Electronic Health Record-Based Genome-Wide Meta-Analysis and Mendelian Randomization Identify Metabolic and Phenotypic Consequences of Non-Alcoholic Fatty Liver Disease

Nooshin Ghodsian
Centre de recherche de l'Institut universitaire de cardiologie et de pneumologie de Québec
Erik Abner
University of Tartu
Connor A. Emdin
Broad Institute
Émilie Gobeil
universitaire de cardiologie et de pneumologie de Québec
Nele Taba
University of Tartu
Mary J. Haas
Broad Institute
Nicolas Perrot
l'Institut universitaire de cardiologie et de pneumologie de Québec
Hasanga D. Manikpurage
Université Laval
Éloi Gagnon
Université Laval
Jérôme Bourgault
Université Laval
Alexis St-Amand
l'Institut universitaire de cardiologie et de pneumologie de Québec
Christian Couture
l'Institut universitaire de cardiologie et de pneumologie de Québec
Patricia Mitchell
l'Institut universitaire de cardiologie et de pneumologie de Québec
Yohan Bossé
Department of Molecular Medicine, Laval University https://orcid.org/0000-0002-3067-3711
Patrick Mathieu
Laboratory of Cardiovascular Pathobiology, Quebec Heart and Lung Institute/Research Center, Department of Surgery, Laval University, Quebec https://orcid.org/0000-0002-3805-2004
Marie-Claude Vohl
Université Laval
André Tchernof

Université Laval

Sébastien Thériault

Institut universitaire de cardiologie et de pneumologie de Québec-Université Laval, Quebec City

https://orcid.org/0000-0003-1893-8307

Amit V. Khera

Broad Institute

Tõnu Esko

University of Tartu

Benoit Arsenault (benoit.arsenault@criucpq.ulaval.ca)

Department of Medicine, Laval University, Quebec

Article

Keywords: Electronic Health Record, Non alcoholic Fatty Liver Disease, Mendelian randomization blood biomarkers and chronic diseases

Posted Date: March 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-97977/v2

Abstract

Non-alcoholic fatty liver disease (NAFLD) is a complex disease linked with several chronic diseases. We aimed at identifying genetic variants associated with NAFLD as well as blood biomarkers that may be causally impacted by NAFLD. We performed a genome-wide meta-analysis of four cohorts of electronic health record-documented NAFLD (8434 cases and 770,180 controls) and confirmed known susceptibility loci (*GCKR*, *MAU2/ TM6SF2*, *APOE*, and *PNPLA3*). We also identified potentially new loci (*LPL*, *FTO* and *TR1B1*) and report an effect of lower *LPL* expression in adipose tissue on NAFLD susceptibility. Mendelian randomization analyses identified an effect of NAFLD on tyrosine metabolism and on blood levels of three proteins. Positive genetic correlations between NAFLD and cardiometabolic traits and negative genetic correlations with parental lifespan, socio-economic factors and ketone bodies were observed. Altogether, this analysis revealed novel susceptibility loci for NAFLD and early biomarkers of NAFLD that could be used to identify patients with NAFLD.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is one of the most prevalent chronic liver diseases.[1,2] According to recent estimates, about 25% of the adult population worldwide may have NAFLD.[3,4] This disease has been predicted to become the most frequent indication for liver transplantation in Western countries by 2030.[5] NAFLD is a progressive liver disease with potential consequences for several other chronic disorders such as cardiovascular disease (CVD) (the leading cause of death in patients with NAFLD),[6-9] type 2 diabetes (T2D),[10,11] dyslipidaemia[12] and other extrahepatic manifestations such as chronic kidney disease[13] and gastrointestinal neoplasms.[14]

According to the National Institutes of Health U.S. National Library of Medicine, there are currently more than 300 ongoing randomized clinical trials (RCTs) enrolling patients with NAFLD. Such RCTs are challenging because NAFLD "diagnosis" often requires invasive methods and/or imaging approaches, which are clinically burdensome and cost-prohibitive, especially since NAFLD has reached epidemic proportions in developing countries that may not have the clinical, financial and infrastructural resources to identify and adequately treat patients with NAFLD. For example, liver biopsy is not only invasive and expensive but is also prone to sampling error.[15] Affordable and easily obtainable tests are required to identify NAFLD patients who may benefit from therapies under investigation. Blood biomarkers of NAFLD that are not modulated by secondary non-causal pathways, are promising candidates for the identification of at-risk individuals and to develop tailored therapy for NAFLD.

Mendelian randomization (MR) is a modern epidemiology investigation technique that is increasingly used to explore whether risk factors associated with disease traits reflect true causal associations or not.[16] Akin to a RCT, MR takes advantage of the random allocation of genetic variation at conception to determine the consequences of human traits that are at least in part under genetic control. MR has also been used to determine whether a genetic susceptibility to certain chronic diseases influences other biological traits such as the blood proteome or the blood metabolome, thereby identifying early biomarkers of disease-related traits.[17,18]

High-quality MR studies rely on the availability of the standardized effect sizes of the top genetic variants associated with a trait of interest when this trait is used as an exposure and on the availability of genome-wide association studies (GWAS) summary statistics with this trait is used as an outcome. Although GWASes have identified genetic variants associated with liver fat accumulation [19,20], liver enzymes [21] and different forms of liver diseases [22,23], less than a handful of small GWAS sought to identify genetic variants associated with a clinical diagnosis of NAFLD. The GWAS of the Electronic Medical Records and Genomics (eMERGE) network included 1106 NAFLD cases and 8571

controls identified only one NAFLD susceptibility locus (*PNPLA3*). The NAFLD GWAS of the UK Biobank included 1664 NAFLD cases and 400,055 controls identified only two regions robustly associated with NAFLD (*PNPLA3* and *PBX4/TM6SF2*). The UK Biobank analysis did not exclude participants with secondary causes of NAFLD (such as hepatitis or alcoholism) and used a rather vague definition of NAFLD (phecode 571.5: other forms of nonalcoholic liver disease). Genetic variation at these two loci also are also associated with NAFLD in the data freeze #4 of the FinnGen cohorts (651 NAFLD cases and 176,248 controls).

Here, we performed a meta-analysis of electronic health record (EHR)-based GWAS to identify genetic variants robustly associated with NAFLD. This analysis included GWAS summary statistics from the eMERGE and FinnGen cohorts, an updated NAFLD GWAS in the UK Biobank (2558 cases and 395,241 controls) and a new GWAS performed in the Estonian Biobank (4119 cases and 190,120 controls) for a total of 8434 NAFLD cases and 770,180 controls. We then used a MR study design to identify novel blood proteins and metabolites that may causally be influenced by the presence of NAFLD and a combination of genetic correlation analysis and phenome-wide MR study to identify other traits and human diseases that may be influenced by NAFLD or variants influencing NAFLD susceptibility.

Results

Identification of genetic variants associated with non-alcoholic fatty liver disease

The study design is presented in Figure 1. In order to identify independent genetic variants robustly associated with NAFLD and suitable for MR analyses, we first performed two new GWASes in the UK Biobank and Estonian Biobank and performed a meta-analysis of four cohorts (UK Biobank, Estonian Biobank, eMERGE and FinnGen) totalling 8434 NAFLD cases, all identified through electronic health records, and 770,180 controls. We identified four genetic loci that harboured at least one SNP that passed the genome-wide significance threshold of $p \leq 5x10^{-8}$ (*TRIB1*, *MAU2* [*TM6SF2*], *APOE* and *PNPLA3*). Figure 2A presents the Manhattan plot of the NAFLD GWAS meta-analysis identifying genetic regions with a p-value for association with NAFLD $\leq 5x10^{-8}$. The associated quantile-quantile plot is presented in Supplementary Figure 1. In order to add more SNPs to our genetic instruments and to identify potentially new relevant NAFLD genetic loci, we used a Bayesian approach (bGWAS) recently described by Mounier and Kutalik [24]. This method seeks to identify new variants associated with complex diseases using inference from risk factors of these diseases. By leveraging GWAS summary statistics from risk factors likely causally associated with NAFLD in a previous MR study[25] (T2D, body mass index [BMI] and triglyceride levels) as priors, this analysis revealed genetic variation at three additional loci (*GCKR*, *LPL*, and *FTO*) associated with NAFLD (Supplementary Table 1). Variation at these new loci act through selected NAFLD risk factors on Bayes Factors (Figure 2B), rather than through direct effects (Figure 2C) or posterior effects (Figure 2D). The association of lead SNPs at these loci with NAFLD as well as those from the conventional GWAS are presented in Supplementary Table 2 in each cohort separately and in the GWAS meta-analysis. Because some of these SNPs showed evidence of heterogeneity, p-values are presented from fixed effects and random effects meta-analysis. This table also presents the F statistic for instrument strength for each SNPs. These range from 16.3 to 212.0. Altogether, through a combination of conventional GWAS and risk factor informed GWAS, our analysis identified genetic variation at seven loci that may influence susceptibility to NAFLD.

