generally considered to be a benign ailment, around a quarter of NAFLD patients progress to nonalcoholic steatohepatitis and liver fibrosis. Inflammatory
responses make the liver more vulnerable to various factors of liver injury, which in turn results in abnormal proliferation of intrahepatic connective tissue, thereby promoting liver fibrosis occurrence and progression of liver cancer [2]. Collagen fibrils proliferation and degradation in the liver can be turned back if 'synthesis' is artificially inhibited and 'degradation' is promoted at preliminary stage, where fibrosis or cirrhosis can be reversed early [3].

Early diagnosis through noninvasive serum measures and prompt treatment have the potential to mitigate the pathological process, and are crucial for favorable clinical outcomes in advanced fibrosis cases [4]. However conventional clinical serum markers of NASH can’t reflect the fibrosis degree correctly, thus identifying individuals with a poor prognosis from the general NAFLD population is of utmost clinical importance [5].

Reprogramming at the transcriptional and genomic levels occurs during the progression of fibrosis accompanied by metabolic changes, which is a multi-factor triggered process [6]. Mendelian randomization (MR) analysis, an epidemiological method enhancing causal inference by modeling exposure effects with genetic variants as instrumental variables, stands out for its capacity to shed light on the enduring impact of risk factors on health outcomes. This aspect proves particularly significant in the context of progressive conditions such as non-alcoholic fatty liver disease (NAFLD) progressing to hepatic fibrosis, characterized by extended induction periods [7]. Furthermore, MR-Phenome-wide association studies (MR-PheWAS) offer a comprehensive approach for probing causal relationships between exposures and clinical outcomes, including metabolites, within large-scale biobank settings [8].

In this study we leveraged multiple public databases to identify transcribed genes potentially involved in liver fibrosis and subsequently evaluated their diagnostic efficiency. Pathway analysis and MR analysis were then employed to delve into the intrinsic links between characteristic transcribed genes and metabolites from previous metabolomics studies, and a select few key predictive gene markers were isolated. To corroborate the findings, clinical samples were scrutinized to pinpoint individuals at a heightened risk of liver fibrosis among Chinese subjects. Through this research, we present a 12-gene marker panel for predicting the progression of liver fibrosis in NASH. Our study underscores the significance of a multi-omics, multi-database approach complemented by Mendelian randomization techniques in forecasting the advancement of progressive diseases.

**Methods**

**Study design and sample collection**

To establish a NASH gene signature model with exceptional precision, we leveraged the advanced capabilities of the MetaIntegrator package in R [9]. This method begins by identifying an initial set of genes strongly associated with NASH, using a sophisticated selection process that emphasizes genes with the highest discriminative power. The model iteratively expands this set by adding genes that significantly improve its predictive accuracy, based on a comprehensive evaluation of their contribution to the model's performance metrics. This addition continues until the inclusion of new genes does not
substantially enhance the model's ability to predict NASH outcomes, as determined by a carefully set threshold. This threshold is strategically defined to stop gene addition when the marginal improvement in the model's performance becomes minimal, ensuring an optimal gene set for NASH prediction. And the whole flow chart is conducted as followed and show in Fig. 1.

We systematically searched for expression profiling by array and whole-genome expression datasets that biopsy diagnosed NASH patients. We identified eight datasets and divided them into four discovery cohorts (GSE151158, GSE24807, GSE48452, and GSE63067) and four validation cohorts (SRP295965, SRP308410, SRP186450, SRP217231) from Gene Expression Omnibus (GEO) repository. In discovery cohorts, microarray was performed to identify differentially expressed genes in the NASH and control group (CON VS NASH). From two existing validation cohorts (SRP186450, SRP295965), which originated from RNA sequencing of snap-frozen liver tissue, patients were divided into NAFLD/control and NASH groups (CON/NAFLD VS NASH). Additional two existing validation cohorts (SRP308410, SRP217231), individuals divide into NAFL and NASH groups (NAFL VS NASH). In two predictive cohorts of Chinese patients, blood samples were collected after centrifugation and stored at −80°C until analyzed.

Then pathway, network analysis and MR analysis was applied to explore the intrinsic association between characteristic transcribed genes and metabolites from previous metabolomics studies. And there by a simplified signature diagnostic of NASH prediction model was developed and validated.

Criteria of predictive cohort

This study was conducted as a retrospective cohort study; Patients with NAFLD diagnosed in Jiangjin District Central Hospital between 2016 and 2021 were enrolled. The Inclusion and Exclusion criteria were as follows: (1) Alcohol intake of less than 20 g/d. (2) Individuals presence of fatty liver on abdominal ultrasound. (3) Patients with any other systemic inflammatory diseases, viral hepatitis type B or C infection, alcoholic hepatitis, or autoimmune hepatitis were excluded. A total of 113 plasma samples were obtained from 51 (the first independent validation dataset) and 62 (the second independent validation dataset) patients with NAFLD at Jiangjin District Central Hospital, an affiliate of the Wuhan University of Science and Technology. Oral consent was obtained from patients, and written informed consent was waived.