Evaluation of the functionality of variants associated with NAFLD

Some of the top variants linked with NAFLD in this analysis may have functional consequences. For instance, the rs1260326 at *GCKR* is a missense variant (p.P446L). The rs1260326 at *APOE* is also a missense variant (p.R130C). The lead variant at MAU2/TM6SF2 rs73001065 is in linkage disequilibrium (r^2=0.90) with the missense variant p.E167K at *TM6SF2* and the lead variant at *PNPLA3* is in high linkage disequilibrium (r^2=0.98) with the missense

variant p.I148M at PNPLA3. Table 1 presents the details of these results as well as the effect of other previously associated variants with NAFLD (p.A165T at *MTARC1*, a splice variant *HSD17B13* and another variant at *MBOAT7*). This analysis confirmed previous NAFLD functional variants at *MTARC1* and *MBOAT7*, but not at *HSD17B13*. Genetic variation at the *PNPLA3*, *TM6SF2*, *APOE* and *GCKR* have been linked with NAFLD-related traits in previous studies[26,27]. Recent studies identified *APOE*, *TR1B1* and *FTO* as potential new loci for liver enzymes [28,29]. To our knowledge, our study is the first to link variation at these loci with a clinical diagnosis of NAFLD and to identify *LPL* as a potential new susceptibility locus for NAFLD. Interestingly, the minor allele (C) at rs13702 associated here with a protection against NAFLD has been predicted to disrupt a micro RNA recognition element seed site for human micro RNA miR-410, resulting in higher *LPL* expression[30]. We therefore sought to determine whether genetically-predicted *LPL* expression was associated with NAFLD. We performed a transcriptome-wide association study for NAFLD to map genetically-regulated genes from the Genotype Tissue Expression (GTEx, version 8) consortium[31] with NAFLD using S-PrediXcan. This analysis did not reveal new NAFLD genes outside those who had a genome-wide signal such as *PNPLA3* and *TM6SF2* (data not shown). Genetically-predicted *LPL* expression could be estimated in 11 tissues. The association between genetically-predicted *LPL* expression in these 11 tissues and NAFLD is presented in (Supplementary Table 3). This analysis suggests a negative association between genetically-predicted *LPL* expression in subcutaneous adipose tissue and NAFLD (p=3.1e-04). The LocusCompare plot (Figure 3) further suggests shared genetic etiology at this locus with the rs13702 variant being significantly associated with both subcutaneous adipose tissue expression of *LPL* and NAFLD[32]. In summary, most of the seven SNPs identified in this analysis or SNPs in close proximity may be considered as functional.

Association of variants associated with NAFLD with NAFLD related phenotypes

We investigated the effect of these variants in another cohort and with NAFLD related traits such as liver fat accumulation and liver enzymes in the UK Biobank. In the Mass General Brigham Biobank, 4312 patients with non-alcoholic steatohepatitis (NASH) or NAFLD (diagnosed by computed tomography and/or magnetic resonance imaging) were compared to 26,404 controls. The direction of the effects of the seven SNPs were concordant with those observed in the GWAS meta-analysis. All SNPs were nominally associated with NAFLD in the Mass General Brigham Biobank, with the exception of the variants at the *FTO* and at the *LPL* loci (Supplementary Table 4). Liver fat accumulation in the UK Biobank was quantified via machine learning of abdominal MRI images as previously described.[33] We analyzed liver fat accumulation as a continuous trait and as a categorical trait (liver fat ≥5.5%) in 32,976 study participants. The direction of the effects of the seven SNPs on liver fat accumulation were concordant with those observed in the GWAS meta-analysis and all SNPs were nominally associated with liver fat accumulation (whether considered as a continuous or as a categorical trait), with the exception of the variant at the *LPL* locus (Supplementary Table 4). Finally, the association between the seven variants associated NAFLD with the liver enzymes ALT (alanine aminotransferase), AST (aspartate aminotransferase), GGT (gamma-glutamyl transferase) and ALP (alkaline phosphatase) was investigated in 361,194 participants of the UK Biobank. Results presented in Supplementary Table 4, suggest that all variants were positively associated with liver enzymes, except that the variant at *GCKR* was not associated with ALT levels, the variant at *APOE* was not associated with AST levels and the variant at *PNPLA3* was not associated with GGT levels. Variants at the *GCKR*, *LPL*, *TRIB1* and *APOE* were positively associated with ALP levels, the variant at *FTO* was not associated with ALP levels and the variants at *MAU2*/ *TM6SF2* and *PNPLA3* were negatively associated with ALP levels. Overall, results of this analysis suggest that the seven variants associated with NAFLD are associated with NAFLD-related traits such as liver fat accumulation and/or liver enzymes.

Impact of non-alcoholic fatty liver disease variants on the blood metabolome and proteome

In order to explore the metabolic effect of variants influencing NAFLD, we investigated the impact of these variants on the blood metabolome using GWAS summary statistics on 123 blood lipids, lipoproteins and metabolites measured in 24,925 individuals from 10 European cohorts, as described by Kettunen et al.[34] and on the blood proteome using GWAS summary statistics from five large-scale protein datasets[35-38] included in the Phenoscanner v2[39]. Figure 4 presents the impact of the NAFLD variants on the blood metabolome. This analysis showed that the lead NAFLD variant at the *GCKR* locus was associated with several blood metabolites such as higher lipoprotein/lipid levels, branched chain amino acids and other amino acids as well as glycolysis and gluconeogenesis metabolites. The lead NAFLD variants at the *LPL, TRIB1 and MAU2/TM6SF2* loci were also associated with higher lipoprotein/lipid levels while the lead NAFLD variant at the *APOE* locus was associated with lower lipoprotein/lipid levels. The impact of NAFLD associated variants at the *FTO* and *PNPLA3* loci did not appear to have a major influence on the blood metabolome. Supplementary Table 5 presents the impact of the NAFLD variants on blood proteins (with p-value for association <$1e^{-05}$). Variants in *GCKR*, *TRIB1*, *MAU2/TM6SF2* and *APOE* were associated with plasma levels of nine, three, two and 61 plasma proteins, respectively, while variants in *LPL*, *FTO* and *PNPLA3* did not show associations with any of the circulating proteins at that threshold. Results of this analysis suggests that the variants influencing NAFLD might have divergent effect on the blood metabolome and a trivial effect on the blood proteome, with the exception of *APOE*.

Mendelian randomization analysis on the impact of non-alcoholic fatty liver disease on the blood metabolome

We used the top SNPs from each of the four loci that showed association with NAFLD in the conventional GWAS and the top SNPs from each of the three loci showed association with NAFLD using the risk factor informed GWAS to create a multilocus genetic instrument for genetically-predicted NAFLD. We performed a two-sample MR analysis to determine the impact of NAFLD (rather than each variant investigated individually) on the blood metabolome. For this purpose, we used the GWAS summary statistics of 123 blood lipids as described above. Using IVW-MR, we found that genetically-predicted NAFLD was robustly associated with higher levels of tyrosine after correction for false-discovery rate (pFDR<0.10) with the Benjamini-Hochberg method (Figure 4 and Supplementary Table 6). We also found an association between NAFLD and the tyrosine precursor phenylalanine, although this association did not pass the FDR-corrected statistical significance threshold. The association between NAFLD and tyrosine and phenylalanine levels was consistent across MR methods and robust to outliers and pleiotropy (Table 2 and Supplementary Figure 2). Because there was sample overlap between the exposure (genetically predicted NAFLD) and outcomes (blood metabolites), with the Estonian Biobank contributing to both datasets, we redid the NAFLD GWAS meta-analysis excluding Estonian Biobank participants. Genetically predicted NAFLD was still association with tyrosine (beta [SE] = 0.085 [0.026], p=9.37E-04]) and phenylalanine levels (beta [SE] = 0.065 [0.026], p=0.011]) using IVW-MR.

We next investigated whether the presence of NAFLD was associated with higher plasma levels of tyrosine and phenylalanine in the Estonian Biobank. Tyrosine and phenylalanine levels were measured in 10,809 individuals including 359 patients with NAFLD (obtained from EHR). Supplementary Figure 3 presents the distribution of tyrosine and phenylalanine levels in cases and controls. Supplementary Table 7 presents the association between tyrosine and phenylalanine levels per one-standard deviation increment before and after multivariable adjustment. After adjusting for age, sex, smoking, education, and BMI, tyrosine levels, but not phenylalanine levels were positively associated with the presence of NAFLD in the Estonian Biobank (odds ratio per 1-SD increment = 1.23 (95% confidence interval = 1.12-1.36, p = 2.19e-05), further suggesting that NAFLD might influence tyrosine metabolism.

Impact of non-alcoholic fatty liver disease on the blood proteome

We used a similar approach as described above to determine the impact of genetic exposure to NAFLD on the blood proteome using GWAS summary statistics on >3000 circulating blood proteins from the INTERVAL study.[37] After FDR correction, we found that NAFLD was associated with higher levels of three circulating proteins: alpha L-iduronidase (encoded by the *IDUA* gene), glutathione-S-transferase A1 (encoded by the *GSTA1* gene), alcohol dehydrogenase 4 (encoded by the *ADH4* gene) (Figure 5A and Supplementary Table 8). The association between NAFLD and plasma levels of these circulating proteins was consistent across MR methods and robust to outliers and pleiotropy (Table 2 and Supplementary Figure 4). In order to gain insight into potential tissue specificity of the genes encoding these proteins, we obtained the tissue-specific gene expression metric (Tau) from the Genotype-Tissue Expression dataset resource, as described by Kryuchkova-Mostacci and Robinson-Rechavi.[40] Genes with evidence of tissue-specific expression have a Tau value closer to 1 while ubiquitous genes have a Tau value closer to 0. This analysis revealed that two of the genes (*ADH4* and *GSTA1*) encoding circulating proteins that may causally be influenced by NAFLD had tissue-specific expression (Tau ≥0.80) (Figure 5B). Altogether, this analysis revealed additional proteins that are influenced by the presence of NAFLD and that may represent new biomarkers of NAFLD.

One of the key assumptions of MR is that genetic variants used as a proxy of the exposure influence the outcome via their effect on the exposure and not via other related traits (horizontal pleiotropy). To identify NAFLD genetic instruments, we leveraged prior effect sizes of NAFLD-related traits, which might increase the chance of finding associations that may be influenced by NAFLD related traits and not by NAFLD per se. We therefore investigated the impact of genetically-predicted NAFLD on blood levels of tyrosine, alpha-L-iduronidase, glutathione S-transferase A1 and alcohol dehydrogenase 4 using all independent (one SNP per region) NAFLD SNPs (with p-value for association <5e-06) using multiple methods (as in Table 2). Results presented in Supplementary Table 9 show strong associations with the other genetic instrument, suggesting that the effect of genetically-predicted NAFLD on these biomarkers may not be driven by horizontal pleiotropy.

Association of non-alcoholic fatty liver disease with human metabolic and phenotypic traits

We performed cross-trait genetic correlation analyses between NAFLD and 240 human traits centralized in the LD Hub database. LD Hub includes GWAS publicly available summary statistics on hundreds of human traits and enables the assessment of LD score regression among those traits. Results presented in Figure 6 show high levels of genetic correlation between NAFLD and cardiometabolic traits such as obesity, insulin resistance and triglycerides and negative genetic correlation with parental lifespan, socio-economic factors and the ketone body acetoacetate.

Impact of non-alcoholic fatty liver disease variants on the human disease-related phenome

In order to determine the effect of the variants associated with NAFLD on the human disease related phenome, we performed phenome-wide MR analyses in the UK Biobank and in the FinnGen cohorts (853 and 1404 disease-specific binary traits in the UK Biobank FinnGen cohorts, respectively). Supplementary Table 10 and Supplementary Table 11 present the effect of the seven NAFLD-associated variants on disease-related traits scaled to their effect on NAFLD after correction for false-discovery rate (pFDR<0.10) with the Benjamini-Hochberg method. This analysis revealed that in accordance to its effect on lipid levels, the NAFLD-associated missense variant at the *GCKR* locus is associated with hypercholesterolemia. This variant is also positively associated with gout and other crystal arthropathies but negatively associated with T2D and cholelithiasis. The NAFLD-associated variants at the *LPL* and *TRIB1* loci were positively associated with cardiometabolic diseases such as hyperlipidemia and CVD. The NAFLD-associated variant the *TRIB1* locus was however negatively associated with cholelithiasis. The NAFLD-associated variant the *FTO* locus was positively associated with several cardiometabolic disorders such as obesity, T2D, sleep apnea, osteoarthrosis and hypertension. The NAFLD-associated variant the *MAU2/TM6SF2* locus was positively

associated with various forms of liver disease and T2D (UK Biobank only) and was negatively associated with hyperlipidemia in both cohorts and with CVD (in FinnGen only). The NAFLD-associated missense variant at the *APOE* locus is negatively associated with the presence of several neurological disorders, hypercholesterolemia and CVD and positively associated with T2D and asthma. Finally, the NAFLD-associated variant at the *PNPLA3* locus is associated with several forms of liver disease such as cirrhosis and alcoholic liver damage and diabetes-related traits (T2D in UK Biobank and diabetic maculopathy in FinnGen) and negatively associated with the presence of gout and other crystal arthropathies. Altogether, this analysis identified several phenotypic consequences of NAFLD associated variants and revealed significant heterogeneity of NAFLD-raising SNPs on the human disease-related phenome.

Discussion

We performed two new genome-wide association studies for NAFLD in the UK Biobank and in the Estonian Biobank and combined these results to those of two publicly available NAFLD GWAS (from the eMERGE network and FinnGen). This GWAS meta-analysis included 8434 NAFLD cases available via EHRs and 770,180 controls, making it the largest genome-wide analysis for a clinical diagnosis of NAFLD. In combination with a risk factor-informed Bayesian GWAS, this analysis identified three known susceptibility loci for NAFLD (*GCKR*, *TM6SF2* and *PNPLA3*) and four new candidate genetic regions for a clinical diagnosis NAFLD based on EHRs (*TRIB1*, *LPL*, *FTO*, *APOE*). We investigated the impact of the top variant at each region with the blood metabolome, the blood proteome and the human disease-related phenome. We established a MR framework aimed at identifying novel early biomarkers of NAFLD that may be causally impacted by the presence of NAFLD. This analysis revealed an intriguing effect of NAFLD on tyrosine metabolism and on the presence of three circulating blood proteins, which may represent clinical biomarkers of NAFLD.

Our conventional GWAS analysis reports that variation at the *TRIB1*, *MAU2/TM6SF2*, *APOE* and *PNPLA3* loci may be linked with NAFLD. While genetic variants at these loci have been associated with some liver phenotypes.[22,26,27,41], this GWAS meta-analysis revealed important information on the genetic architecture of NAFLD. Using bGWAS, our study identified known, and potentially new loci for NAFLD (*GCKR*, *LPL*, and *FTO)* that may be associated with NAFLD through their effects on NAFLD risk factors (BMI, T2D and triglycerides). Although the biological relevance of variation at the *FTO* locus is still a matter of debate, *FTO* is a well-characterized genetic locus for obesity.[42] Other studies reported an association of variants at the *GCKR* loci and liver fat accumulation[19] and liver enzymes[21]. Lipoprotein lipase (LPL) on the other hand is a key enzyme that regulates the catabolism of triglycerides-rich lipoproteins such as chylomicrons and very-low-density lipoproteins in adipose tissue, skeletal muscle and heart. Gain-of-function mutations in LPL were associated with lower triglyceride levels and lower risk of coronary artery disease.[43] In this study, we found a potentially causal association between genetically-predicted *LPL* expression in subcutaneous adipose tissue and NAFLD. These results are in line with the recent study of Maltais et al.[44] who have reported that 4 out of 10 patients with familial chylomicronemia syndrome and almost 3 out of 4 patients with multifactorial chylomicronemia syndrome (two disorders of impaired *LPL* function) met the criteria of NAFLD, independently of their BMI. It should be noted that although additional associations the variant at the *LPL* locus linked with higher NAFLD was associated with liver enzymes in the UK Biobank, it was not associated with liver fat accumulation in the UK Biobank or with NAFLD in the Mass General Brigham Biobank. Additionally, although these results did not reach the level of genome-wide significance, we found nominally significant associations at the *MTARC1* and *MBOAT7* loci, thereby confirming the role of these genes in the etiology of NAFLD.

Several observational studies have suggested that liver fat accumulation or NAFLD negatively impacts triglyceride-rich lipoprotein metabolism, glucose-insulin homeostasis as well as branched-chain amino acid levels.[45-49] Sliz et

al.[50] also documented the individual impact of 4 variants (at the *PNPLA3*, *TM6SF2*, *GCKR* and *LYPLAL1* loci) on the blood metabolome and found inconsistent associations. Here, we used a similar approach to investigate the effect of the top NAFLD variant (included 3 already investigated by Sliz et al.). This analysis confirmed that the lead NAFLD variant at the *GCKR* locus is linked with several blood metabolites levels such as higher lipoprotein/lipid, branched chain amino acids and other amino acids as well as glycolysis and gluconeogenesis metabolites. Variation at the *LPL, TRIB1 and MAU2/TM6SF2* loci were also associated with higher lipoprotein/lipid levels. On the other hand, variation at the *APOE* locus was associated with lower lipoprotein/lipid levels. We investigated whether the presence of NAFLD impacted lipoprotein levels and metabolites of these pathways to identify early biomarkers of NAFLD using MR. This analysis did not find evidence of a causal association of NAFLD with triglyceride-rich lipoprotein metabolism, which is expected since some variants were associated with higher lipid levels while other variants were associated with lower lipid levels. NAFLD was not however associated with glucose-insulin homeostasis markers or branched-chain amino acids. We did however find a significant impact of NAFLD on tyrosine and, to a lesser extent, its metabolic precursor phenylalanine. Although the impact of NAFLD on tyrosine metabolism has been reported decades ago[51], our analysis adds to this body of evidence by suggesting that the impact of NAFLD on tyrosine metabolism might be a direct consequence NAFLD, and that this association might not be driven by secondary causes of NAFLD. We also reported that patients with higher blood levels of tyrosine had a higher prevalence of NAFLD in the Estonian Biobank.

Our MR analysis identified three proteins that may be causally impacted by NAFLD. Glutathione-S-transferase A1 (encoded by the *GSTA1* gene) has been shown to be a sensitive biomarker of hepatocellular damage.[52] Alcohol dehydrogenase 4 (encoded by the *ADH4* gene) is also a liver expressed enzyme that mediates oxidative pathways involved in alcohol metabolism.[53] Finally, Alpha L-iduronidase (encoded by the IDUA gene), is a lysosomal enzyme involved in glycosaminoglycan degradation. Additional studies are required to determine if these blood-based biomarkers could predict NAFLD onset or identify patients with more severe forms of liver diseases.