Follow-up evaluation: The Finnish cohorts consisted of 113 patients. The presence of liver fibrosis was confirmed by liver biopsy. The clinical characteristics of the two cohorts are shown in Supplementary Table 1. A total of 38 (74.5%, 51) patients with advanced fibrosis were involved in the independent validation of one study. The second independent validation dataset included 47 (75.8%, 62) patients with NASH (Supplementary Table S1).

RNA extraction and gene expression quantification

We extracted RNA from plasma using Trizol reagent (Invitrogen) and assessed its quality with an Agilent 2100 Bioanalyzer. RNA-seq libraries were prepared using the TruSeq RNA Sample Preparation Kit
(Illumina) and sequenced on an Illumina HiSeq 2500 with 150bp paired-end reads.

For data analysis, we filtered raw reads using Trimmomatic, aligned them to the human genome (GRCh38) with HISAT2, and quantified gene read counts using featureCounts. Differential expression was analyzed using DESeq2 in R, considering genes with an adjusted p-value < 0.05 and an absolute log2 fold change $\geq 1$ as significantly differentially expressed [10].

**Differentially expressed metabolites identification of liver fibrosis**

We compared metabolic changes between human liver fibrosis patients and a control group. First of all, articles were collected related to metabolic characteristics associated with human liver fibrosis from databases such as PubMed and Google Scholar, and then five research articles were included [11-15]. Our research encompassed several steps: First, we examined studies that compared metabolite levels in serum samples between liver fibrosis patients and controls, analyzed the data and changes using MS-based techniques, and provided our interpretations; Second, study information and metabolite data were extracted from the selected studies, including recruitment area, sample size, mean age, percent of females, disease severity measure, and the analytical platform used. Third, curated metabolite data encompassed names and statistical information, and p-values from t-tests and fold change were recorded. We used R-studio software (R Core Team, 2022) to conduct Meta-analyses [16]. Metabolites appearing in at least two different datasets were selected for analysis, and statistical analysis was performed based on the original data provided in the literatures (Supplementary Table S2).

**Pathway enrichment analysis and network analysis**

Although there was considerable clinical heterogeneity, we sought a high diagnostic performance. It was essential to maximize the performance of the diagnostics. A standardized enrichment score was then calculated using the same method as the z score \[\text{Mean (upregulated genes)} - \text{Mean (downregulated genes)}\]. DisGeNET enrichment, GO enrichment analysis (MF), and GO enrichment analysis (BP) were used here for pathway enrichment analysis (Table 1).

Pathway and network analyses were employed, to uncover the biological functions of different metabolites and their interactions with genes and with each other. Detailed methods can be found in the Supplementary Materials. Altered metabolic pathways were identified using MetaboAnalyst 5.0, whose pathway analysis currently supports 15 model organisms with $\sim$1,200 precompiled Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [17]. Furthermore, network analysis was performed to construct molecular networks based on interactions among input metabolites and relevant genes. A pathway with a p-value less than 0.05 in both the metabolic pathway analysis and network analysis was considered to be significantly enriched.

**Method of MR study design**

MR analysis is a genetic method employed to infer causal effects by utilizing the random allocation of genetic variants at conception. The Single Nucleotide Polymorphisms (SNPs) serving as Instrumental
Variables (IVs) in this study satisfied the three specific assumptions: 1) a robust association between the SNPs and the exposure; 2) not connect to the outcome through confounding variables; 3) don't exert a direct impact on the outcome. The design of this research adhered to the STROBE-MR guidelines [18].

Gene, metabolites, and disease sources

Genetic instruments for Gene, metabolites, and NASH were obtained from the GWAS Catalog [19]. IEU open GWAS project, FinnGen, and UKBiobank. All eQTL data of the target gene were obtained from IEU open GWAS project using Gene stable ID (Supplementary Table S3) [20]. Metabolite sources were derived from large GWASs, encompassing 1,091 metabolites and 309 metabolite ratios in 8,299 individuals from the Canadian Longitudinal Study on Aging (CLSA) cohort, with plasma samples analyzed by Metabolon using ultrahigh performance liquid chromatography–tandem mass spectrometry (UPLC–MS/MS) (Supplementary Table S4) [21]. NASH and liver fibrosis GWAS data were obtained from two GWAS databases, including 3,108 UK Biobank participants for Fibrosis and cirrhosis of liver with Sex-specific and three phenotype named Fibrosis and cirrhosis of liver (368,291 European), Fibrosis of liver (373,453 European) and Nonalcoholic steatohepatitis (377,277 European) from FinnGen (Supplementary Table S5) [22]. The GWAS summary data in UK Biobank are publicly available at https://www.nealelab.is/uk-biobank.

Genetic instrument selection

To adhere to the MR assumption, instrumental variables (IVs) were chosen with a linkage disequilibrium threshold of $r^2 < 0.01$ and a distance $> 1000$ kb, ensuring genome-wide significance ($p < 1.0 \times 10^{-11}$). The linkage disequilibrium reference panel was constructed using the 1000 Genomes Project European sample. The strength of each IV was assessed using the F-statistic $= \beta^2 / se^2$, with a threshold set at $>10$ to ensure adequate strength.

Statistical analysis

The primary statistical analysis method was the inverse variance weighted (IVW) method under random effects. This method was supplemented with weighted median analysis [23], MR-Egger regression and MR-PRESSO methods [24]. IVW assumes that all genetic variation SNPs are valid IVs with an overall bias of zero. As for the weighted median analysis, this estimate is consistent even if up to half of the weights are from invalid instruments. Besides, MR-Egger analysis can identify horizontal pleiotropy through the intercept ($p < 0.05$ for the intercept indicates pleiotropy) [25]. The MR-PRESSO method can detect possible outliers and generate causal estimates after the removal of outlying IVs [26]. To measure the degree of heterogeneity, the Q value from Cochrane was applied. The causal relationship is considered significant if: 1) the p-value of the IVW method is less than $< 0.05$; 2) the estimations obtained using the MR-Egger, weighted median, and IVW methods all have the same direction; and 3) neither the MR-Egger intercept test nor the MR-PRESSO global test has statistical significance ($p > 0.05$) [27]. Each test was two-sided and conducted utilizing the Two-Sample MR and MR-PRESSO packages in the R software (version 4.3.1) [28].
Results

The 25-gene signature diagnostic of NASH and liver fibrosis in existing discovery, validation cohorts and Chinese cohorts

As described in the methods, eight datasets were identified and divided into four discovery and validation cohorts (Fig. 1). Based on the multi-cohort analysis, twenty-five genes, including CCL23, RPL19, KRT10, IQCH-AS1, NDUFA2, CBS, IFIT1, RPS7, DNAJC12, STXB6, SULT1A2, ANKRD37, GNAO1, KRT8, PKMYT1, ARHGAP9, ANXA5, DECR1, TNFRSF14, CDCP1, KIAA1522, STRADA, LGALS3, WARS2, NPM2, were significantly differentially expressed between patients who progress to NASH and CON/NAFLD in discovery and validation cohorts (false discovery rate [FDR] < 40%). We calculated the liver fibrosis score for each sample and meta-scores for each cohort according to the criteria in Methods (Fig. 2A). The Liver fibrosis scores differentiated NASH from CON/NAFLD by the receiver operating characteristic curve value (ROC) shown as a summary area under the curve (AUC) = 0.95 with 95% confidence interval (CI) [0.79–0.99] in the discovery cohorts (Fig. 2B). Then we validated the Liver fibrosis scores using the validation cohorts, the liver fibrosis score of the representative genes are shown in Fig. 3A, the others are in Supplementary Fig. S1. The Liver fibrosis scores differentiated NASH from NAFLD with a summary area under the curve (AUC) = 0.9 with 95% CI [0.86–0.94] in the validation cohorts (Fig. 3B).

The Liver fibrosis scores were further validated by collecting plasma from Chinese patients with NASH or liver fibrosis (Fig. 3A). In the validation 2 (NASH and NAFLD group), the Liver fibrosis scores differentiated NASH from NAFLD with AUC=0.94, 95% CI [0.89–0.98]. Next, we evaluated the predictive power of the Liver fibrosis scores in independent validation 1 (Liver fibrosis cohort) with AUC= 0.84, 95% CI [0.77–0.91] (Fig. 3B).

Pathway analysis of the 25-gene signature predictive of liver fibrosis

To investigate the functions of these genes, we performed Gene Ontology (GO) term enrichment analysis and DisGeNET enrichment analysis (Fig. 4A). In DisGeNET enrichment analysis, the most significantly enriched pathway was Liver Failure. The most significant GO term was NADP binding (Molecular Function, BP) and monocyte chemotaxis (Biological Process, BP). The NAD (P)-binding is the most enriched pathway associated with early liver fibrosis diagnosis. Then some other pathways were listed, such as oxidoreductase activity acting on other nitrogenous compounds as donors, nitrite reductase (NO-forming) activity, carbon monoxide binding, nitrite reductase (NO-forming) activity cystathionine beta-synthase activity, 2,4-dienoyl-CoA reductase (NADPH) activity. Tryptophan-tRNA ligase activity, mitochondrial tryptophanyl-tRNA aminoacylation, and tryptophanyl-tRNA aminoacylation were also identified in GO enrichment. As shown in our GO-KEGG results (Table 1), CBS is Oxidoreductase activity, acting on other nitrogenous compounds as donors, cytochrome as acceptor, Nitrite reductase (NO-forming) activity, Carbon monoxide binding and Nitrite reductase activity, and also cystathionine beta-synthase activity. NADP binding, 2, 4-dienoyl-CoA reductase (NADPH) activity, oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor were attributed to DECR1.
**Differentially expressed metabolites associated with liver fibrosis and network analysis**

A total of 63 metabolites, which were repeated at least twice across the five different datasets, were used for the differential metabolite identification (Supplementary Table S2). Among them, some frequently occurring metabolites, especially amino acids, most are amino acids, were found to be significantly differentially regulated between the liver fibrosis and control groups (Fig. 4C). Patients were characterized by uniformly higher or lower levels of asymmetric metabolites. For instance, Tryptophan, Asparagine, Methionine, Phenylalanine, Tyrosine, Threonine consistently manifest elevated levels across the first three articles under scrutiny [11-13]. This uniform upregulation not only signifies their importance but also alludes to their pivotal role in the context being explored.