Previous studies have shown that NAFLD could be associated with, or predict the risk of chronic diseases such as CVD or T2D. Our genetic correlation analyses revealed associations with these diseases as well as risk factors for these diseases such as obesity and insulin resistance. We also report interesting negative correlations between NAFLD and the ketone body acetoacetate (as previously suggested in an observational study)[54] as well as parental lifespan, suggesting that NAFLD may be a critical component of long-term disease risk potentially influencing human lifespan. Similar to the impact on the blood metabolome, our study also documented considerable heterogeneity with regards to disease-traits associated with NAFLD variants. It appears that only the *LPL* variant and to a lesser extent *TR1B1* seemed to be associated with diseases in the same direction as the NAFLD association, thereby suggesting that targeting the LPL pathway may prevent NAFLD as well as other diseases such as hyperlipidemia and CVD without increasing the risk of other human diseases. Drugs targeting the LPL pathway under investigation for NAFLD include the angiopoietin-like protein-3 (ANGPTL3) inhibitors[55], glucagon-like peptide-1 (GLP-1) receptor agonists[56] and dual glucose-dependent insulinotropic peptide (GIP)/GLP-1 receptor agonists[57].

Our study has limitations. For instance, although we have excluded secondary causes of NAFLD whenever possible, an EHR-based diagnosis of complex diseases such as NAFLD might be prone to misclassification of cases and controls. The prevalence of NAFLD was also not available in some of the cohorts used to document the impact of NAFLD on the blood metabolome (24,925 individuals from 10 European cohorts) and the blood proteome (INTERVAL). We also did not have a validation cohort to replicate the effect of NAFLD on the blood proteome that we have identified nor could we determine if these biomarkers were only elevated in specific NAFLD stages or subtypes. Studies documenting the impact of NAFLD resolution on these biomarkers could also consolidate the causal effect

of NAFLD on the blood metabolome and proteome. Finally, there was also sample overlap as subjects in the UK Biobank and of the FinnGen cohorts were used to create our study exposure and were used to assess the phenotypic consequences of variants linked with NAFLD.

In conclusion, we conducted a large NAFLD GWAS based on EHRs from four cohorts to identify genetic variants of NAFLD susceptibility. We identified known NAFLD variants and show that variants associated with liver fat accumulation and liver enzymes may also be associated with the presence of NAFLD. Our analysis revealed a potentially causal effect of lower adipose-tissue expression of *LPL* and NAFLD that will need confirmation by other larger studies. We also identified plasma metabolites and proteins that may be causally influenced by the presence of NAFLD, including a potential effect of NAFLD on tyrosine metabolism. These finding shed light on the metabolic consequences of NAFLD but also identifies potential early biomarkers of NAFLD that could be used to identify patients who may benefit from therapies targeting NAFLD and/or for risk stratification in this population. Additional studies will be required to determine whether our findings could be helpful optimize NAFLD risk prediction as well as patient recruitment for trials aiming at preventing and/or treating NAFLD.

Methods

Genome-wide association study summary statistics NAFLD

To obtain a comprehensive set of NAFLD GWAS summary statistics, we performed a GWAS meta-analysis of four cohorts: The Electronic Medical Records and Genomics (eMERGE)[58] network, the UK Biobank, the Estonian Biobank and FinnGen. The NAFLD GWAS in the eMERGE network has previously been published.[59] The study sample included 1106 NAFLD cases and 8571 controls participants of European ancestry. Of them, 396 NAFLD cases and 846 controls participants (47% males) were derived from a pediatric population and 710 NAFLD cases and 7725 controls participants (42% males) were derived from an adult population. NAFLD was defined by the used of EHR codes (ICD9: 571.5, ICD9: 571.8, ICD9: 571.9, ICD10: K75.81, ICD10: K76.0 and ICD10: K76.9. Exclusion criteria included, but were not limited to alcohol dependence, alcoholic liver disease, alpha-1 antitrypsin deficiency, Alagille syndrome, liver transplant, cystic fibrosis, hepatitis, abetalipoproteinemia, LCAT deficiency, lipodystrophy, disorders of copper metabolism Reye's syndrome, inborn errors of metabolism, HELLP syndrome, starvation and acute fatty liver (as suggested by the American Association for the Study of Liver Disease [AASLD]) Logistic regression analysis was performed on over 7 million SNPs with MAF >1% adjusted for age, sex, body mass index, genotyping site and the first three ancestry based principal components. We performed a new GWAS for NAFLD in the UK Biobank (data application number 25205). NAFLD diagnosis was established from hospital records (ICD10: K74.0 and K74.2 (hepatic fibrosis), K75.8 (NASH), K76.0 (NAFLD) and ICD10: K76.9 (other specified diseases of the liver). Exclusion criteria were the same as those used in the eMERGE study. In the UK Biobank genome-wide genotyping was available for over 28 million genetic markers directly genotyped or imputed by the Haplotype Reference Consortium (HRC) panel. We used the SAIGE (Scalable and Accurate Implementation of Generalized Mixed Models) method to performed the GWAS. This method is based on generalized mixed models and was developed to control for case-control imbalance, sample relatedness and population structure. In this analysis, sex, age and the 10 main ancestry-based principal components were used as covariates. This UK Biobank analysis included 2558 NAFLD cases and 395,241 controls. We also performed a GWAS for NAFLD using SAIGE in the Estonian Biobank. This study and the use of data from 4119 cases and 190,120 controls was approved by the Research Ethics Committee of the University of Tartu (Approval number 288/M-18). We used the same case definition and inclusion/exclusion criteria as in the UK Biobank. Age, sex and the 10-main ancestry-based PCs were used as covariates. Finally, SAIGE was also used to obtain GWAS summary statistics of the FinnGen cohort. In this study, GWAS was performed using over 16 million

genetic markers genotyped with the Illumina or Affymetrix arrays or imputed using the population specific SISu v3 reference panel. Variables included in the models were sex, age, the 10-main ancestry-based principal components and genotyping batch. In the FinnGen data freeze 4 (November 30, 2020), 651 patients had a NAFLD diagnosis (EHR code K76.0). They were compared to 176,248 controls. We performed a fixed-effect GWAS meta-analysis of the eMERGE, UK Biobank, FinnGen and Estonian Biobank cohorts using the METAL package. [60] When variants showed evidence of heterogeneity, we performed a random effect meta-analysis. A total of 6,797,908 SNPs with a minor allele frequency equal or above 0.01 were investigated.

Risk-factor informed Bayesian genome-wide association study

We used bGWAS to identify more SNPs associated with NAFLD.[24] The aim of bGWAS is to identify new variants associated with complex diseases using inference from risk factors of focal traits. We used GWAS summary statistics from three risk factors causally associated with NAFLD in a previous MR study[25] (T2D, BMI and triglyceride levels) as priors and worked with default parameters of the package. GWAS summary statistics for these risk factors are included in the bGWAS package. These were obtained from the Global Lipids Genetic Consortium, Genetics of Anthropometric Traits (GIANT) and the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortia. Briefly, bGWAS derives informative prior effects from these risk factors and their causal effect on NAFLD using multivariable MR. Prior estimates (mu) are calculated for each SNP by multiplying the SNP-risk factor effect by the SNP-NAFLD causal effect estimates. By combining observed effects from the NAFLD GWAS meta-analysis and prior effects, Bayes factors, posterior effects and direct effects and their corresponding p-values are generated.

Transcriptome-wide association study of NAFLD

Tissues from the GTEx consortium (version 8) with less than 70 samples were not used to provide sufficient statistical power for eQTL discovery, resulting in a set of 48 tissues. Only non-sex-specific tissues (N=43) were analyzed. Alignment to the human reference genome hg28/GRCh38 was performed using STAR v2.6.1d, based on the GENCODE v30 annotation. RNA-seq expression outliers were excluded using a multidimensional extension of the statistic described by Wright et al.[61] Samples with less than 10 million mapped reads were removed. For samples with replicates, replicate with the greatest number of reads were selected. Expression values were normalized between samples using TMM as implemented in edgeR[62]. For each gene, expression values were normalized across samples using an inverse normal transformation. eQTL prediction models were performed using elastic net, a regularized regression method, as implemented in S-PrediXcan[63,64]. We used SNPs with a minor allele frequency greater than 1% from European ancestry participants. *Locuscompare* function from the *LocuscompareR* R package[65] was used to depict the colocalization event at the *LPL* locus. *Locuscompare* enables visualization of the strengths of eQTLs and outcomes associations by plotting p-values for each within a given genomic location, thereby contributing to distinguish candidates from false-positive genes.

Replication of variants associated with NAFLD in the Mass General Brigham Biobank

The Mass General Brigham Biobank is a hospital-based biorepository with genetic data linked to clinical records as previously described.[66] Patients were defined as having NAFLD or NASH according to diagnosis codes in the electronic health care record and were compared to controls without such diagnoses. In this cohort, genotyping was performed using the Illumina MEGA array. Association of each the seven variants associated with NAFLD was assessed using logistic regression of disease status with age, sex and five principal components of ancestry as covariates.