Contrastingly, Valine's marked downregulation in the first three articles accentuates its potential reduced role or consumption under the studied conditions. While the other two branched-chain amino acids (BCAAs), Isoleucine and Leucine showed different and inconsistent changes. Other metabolites showed inconsistent trends across studies. BCAAs have been associated with the improvement of protein malnutrition and have shown potential to reduce the risk of hepatocellular carcinoma in patients with cirrhosis [29].

The network analysis results revealed associations between key genes and many amino acids, as DECR1 concurrently links to lysine, glutamate, methionine, arginine etc. And CBS gene is associate with serine, taurine and methionine (Fig. 4C, Supplementary Fig. S2).

**Mendelian randomization (MR) analysis**

To explore the potential causal links of selected Gene/Metabolite to NASH, we further conducted Mendelian randomization (MR) analysis utilizing single nucleotide polymorphism (SNP) data from four well-known GWAS databases. The analysis encompassed 3 parts: 1) to explore the causal link of metabolites and genes, with each of the selected 49 metabolites as the outcome and each of the 25 target genes as the exposure (Supplementary Table S6); 2) to explore the causal link of genes and NASH, with genes and NASH/Liver fibrosis as exposure and outcome respectively (Table 2); 3) to explore the causal link of metabolites and NASH, with the 49 metabolites as the exposure and NASH/Liver fibrosis as outcome(Supplementary Table S7). Eventually, 21 metabolites and 12 genes were identified with a causal association with NASH/liver fibrosis in a directed network diagram (Fig. 4B).

We found that three genes KIAA1522 (OR:0.90, 95%CI:0.83~0.98, p=0.01), KRT8(OR:5.52, 95%CI:2.12~14.38, p=0.02), and PKMYT1(OR:1.00, 95%CI:1.00~1.00, p=0.04) have direct causal relationships with NASH and liver fibrosis. Eleven out of all the 12 genes (except for KRT8) show causal relationships with NASH and liver fibrosis through the mediation of metabolites. Notably, WARS2 has a negative causal relationship with Glutamine (β=−0.1±0.08, p=0.04), which has an inhibitory effect on liver fibrosis (OR: 0.87, 95%CI: 0.79~0.97, p<0.01). The CBS gene may lead to the occurrence of liver fibrosis by upregulating Margarate (β=0.13±0.06, p=0.04) and downregulating Leucine (β=−0.15±0.06, p=0.02). Interestingly, the DECR1 gene may has dual effects on NASH or liver fibrosis. It may affect the levels of
Linolenate (β=0.06±0.03, p=0.02) and Alanine (β=-0.06±0.03, p=0.03), which are potential promoters of NASH (OR: 3.35, 95%CI: 1.09~10.29, p=0.03 for Linolenate) and liver fibrosis (OR: 1.62, 95%CI: 1.09~2.42, p=0.03 for Alanine). Meanwhile, it may lead to an increase of the levels of Methionine (β=-0.07±0.03, p<0.01), which has an inhibitory effect on liver fibrosis (OR: 0.05, 95%CI: 0.01~0.35, p<0.01).

**Network analysis of key genes and metabolites**

Network analysis using MetaboAnalyst showed that the genes DECR1 and CBS had the highest degree and betweenness centrality, respectively. DECR1 interacted with 10 metabolites, while CBS interacted with 3 metabolites. Methionine is associated with both of the two genes (Fig. 4C, Supplementary Fig.S2).

**The 12-gene signature predictive of NASH and liver fibrosis with Pathway analysis**

The twelve genes including KRT10, NDUFA2, CBS, IFIT1, GNAO1, KRT8, PKMYT1, DECR1, TNFRSF14, KIAA1522, STRADA and WARS2, were significantly differentially expressed between patients who progress to NASH and CON/NAFLD in all the discovery and validation cohorts (Fig. 5A). The Liver fibrosis scores differentiated NASH from NAFLD with a summary AUC=0.8, 95% CI [0.75−0.85] in the four SRP cohorts. Next, we evaluated the predictive power of the Liver fibrosis scores in the two Chinese independent Liver fibrosis cohorts (Validation 2 AUC=0.8, 95% CI [0.73−0.87], Validation 1 AUC=0.75, 95% CI [0.67−0.83]), with a summary AUC= 0.76, 95% CI [0.63−0.86] (Fig. 5B).