Impact of NAFLD variants on liver fat accumulation in the UK Biobank

As part of the study protocol of the UK Biobank, a subset of individuals underwent detailed imaging between years 2014 and 2019 including abdominal MRI.[67] Liver fat in this cohort was quantified via machine learning of abdominal MRI images as previously described.[33] We excluded samples that had no imputed genetic data, a genotyping call rate < .98, a mismatch between submitted and inferred sex, sex chromosome aneuploidy, exclusion from kinship inference, excessive third-degree relatives, or that were outliers in heterozygosity or genotype missingness rates, all of which were previously defined centrally by the UK Biobank[68] Due to the small percentage of samples of non-European ancestries, to avoid artifacts from population stratification we restricted our GWAS to samples of European ancestries, determined via self-reported ancestry of British, Irish, or other white and outlier detection using the R package *aberrant*, resulting in a total of 32,976 individuals. We did not remove related individuals from this analysis as we used a linear mixed model able to account for cryptic relatedness in common variant association studies.[69] For analysis of liver fat as a continuous trait, we applied a rank-based inverse normal transformation. We took the residuals of liver fat in a linear model that included sex, year of birth, age at time of MRI, age at time of MRI squared, genotyping array, MRI device serial number, and the first ten principal components of ancestry. We then performed the inverse normal transform on the residuals from this model, yielding a standardized output with mean 0 and standard deviation of 1. We measured the association of genetic variants with rank inverse normal transformed liver fat via a linear mixed model using BOLT-LMM (version 2.3.4) to account for ancestry, cryptic population structure, and sample relatedness. The default European linkage disequilibrium panel provided with BOLT was used. We also determined the effects of each of the seven variants on presence of hepatic steatosis (liver fat >5.5%)[70] using logistic regression in the same 32,976 individuals, adjusting for sex, year of birth, age at time of MRI, age at time of MRI squared, genotyping array, MRI device serial number, and the first ten principal components of ancestry.

Impact of NAFLD variants on liver enzymes in the UK Biobank

Age, sex and ancestry-based principal components-adjusted GWAS summary statistics on ALT, AST, GGT and ALP concentrations in 361,194 participants of the UK Biobank of European ancestry, were obtained from the Neale lab. Details on the protocols used to measure these biomarkers is available on the UK Biobank website: https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/serum_biochemistry.pdf.

Impact of NAFLD on the blood metabolome

We used GWAS summary statistics from the study of Kettunen et al.[34] In this study, 123 blood lipids and metabolites were measured in 24,925 individuals from 10 European cohorts using high-throughput nuclear magnetic resonance spectroscopy. Metabolites measured using this platform represent a broad molecular signature of systemic metabolism and include metabolites from multiple metabolic pathways (lipoprotein lipids and subclasses, fatty acids as well as amino acids, glycolysis precursors, etc.). The association of each the seven variants associated with NAFLD was assessed using logistic regression and the association of genetically-determined NAFLD and the blood metabolome was assessed using the IVW-MR with the *mr* function from *TwoSampleMR* package in R.[16] The IVW-MR is comparable to performing a meta-analysis of each Wald ratio (the effect of the genetic instrument on the outcome divided by its effect on the exposure). Additional MR analysis were performed to evaluate heterogeneity (intercept p-value from MR Egger[71]) and the presence of outliers. We used MR-PRESSO[72], an outlier-robust method, to detected the presence of outliers (variants potentially causing pleiotropy and influencing causal estimates) and causal

estimates were obtained before and after excluding outliers. We also used the simple median and weighted median consensus methods, which give more weight to more precise genetic instruments.

Impact of NAFLD on tyrosine and phenylalanine levels in the Estonian Biobank

Blood plasma levels of tyrosine and phenylalanine were measured using nuclear magnetic resonance spectroscopy in 10,809 participants of the Estonian Biobank. Odds-ratios and corresponding p-values were estimated using logistic regression model implemented in R version 3.6.1. Metabolite values were scaled and centered prior to analysis. Two models were run: raw model with adjusting for age and sex; and adjusted model, which was additionally adjusted for smoking status, education and body-mass index.

Impact of NAFLD on the blood proteome

A comparable analytical framework as the one used above for the discovery of NAFLD-associated metabolites was used to identify NAFLD-associated proteins. For that purpose, we used GWAS summary statistics from the INTERVAL cohort. In that study, the relative concentrations of 3,622 plasma proteins or protein complexes were assayed using 4,034 modified aptamers (SomaSCAN) in 3,301 participants from the INTERVAL study, as described by Sun et al.[37]

Tissue-specificity of gene expression of proteins influences by non-alcoholic fatty liver disease

The tissue-specific gene expression metric (Tau) was obtained from all genes encoding proteins causally influenced by NAFLD. We used the formula from Yanai et al.[73] to compare the level of gene expression across selected tissues based on RNA sequencing data from European ancestry donors from GTEx. All the genes with expression <1 RPKM were set as not expressed. The RNA-sequencing data were first log-transformed. After the normalization, a mean value from all replicates for each tissue separately was calculated. A Tau value closer to 1 indicates tissue-specificity while a Tau value closer to 0 indicates ubiquitous gene expression. We considered that genes encoding proteins found to be causally impacted by NAFLD had tissue-specific expression when their Tau statistic was ≥0.80.

Phenome-wide Mendelian randomization studies in the UK Biobank and FinnGen cohorts

A recent study performed in the UK Biobank generated GWAS summary statistics on 1403 disease-specific binary traits in 408,961 white British participants.[74] A scheme was used to defined disease-specific binary traits by combining International Classification of Diseases (ICD)-9 codes into hierarchical “PheCodes”. UK Biobank participants were assigned a PheCode if they had one or more of the PheCode-specific ICD codes. A detailed description of the EHR codes included in this phecode are available on the Center for Precision Health Data Science of the University of Michigan website: http://prsweb.sph.umich.edu:8080/phecodeData/searchPhecode. In the present analysis, outcomes with a case: control ratio <1:1000 were excluded leaving 853 traits for PheWAS. In the FinnGen cohorts (data freeze 4), outcomes with <500 cases were excluded leaving 1404 traits for PheWAS. In both datasets, we determined the effect of each SNP on disease related traits, scaled on its effect on NAFLD using Wald ratios where the effect of each variant on the disease-related trait divided by its effect on NAFLD.

Data availability

GWAS summary statistics of the genome-wide meta-analysis of NAFLD will be made available on the GWAS catalog at time of publication.

The GWAS summary statistics for NAFLD of the eMERGE network are available here: https://www.ebi.ac.uk/gwas/studies/GCST008468

The GWAS summary statistics for NAFLD of FinnGen are available here: https://www.finngen.fi/en/access_results

The bGWAS R package is available at: https://github.com/n-mounier/bGWAS

The *LocusCompareR* R package is available at https://github.com/boxiangliu/locuscomparer.

GWAS summary statistics on the liver enzymes measured in participants of the UK Biobank are available here: http://www.nealelab.is/blog/2019/9/16/biomarkers-gwas-results

GWAS summary statistics for the proteins of the INTERVAL cohort are available for download at: https://www.phpc.cam.ac.uk/ceu/proteins/

GWAS summary statistics for lipoprotein metabolomics parameters, from Kettunen et al. are available for download at: http://www.computationalmedicine.fi/data#NMR_GWAS.

GTEx data is available to download at https://gtexportal.org/home/datasets. The data used for the analyses described in this manuscript were obtained from dbGaP, accession number phs000424.vN.pN.

The GWAS summary statistics for >1400 binary phenotypes in the UK Biobank by SAIGE are available to download at https://www.leelabsg.org/resources.

The GWAS summary statistics for >2000 binary phenotypes in the FinnGen cohorts by SAIGE are available to download at https://www.finngen.fi/en/access_results.

Declarations

Acknowledgements

We would like to thank all study participants as well as all investigators of the studies that were used throughout the course of this investigation (eMERGE, UK Biobank, FinnGen, GTEx, INTERVAL and the European cohorts that have contributed to the metabolomics dataset). NP holds a doctoral research award from the *Fonds de recherche du Québec: Santé* (FRQS). EG holds a master's research award from FRQS. BJA and ST hold junior scholar awards from the FRQS. PM holds a FRQS Research Chair on the Pathobiology of Calcific Aortic Valve Disease. YB holds a Canada Research Chair in Genomics of Heart and Lung Diseases. MCV is Canada Research Chair in Genomics applied to Nutrition and Metabolic Health. Part of this study was supported by the European Union through the European Regional Development fund. The work of Estonian Genome Center, Univ. of Tartu has been supported by the European Regional Development Fund and grants No. GENTRANSMED (2014-2020.4.01.15-0012), MOBERA5 (Norface Network project no 462.16.107) and 2014-2020.4.01.16-0125. This study was also funded by the European Union through Horizon 2020 research and innovation programme under grant no 810645 and through the European Regional Development Fund project no. MOBEC008 and Estonian Research Council Grant PUT1660. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

References

1 Sumida, Y. & Yoneda, M. Current and future pharmacological therapies for NAFLD/NASH. *Journal of gastroenterology***53**, 362-376 (2018).

2 Stefan, N., Häring, H.-U. & Cusi, K. Non-alcoholic fatty liver disease: causes, diagnosis, cardiometabolic consequences, and treatment strategies. *The lancet Diabetes & endocrinology***7**, 313-324 (2019).

3 Younossi, Z. M. *et al.* Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology***64**, 73-84 (2016).

4 Eguchi, Y. *et al.* Prevalence and associated metabolic factors of nonalcoholic fatty liver disease in the general population from 2009 to 2010 in Japan: a multicenter large retrospective study. *Journal of gastroenterology***47**, 586-595 (2012).

5 Pais, R. *et al.* NAFLD and liver transplantation: current burden and expected challenges. *Journal of hepatology***65**, 1245-1257 (2016).

6 Yoshitaka, H. *et al.* Nonoverweight nonalcoholic fatty liver disease and incident cardiovascular disease: a post hoc analysis of a cohort study. *Medicine***96** (2017).

7 Brouwers, M. C., Simons, N., Stehouwer, C. D. & Isaacs, A. Non-Alcoholic fatty liver disease and cardiovascular disease: assessing the evidence for causality. *Diabetologia*, 1-8 (2020).

8 Kotronen, A. & Yki-Järvinen, H. Fatty liver: a novel component of the metabolic syndrome. *Arteriosclerosis, thrombosis, and vascular biology***28**, 27-38 (2008).

9 Targher, G., Day, C. P. & Bonora, E. Risk of cardiovascular disease in patients with nonalcoholic fatty liver disease. *New England Journal of Medicine***363**, 1341-1350 (2010).

10 Anstee, Q. M., Targher, G. & Day, C. P. Progression of NAFLD to diabetes mellitus, cardiovascular disease or cirrhosis. *Nature reviews Gastroenterology & hepatology***10**, 330 (2013).

11 Lonardo, A., Ballestri, S., Marchesini, G., Angulo, P. & Loria, P. Nonalcoholic fatty liver disease: a precursor of the metabolic syndrome. *Digestive and Liver disease***47**, 181-190 (2015).

12 Neuschwander-Tetri, B. A. *et al.* Clinical, laboratory and histological associations in adults with nonalcoholic fatty liver disease. *Hepatology***52**, 913-924 (2010).

13 Kaps, L. *et al.* Non-alcoholic fatty liver disease increases the risk of incident chronic kidney disease. *United European Gastroenterology Journal*, 2050640620944098 (2020).

14 Armstrong, M. J., Adams, L. A., Canbay, A. & Syn, W. K. Extrahepatic complications of nonalcoholic fatty liver disease. *Hepatology***59**, 1174-1197 (2014).

15 Estep, J., Birerdinc, A. & Younossi, Z. Non-invasive diagnostic tests for non-alcoholic fatty liver disease. *Current molecular medicine***10**, 166-172 (2010).

16 Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. (Clinical report). *eLife***7**, doi:10.7554/eLife.34408 (2018).

17 Mohammadi-Shemirani, P. *et al.* A Mendelian randomization-based approach to identify early and sensitive diagnostic biomarkers of disease. *Clinical chemistry***65**, 427-436 (2019).

18 Ritchie, S. C. *et al.* Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases. *BioRxiv* (2019).

19 Speliotes, E. K. *et al.* Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet***7**, e1001324 (2011).

20 Parisinos, C. A. *et al.* Genome-wide and Mendelian randomisation studies of liver MRI yield insights into the pathogenesis of steatohepatitis. *Journal of hepatology***73**, 241-251 (2020).

21 Chambers, J. C. *et al.* Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nature genetics***43**, 1131-1138 (2011).

22 Emdin, C. A. *et al.* A missense variant in Mitochondrial Amidoxime Reducing Component 1 gene and protection against liver disease. *PLoS genetics***16**, e1008629 (2020).

23 Anstee, Q. M. *et al.* Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort. *Journal of hepatology***73**, 505-515 (2020).

24 Mounier, N. & Kutalik, Z. bGWAS: an R package to perform Bayesian Genome Wide Association Studies. *Bioinformatics* (2020).

25 Liu, Z. *et al.* Causal relationships between NAFLD, T2D and obesity have implications for disease subphenotyping. *Journal of Hepatology* (2020).

26 Romeo, S. *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nature genetics***40**, 1461-1465 (2008).

27 Kozlitina, J. *et al.* Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nature genetics***46**, 352-356 (2014).

28 Chen, V. L. *et al.* Genome-wide association study of serum liver enzymes implicates diverse metabolic and liver pathology. *Nature communications***12**, 1-13 (2021).

29 Jamialahmadi, O. *et al.* Exome-Wide Association Study on Alanine Aminotransferase Identifies Sequence Variants in the GPAM and APOE Associated With Fatty Liver Disease. *Gastroenterology* (2021).

30 Richardson, K. *et al.* Gain-of-function lipoprotein lipase variant rs13702 modulates lipid traits through disruption of a microRNA-410 seed site. *The American Journal of Human Genetics***92**, 5-14 (2013).

31 Consortium, G. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science***348**, 648-660 (2015).

32 Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nature genetics***51**, 768-769 (2019).

33 Haas, M. E. *et al.* Machine learning enables new insights into clinical significance of and genetic contributions to liver fat accumulation. *medRxiv* (2020).

34 Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature communications***7**, 1-9 (2016).

35 Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS genetics***13**, e1006706 (2017).

36 Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature communications***8**, 1-14 (2017).

37 Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature***558**, 73-79 (2018).

38 Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature communications***9**, 1-11 (2018).

39 Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics***35**, 4851-4853 (2019).

40 Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Briefings in bioinformatics***18**, 205-214 (2017).

41 Parisinos, C. A. *et al.* Genome-wide and Mendelian randomisation studies of liver MRI yield insights into the pathogenesis of steatohepatitis. *Journal of Hepatology* (2020).

42 Scuteri, A. *et al.* Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet***3**, e115 (2007).

43 Genetics, M. I. & Investigators, C. E. C. Coding variation in ANGPTL4, LPL, and SVEP1 and the risk of coronary disease. *The New England journal of medicine***374**, 1134 (2016).

44 Maltais, M., Brisson, D. & Gaudet, D. Non-Alcoholic Fatty Liver in Patients with Chylomicronemia. *Journal of Clinical Medicine***10**, 669 (2021).

45 Grzych, G. *et al.* Plasma BCAA changes in Patients with NAFLD are Sex Dependent. *The Journal of Clinical Endocrinology & Metabolism***105**, dgaa175 (2020).

46 Lovric, A. *et al.* Characterization of different fat depots in NAFLD using inflammation-associated proteome, lipidome and metabolome. *Scientific reports***8**, 1-14 (2018).

47 Lim, S., Taskinen, M. R. & Borén, J. Crosstalk between nonalcoholic fatty liver disease and cardiometabolic syndrome. *Obesity Reviews***20**, 599-611 (2019).

48 Jin, R. *et al.* Amino acid metabolism is altered in adolescents with nonalcoholic fatty liver disease—An untargeted, high resolution metabolomics study. *The Journal of pediatrics***172**, 14-19. e15 (2016).

49 Lake, A. D. *et al.* Branched chain amino acid metabolism profiles in progressive human nonalcoholic fatty liver disease. *Amino acids***47**, 603-615 (2015).

50 Sliz, E. *et al.* NAFLD risk alleles in PNPLA3, TM6SF2, GCKR and LYPLAL1 show divergent metabolic effects. *Human molecular genetics***27**, 2214-2223 (2018).

51 Andersson, S. M., Salaspuro, M. & Ohisalo, J. J. Metabolic basis of hypertyrosinemia in liver disease. *Gastroenterology***82**, 554-557 (1982).

52 Knapen, M. F. *et al.* Plasma glutathione S-transferase alpha 1-1: a more sensitive marker for hepatocellular damage than serum alanine aminotransferase in hypertensive disorders of pregnancy. *American journal of obstetrics and gynecology***178**, 161-165 (1998).

53 Lieber, C. S. Metabolism of alcohol. *Clinics in liver disease***9**, 1-35 (2005).

54 Männistö, V. T. *et al.* Ketone body production is differentially altered in steatosis and non-alcoholic steatohepatitis in obese humans. *Liver International***35**, 1853-1861 (2015).

55 Gaudet, D. *et al.* Vupanorsen, an N-acetyl galactosamine-conjugated antisense drug to ANGPTL3 mRNA, lowers triglycerides and atherogenic lipoproteins in patients with diabetes, hepatic steatosis, and hypertriglyceridaemia. *European heart journal***41**, 3936-3945 (2020).

56 Vergès, B. *et al.* Liraglutide Increases the Catabolism of Apolipoprotein B100–Containing Lipoproteins in Patients With Type 2 Diabetes and Reduces Proprotein Convertase Subtilisin/Kexin Type 9 Expression. *Diabetes Care* (2021).

57 Wilson, J. M. *et al.* The dual glucose-dependent insulinotropic peptide and glucagon-like peptide-1 receptor agonist, tirzepatide, improves lipoprotein biomarkers associated with insulin resistance and cardiovascular risk in patients with type 2 diabetes. *Diabetes, Obesity and Metabolism***22**, 2451-2459 (2020).