**Discussion**

Liver fibrosis, actually a kind of hepatic scar tissue caused by chronic inflammation, leading to liver cirrhosis, hepatocellular carcinoma, and even death. Early detection and accurate diagnosis are crucial but there are no blood markers other than liver biopsy [30]. In this study, we firstly identified 25 differentially expressed transcriptional genes between the NASH and control groups (Fig. 2, 3). Then pathway analysis was conducted to gain a deeper understanding of the interactions involving these genes (Fig. 4A, Table 1). In terms of amino acids, several anabolic pathways and amino acids themselves are implicated in liver fibrosis prediction. Elevated tryptophan levels observed across the first three metabolomics studies potentially contribute to the tryptophanyl-tRNA aminoacylation, a critical indicator of impending hepatic fibrosis (Table S2). So the WARS2 gene, coding mitochondrial tryptophanyl-tRNA synthetase, was reported leading to severe hepatopathy with WARS2 deficiency [31]. It is sure found in GO and also has a negative causal relationship with glutamine, which has an inhibitory effect on liver fibrosis (Fig. 4B).

The CBS gene, encoding cystathionine-β, plays a role in the first step of the transsulfuration pathway of homocysteine, converting homocysteine to cystathionine. This catabolic route facilitates the elimination of L-methionine and the toxic metabolite L-homocysteine [32]. In addition, homocysteinemia, cystathionine beta – synthase activity, and cysteine biosynthetic processes from serine were identified.
by DisGeNET and GO enrichment analysis (Fig. 4A, Table 1). Higher homocysteine level have been associated with the severity of insulin resistance in NAFLD patients [33].

Homocysteine/cysteine pathway might be a critical predictive factor for liver fibrosis. Here we also observed significantly increased methionine levels in the first three metabolomics studies (Table S2). Furthermore, it has been reported that Hepatic sulfur amino acid metabolism is a key determinant of sensitivity to methionine restriction and chemotherapy in liver cancer [34]. Through the transsulfuration pathway, methionine is converted to downstream products including taurine [35]. Glutamate was shown to induce suppression in the expression and activity of cystathionine-β-synthetase (CBS) [36]. CBS is highly expressed in liver, and deficient expression of CBS has been observed to cause hepatic steatosis, inflammation, and fibrosis in animal models [37]. In this study, we have identified associations between CBS and glutamate, methionine, and taurine (Fig. 4B, C).

From our GO-KEGG results (Table 1), NADP binding, 2, 4-dienoyl-CoA reductase (NADPH) activity were attributed to DECR1, which is the key enzyme in fatty acid metabolic pathways. DECR1 is involved in maintaining REDOX homeostasis by controlling the balance between saturated and unsaturated phospholipids, and in vivo [38]. The enzyme DECR1 acts on unsaturated fatty enoyl-CoA esters and its absence is closely related to a hepatic steatosis [39]. In MR analysis, DECR1 shows causal relationships with alanine and methionine. NADP serves as a substrate for DECR1 in lysine degradation and beta-oxidation of fatty acids [40]. Intriguingly, glutamate and aspartic acid participate Urea Cycle (arginine, citrulline and ornithine), which is associated with progression mitochondrial dysfunction and mild versus severe fibrosis changes in NAFLD [41].

Branched-chain amino acids (BCAAs), including leucine, isoleucine, and valine, derived from dietary protein, are believed to be closely associated with liver health. BCAAs have been identified to play a role in glucose homeostasis and lipid metabolism, both of which are crucial for liver health [42]. Among the three studies, only the valine level consistently decreased, indicating the other two BCAAs may change differently with disease progression. Beside, other amino acids from diet protein, including phenylalanine, methionine, threonine, tryptophan, were all significantly upregulated in liver fibrosis groups across the three studies definitively, suggesting the degradation process of essential amino acids might be hindered in fibrotic liver (Supplementary Table S2). Moreover, the spotlight has now shifted to the metabolic axis encompassing serine, glycine and threonine. As supported by seminal works, augmenting this axis with specific fasting regimens appears to have the potential to enhance longevity [43]. From results of MR analysis, serine and glycine were found to have a causal association with GNAO1 and NDUFA2 respectively. Threonine is the only metabolite which was consistently upregulated in liver fibrosis groups across three metabolomics studies, and also a metabolite possessing the strongest causal association with NASH (Fig. 4B).

IFIT1 gene, encoding a protein known for its role in response to interferon treatment, has been associated with liver conditions such as hepatitis C-mediated cirrhosis and liver cancer [44]. IFIT1 was reported to possessed a significant inhibitory effect on alleviating oxidative stress damage, and
mediating liver cell apoptosis [45]. From our MR analysis, IFIT1 possesses the highest bold arrow line to alanine and the latter to live fibrosis, suggesting IFIT1 might have a signature response to hepatic lesions (Fig. 4B).

As for PKMYT1 gene, a newly published article showed it exhibited higher expression in human HCC tissues compared to adjacent tissues [46]. GNAO1 gene was another simultaneously captured by both GO and MR, being verified to be strongly related to the largest number of metabolites (Table 1, Fig. 4). Previous study explained that GNAO1 is not only highly associated with hepatoma carcinoma cells but also a reliable biomarker of relapse prediction for HCC [47, 48].

KIAA1522 gene was reported an increased expression in HCC tissues related to worse overall survival [49]. Our MR results demonstrated it has the highest causal association with fibrosis and cirrhosis among all the twelve genes. Besides, it has a strong associations with histidine and a weaker link with creatinine, both of which participate in Urea Cycle through arginine, which also associated with DECR1 and STRADA (Fig. 4B, C).