58 Jongstra-Bilen, J. *et al.* Low-grade chronic inflammation in regions of the normal mouse arterial intima predisposed to atherosclerosis. **203**, 2073-2083, doi:10.1084/jem.20060245 %J The Journal of Experimental Medicine (2006).

59 Namjou, B. *et al.* GWAS and enrichment analyses of non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE Network. *BMC medicine***17**, 135 (2019).

60 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics***26**, 2190-2191 (2010).

61 Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nature genetics***46**, 430-437 (2014).

62 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology***11**, 1-9 (2010).

63 Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics***47**, 1091 (2015).

64 Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature communications***9**, 1-20 (2018).

65 Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet***51**, 768-769, doi:10.1038/s41588-019-0404-0 (2019).

66 Karlson, E. W., Boutin, N. T., Hoffnagle, A. G. & Allen, N. L. Building the partners healthcare biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. *Journal of personalized medicine***6**, 2 (2016).

67 Littlejohns, T. J. *et al.* The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications***11**, 1-12 (2020).

68 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature***562**, 203-209 (2018).

69 Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics***47**, 284 (2015).

70 Wilman, H. R. *et al.* Characterisation of liver fat in the UK Biobank cohort. *PloS one***12**, e0172921 (2017).

71 Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in medicine***36**, 1783-1802 (2017).

72 Verbanck, M., Chen, C.-y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics***50**, 693-698 (2018).

73 Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics***21**, 650-659 (2005).

74 Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics***50**, 1335-1341 (2018).

Tables

Table 1. Association of previously identified functional variants linked with liver diseases in the present genome-wide association study.

Gene	CHR	SNP	Impact on protein	Minor allele	Major allele	Association with NAFLD		
						Beta (minor allele)	SE	P-value
MTARC1	1	rs2642438	Missense (p.A165T)	A	G	-0.0674	0.0178	1.54E-04
GCKR	2	rs1260326	Missense (p.P446L)	T	C	0.0755	0.0167	5.98E-06
*HSD17B13**	4	rs72613567	Splice variant	C	G	-0.0304	0.0186	1.02E-01
MBOAT7	19	rs641738	Linked to 3' UTR	T	C	0.0519	0.0164	1.53E-03
APOE	19	rs429358	Missense (p.R130C)	C	T	-0.1366	0.0239	1.14E-08
TM6SF2	19	rs58542926	Missense (p.E167K)	T	C	0.2676	0.0320	6.90E-17
PNPLA3	22	rs738409	Missense (p.I148M)	G	C	0.2869	0.0198	1.23E-47

*The effect of a SNP in linkage disequilibrium (r^2=0.96) with this variant (rs10433879) is presented.

Table 2. Association of non-alcoholic fatty liver with blood metabolites and proteins across multiple Mendelian randomization methods.

Metabolites/proteins	Inverse-variance weighted				Simple median			Weighted median			Mr-Egger			MR_PRESSO outlier test
	N SNPs	Beta	SE	P-value	Beta	SE	P-value	Beta	SE	P-value	Intercept	P-value intercept		P-value
Tyrosine	7	0.131	0.033	6.75E-05	0.142	0.045	0.002	0.137	0.038	2.82E-04	0.010	0.515		0.334
ADH4	7	0.298	0.073	4.88E-05	0.270	0.109	0.013	0.269	0.092	0.003	0.005	0.831		0.883
GSTA1	7	0.305	0.073	3.16E-05	0.275	0.108	0.011	0.287	0.090	0.001	0.001	0.948		0.934
IDUA	7	0.327	0.081	5.46E-05	0.307	0.128	0.017	0.362	0.085	2.18E-05	-0.035	0.147		0.367

Figures

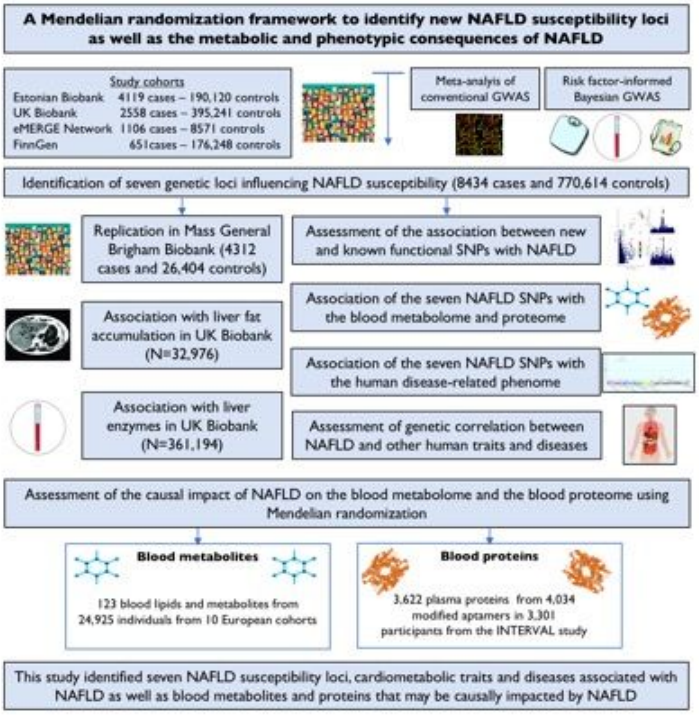

Figure 1

Figure 1

Schematic overview of the analytical framework used to identify novel genetic loci for non-alcoholic fatty liver disease, their metabolic and phenotypic effects and to identify the metabolic consequences of NAFLD.

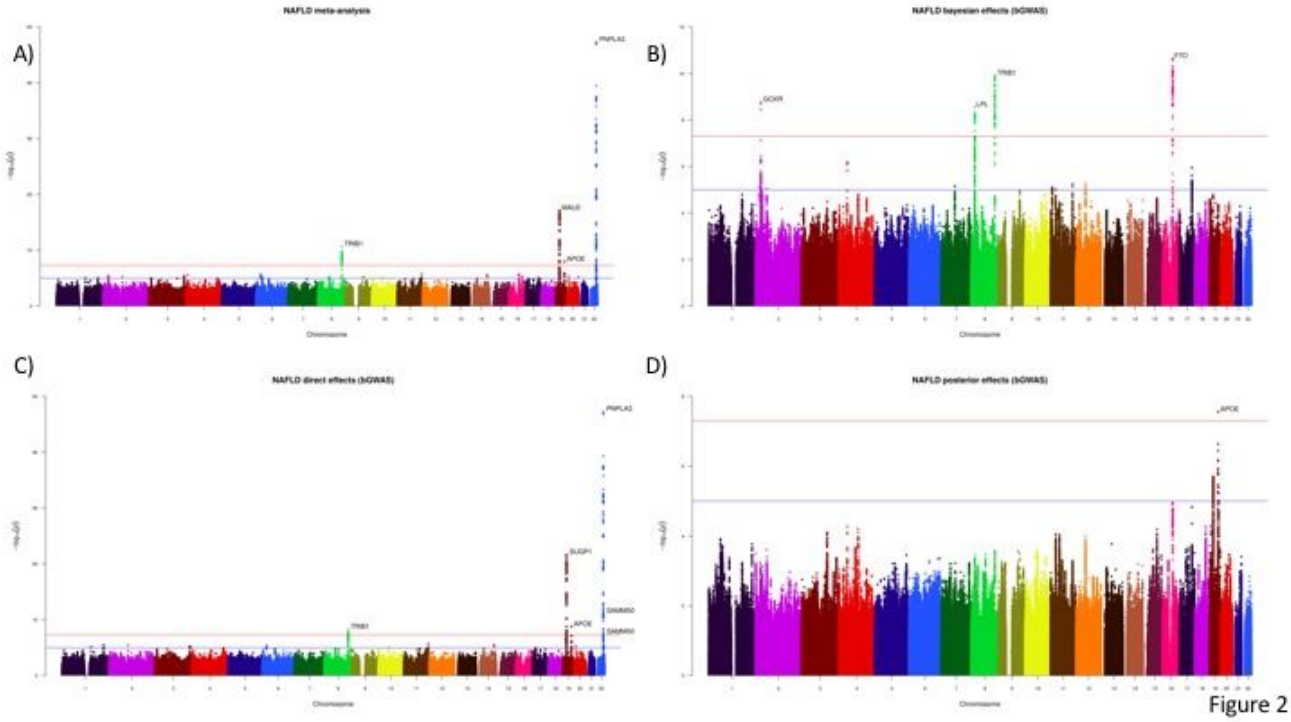

Figure 2

Figure 2

Main results of the meta-analysis of genome-wide association studies (GWAS). A) Manhattan plot depicting single-nucleotide polymorphisms (SNPs) associated with non-alcoholic fatty liver disease in the GWAS meta-analysis of the eMERGE, FinnGen, UK Biobank and Estonian Biobank cohorts. Identification of genetic variants linked with NAFLD

via a risk factor informed Bayesian GWAS based on B) Bayes Factors (BFs), C) direct effects and D) posterior effects. Genetic loci harboring SNPs associated with NAFLD (p<5.0x10-8) are shown.