Overall, the above optimized results with network and MR analysis establish a 12-genes liver fibrosis diagnosis model which successfully predicts the progression to fibrosis from interracial Chinese cohorts (Fig. 5B), confirming the robustness of these gene markers, providing valuable insights for further potential diagnostic applications.

**Conclusion**

This study identified noninvasive 12-gene signature serological biomarkers associated with the progression to liver fibrosis. Pathway, network and MR analysis revealed several key genes like DECR1, GNAO1 and CBS, were closely associated with amino acid metabolism. This integrated approach may facilitate the development of biomarkers for predicting the fibrosis stage with highly potential clinical value in multiple regions of the world.

**Declarations**

The authors have no competing interests to declare that are relevant to the content of this article.

**Acknowledgements**

The IEU OpenGWAS and NHGRI-EBI GWAS catalog Database are databases of publicly available datasets, the University of Helsinki is the organization responsible for the FinnGen Project, GWAS data of UKBiobank were published by Neale lab (http://www.nealelab.is/uk-biobank), and each study included in it was approved by the local institutional review board and ethics committee. The authors thank all investigators for sharing these data.

**Author contributions**
All authors read and approved the final version of the manuscript. X. Wang and Y. Deng had designed and launched the whole study. S. Chen performed the open access datasets analysis; L. Zhang performed clinical samples collection and RNA sequencing; T. Chen performed gene data validation, composed the manuscript of method and results; Y. Chang processed metabolomics data; J. Teng performed pathway analysis; G. Yang performed network analysis; W. Xie performed network construction; Y. Guo and M. Zhu performed MR analysis and construction. X. Wang and W. Jia composed the manuscript of the metabolomics and network section; J. Shen polish language of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (Grant no. 2022YFC2407003), Interdisciplinary Program of Shanghai Jiao Tong University (YG2022QN015).

Data sharing statement

The RNA-seq raw data supporting this study are available from the Gene Expression Omnibus (GEO), including the following datasets: GSE151158, GSE24807, GSE48452, GSE63067, SRP295965, SRP308410, SRP186450, and SRP217231.

All GWAS data are publicly available as summary-level data and can be obtained from their respective websites: IEU OpenGWAS project (https://gwas.mrcieu.ac.uk/), NHGRI-EBI GWAS catalog (https://www.ebi.ac.uk/gwas/), FinnGen (https://www.finngen.fi/en), UKBiobank (http://www.nealelab.is/uk-biobank).

References


Tables

Table 1. The top of the list of GO enrichment analysis results in the 25 genes
<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>BgRatio</th>
<th>pvalue</th>
<th>p.adjust</th>
<th>qvalue</th>
<th>geneID</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0016662</td>
<td>oxidoreductase activity, acting on other nitrogenous compounds as donors, cytochrome as acceptor</td>
<td>1/18352</td>
<td>0.001</td>
<td>0.043</td>
<td>0.028</td>
<td>CBS</td>
</tr>
<tr>
<td>GO:0050421</td>
<td>nitrite reductase (NO-forming) activity</td>
<td>1/18352</td>
<td>0.001</td>
<td>0.043</td>
<td>0.028</td>
<td>CBS</td>
</tr>
<tr>
<td>GO:0070025</td>
<td>carbon monoxide binding</td>
<td>1/18352</td>
<td>0.001</td>
<td>0.043</td>
<td>0.028</td>
<td>CBS</td>
</tr>
<tr>
<td>GO:0098809</td>
<td>nitrite reductase activity</td>
<td>1/18352</td>
<td>0.001</td>
<td>0.043</td>
<td>0.028</td>
<td>CBS</td>
</tr>
<tr>
<td>GO:0050661</td>
<td>NADP binding</td>
<td>54/18352</td>
<td>0.002</td>
<td>0.043</td>
<td>0.028</td>
<td>DECR1</td>
</tr>
<tr>
<td>GO:0004122</td>
<td>cystathionine beta-synthase activity</td>
<td>2/18352</td>
<td>0.003</td>
<td>0.043</td>
<td>0.028</td>
<td>CBS</td>
</tr>
<tr>
<td>GO:0004830</td>
<td>tryptophan-tRNA ligase activity</td>
<td>2/18352</td>
<td>0.003</td>
<td>0.043</td>
<td>0.028</td>
<td>WARS2</td>
</tr>
<tr>
<td>GO:0008670</td>
<td>2,4-dienoyl-CoA reductase (NADPH) activity</td>
<td>2/18352</td>
<td>0.003</td>
<td>0.043</td>
<td>0.028</td>
<td>DECR1</td>
</tr>
<tr>
<td>GO:0047894</td>
<td>flavonol 3-sulfotransferase activity</td>
<td>3/18352</td>
<td>0.004</td>
<td>0.050</td>
<td>0.032</td>
<td>SULT1A2</td>
</tr>
<tr>
<td>GO:0070026</td>
<td>nitric oxide binding</td>
<td>3/18352</td>
<td>0.004</td>
<td>0.050</td>
<td>0.032</td>
<td>CBS</td>
</tr>
<tr>
<td>GO:0031852</td>
<td>mu-type opioid receptor binding</td>
<td>4/18352</td>
<td>0.005</td>
<td>0.050</td>
<td>0.032</td>
<td>GNAO1</td>
</tr>
<tr>
<td>GO:0051430</td>
<td>corticotropin-releasing hormone receptor 1 binding</td>
<td>4/18352</td>
<td>0.005</td>
<td>0.050</td>
<td>0.032</td>
<td>GNAO1</td>
</tr>
<tr>
<td>GO:0016661</td>
<td>oxidoreductase activity, acting on other nitrogenous compounds as donors</td>
<td>5/18352</td>
<td>0.007</td>
<td>0.050</td>
<td>0.032</td>
<td>CBS</td>
</tr>
<tr>
<td>GO:0070402</td>
<td>NADPH binding</td>
<td>16/18352</td>
<td>0.021</td>
<td>0.089</td>
<td>0.058</td>
<td>DECR1</td>
</tr>
<tr>
<td>GO:1904047</td>
<td>S-adenosyl-L-methionine binding</td>
<td>16/18352</td>
<td>0.021</td>
<td>0.089</td>
<td>0.058</td>
<td>CBS</td>
</tr>
<tr>
<td>GO:0016628</td>
<td>oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor</td>
<td>26/18352</td>
<td>0.033</td>
<td>0.104</td>
<td>0.067</td>
<td>DECR1</td>
</tr>
<tr>
<td>GO:0019825</td>
<td>oxygen binding</td>
<td>36/18352</td>
<td>0.046</td>
<td>0.125</td>
<td>0.080</td>
<td>CBS</td>
</tr>
</tbody>
</table>

Table 2. Causal effects of Gene-to-NASH or liver fibrosis
<table>
<thead>
<tr>
<th>Exposure</th>
<th>Outcome</th>
<th>Method</th>
<th>nSNP</th>
<th>Beta</th>
<th>SE</th>
<th>P</th>
<th>Het</th>
<th>Ple</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIAA1522</td>
<td>Nonalcoholic fatty liver disease</td>
<td>MR Egger</td>
<td>23</td>
<td>-0.20</td>
<td>0.084</td>
<td>0.027</td>
<td>0.438</td>
<td>0.215</td>
</tr>
<tr>
<td>KIAA1522</td>
<td>Nonalcoholic fatty liver disease</td>
<td>Weighted median</td>
<td>23</td>
<td>-0.12</td>
<td>0.059</td>
<td>0.040</td>
<td>0.438</td>
<td>0.215</td>
</tr>
<tr>
<td>KIAA1522</td>
<td>Nonalcoholic fatty liver disease</td>
<td>IVW</td>
<td>23</td>
<td>-0.11</td>
<td>0.042</td>
<td>0.012</td>
<td>0.438</td>
<td>0.215</td>
</tr>
<tr>
<td>KRT8</td>
<td>Nonalcoholic steatohepatitis</td>
<td>MR Egger</td>
<td>7</td>
<td>1.71</td>
<td>0.489</td>
<td>0.017</td>
<td>0.039</td>
<td>0.019</td>
</tr>
<tr>
<td>NDUFA2</td>
<td>Liver fat</td>
<td>Weighted median</td>
<td>4</td>
<td>-0.05</td>
<td>0.026</td>
<td>0.048</td>
<td>0.832</td>
<td>0.886</td>
</tr>
<tr>
<td>NDUFA2</td>
<td>Liver fat</td>
<td>IVW</td>
<td>4</td>
<td>-0.05</td>
<td>0.025</td>
<td>0.049</td>
<td>0.832</td>
<td>0.886</td>
</tr>
<tr>
<td>WARS2</td>
<td>Nonalcoholic fatty liver disease</td>
<td>MR Egger</td>
<td>11</td>
<td>0.21</td>
<td>0.066</td>
<td>0.012</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>NDUFA2</td>
<td>Nonalcoholic fatty liver disease</td>
<td>Weighted median</td>
<td>4</td>
<td>-0.13</td>
<td>0.054</td>
<td>0.013</td>
<td>0.912</td>
<td>0.608</td>
</tr>
<tr>
<td>NDUFA2</td>
<td>Nonalcoholic fatty liver disease</td>
<td>IVW</td>
<td>4</td>
<td>-0.12</td>
<td>0.050</td>
<td>0.014</td>
<td>0.912</td>
<td>0.608</td>
</tr>
<tr>
<td>KIAA1522</td>
<td>Diagnoses - main ICD10: K74 Fibrosis and cirrhosis of liver_both_sexes</td>
<td>IVW</td>
<td>24</td>
<td>0.00</td>
<td>0.000</td>
<td>0.024</td>
<td>0.223</td>
<td>0.575</td>
</tr>
<tr>
<td>PKMYT1</td>
<td>Diagnoses - main ICD10: K74 Fibrosis and cirrhosis of liver_both_sexes</td>
<td>Weighted median</td>
<td>16</td>
<td>0.00</td>
<td>0.000</td>
<td>0.039</td>
<td>0.139</td>
<td>0.256</td>
</tr>
<tr>
<td>PKMYT1</td>
<td>Diagnoses - main ICD10: K74 Fibrosis and cirrhosis of liver_both_sexes</td>
<td>IVW</td>
<td>16</td>
<td>0.00</td>
<td>0.000</td>
<td>0.037</td>
<td>0.139</td>
<td>0.256</td>
</tr>
<tr>
<td>PKMYT1</td>
<td>Diagnoses - main ICD10: K74 Fibrosis and cirrhosis of liver_both_sexes</td>
<td>Weighted mode</td>
<td>16</td>
<td>0.00</td>
<td>0.000</td>
<td>0.039</td>
<td>0.139</td>
<td>0.256</td>
</tr>
<tr>
<td>PKMYT1</td>
<td>Diagnoses - main ICD10: K74 Fibrosis and cirrhosis of liver_female</td>
<td>IVW</td>
<td>16</td>
<td>0.00</td>
<td>0.000</td>
<td>0.033</td>
<td>0.138</td>
<td>0.891</td>
</tr>
<tr>
<td>KIAA1522</td>
<td>Diagnoses - main ICD10: K74 Fibrosis and cirrhosis of liver_male</td>
<td>Weighted median</td>
<td>24</td>
<td>0.00</td>
<td>0.000</td>
<td>0.010</td>
<td>0.519</td>
<td>0.859</td>
</tr>
<tr>
<td>Gene</td>
<td>Diagnosis</td>
<td>Study Design</td>
<td>nSNP</td>
<td>Het P value</td>
<td>Ple P value</td>
<td>Note: nSNP, number of SNPs fit MR’s three hypotheses; Het, Heterogeneity test P value; Ple, Horizontal Pleiotropy test P value.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------</td>
<td>--------------</td>
<td>------</td>
<td>-------------</td>
<td>-------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIAA1522</td>
<td>Diagnoses - main ICD10: K74 Fibrosis and cirrhosis of liver_male</td>
<td>IVW</td>
<td>24</td>
<td>0.00</td>
<td>0.000</td>
<td>0.011 0.519 0.859</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIAA1522</td>
<td>Diagnoses - main ICD10: K74 Fibrosis and cirrhosis of liver_male</td>
<td>Weighted mode</td>
<td>24</td>
<td>0.00</td>
<td>0.000</td>
<td>0.023 0.519 0.859</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRT8</td>
<td>Nonalcoholic fatty liver disease_both sexes</td>
<td>MR Egger</td>
<td>7</td>
<td>0.00</td>
<td>0.000</td>
<td>0.042 0.175 0.049</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANXA5</td>
<td>Nonalcoholic fatty liver disease_male</td>
<td>Weighted median</td>
<td>10</td>
<td>0.00</td>
<td>0.000</td>
<td>0.015 0.712 0.387</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANXA5</td>
<td>Nonalcoholic fatty liver disease_male</td>
<td>IVW</td>
<td>10</td>
<td>0.00</td>
<td>0.000</td>
<td>0.027 0.712 0.387</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANXA5</td>
<td>Nonalcoholic fatty liver disease_male</td>
<td>Weighted mode</td>
<td>10</td>
<td>0.00</td>
<td>0.000</td>
<td>0.030 0.712 0.387</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: nSNP, number of SNPs fit MR’s three hypotheses; Het, Heterogeneity test P value; Ple, Horizontal Pleiotropy test P value.

**Figures**

**Figure 1**

Flowchart of the study
Figure 2

Discovery of the 25-Gene signature Predictive of Liver Fibrosis

A: Illustration of the 25 genes in all the four discovery cohorts

B: A violin plot illustrating the power of the 25-gene signature for separating NASH from NAFLD in one of the discovery cohorts; and A ROC curve illustrating the power of 25-gene signature for separating NASH from NAFLD in all discovery cohorts.
Figure 3

Validation of the 25-Gene signature Predictive of Liver Fibrosis

A: Illustration of representative genes in all the validation cohorts (The other shown in Supplementary Fig. S1)

B: A violin plot illustrating the power of the 25-gene signature for separating NASH from NAFLD in one of the validation cohorts; A ROC curve illustrating the power of 25-gene signature for separating NASH from NAFLD in all validation cohorts; and A ROC curve illustrating the power of 25-gene signature for predicting liver fibrosis from NAFLD in the two independent cohorts.
Figure 4

Function enrichment analysis and Gene-Metabolite association

A: DisGeNET enrichment analysis of the 25-gene signature; GO enrichment analysis (molecular function) of the 25-gene signature; GO enrichment analysis (biological process) of the 25-gene signature

B: Causal Associations of Genes, Metabolites and NASH Diseases by MR Analysis

C: Association of key transcriptional genes with amino acid metabolites
Figure 5

Validation of the 12-Gene signature Predictive of Liver Fibrosis

A: Illustration of the 12 genes in all the cohorts
B: A ROC curve illustrating the power of 12-gene signature for predicting Liver Fibrosis from NAFLD in the two independent cohorts.

Supplementary Files

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

- SIfigtableforJTM.docx