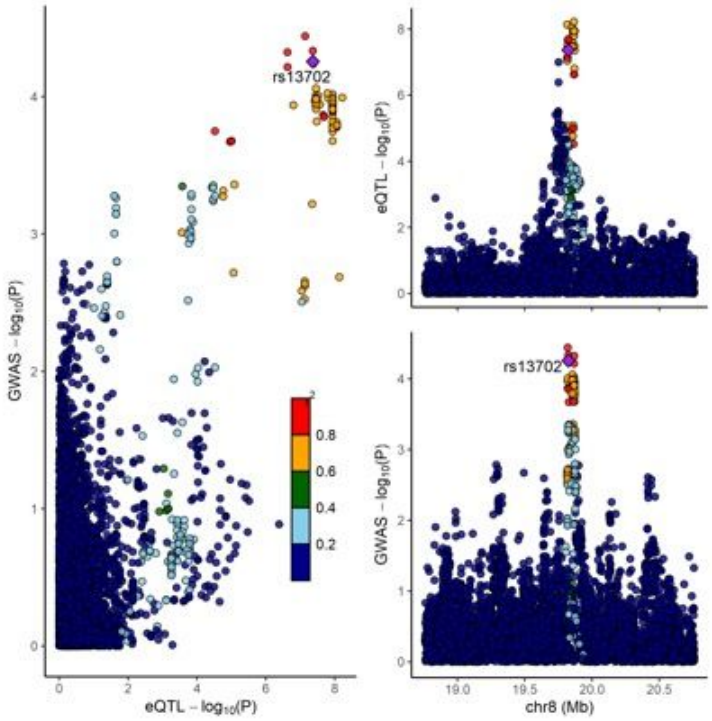

Figure 3

Figure 3

Shared genetic etiology at the LPL locus. LocusCompare plot depicting colocalization of the top SNP associated with subcutaneous adipose tissue LPL expression and non-alcoholic fatty liver disease (NAFLD). Each dot represents a single-nucleotide polymorphism (SNP) at the LPL locus. In the left panel, these SNPs are plotted to represent their effect on LPL expression (top right) against their effect on NAFLD (bottom right).

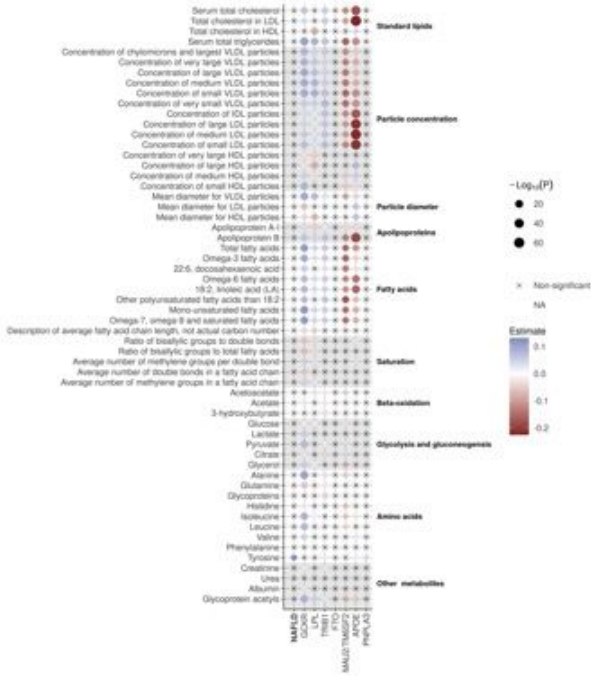

Figure 4

Figure 4

Causal impact of non-alcoholic fatty liver disease (NAFLD) and NAFLD variants on the blood metabolome. Balloon plot depicting the effect of the seven NAFLD variants on blood metabolites and of the effect of NAFLD on blood metabolites using inverse-variance weighted Mendelian randomization.

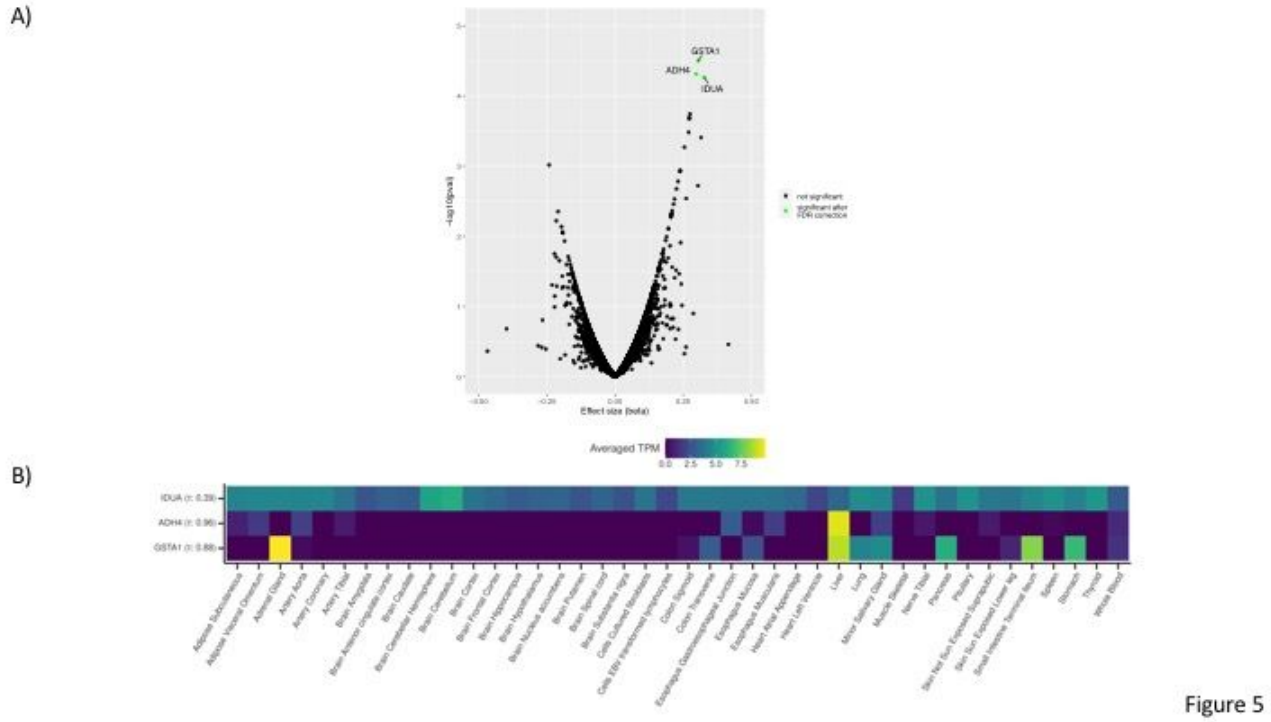

Figure 5

Figure 5

Causal impact of non-alcoholic fatty liver disease on the blood proteome. A) Volcano plot depicting and blood proteins influenced by the presence of non-alcoholic fatty liver disease using inverse-variance weighted Mendelian randomization. Green dots represent proteins significant influenced by the presence of NAFLD following correction for false discovery rate (FDR). B) Tissue-specificity of genes encoding proteins influenced by the presence of NAFLD. Heat map showing the tissue-specificity of genes encoding proteins influenced by the presence of NAFLD. Tau value is shown in parentheses after the gene name. TPM indicates transcript per per million mapped reads.

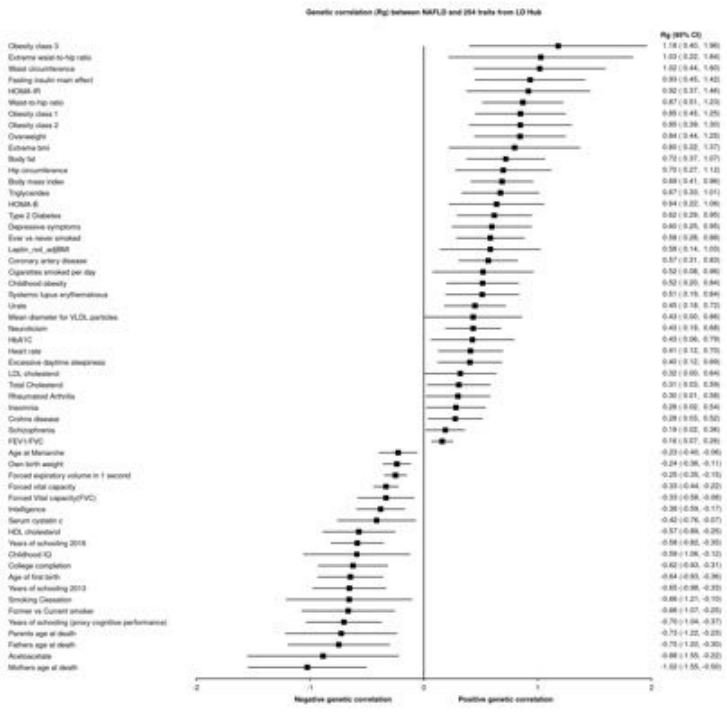

Figure 6

Figure 6

Results of the LD regression analysis between non-alcoholic fatty liver disease and other human diseases and traits. LD regression analyses were performed in LD Hub to test the genetic correlation of NAFLD with 240 human diseases and traits. Nominally significant (p<0.05) genetic correlation coefficients (Rg) and their 95% confidence interval are presented. HOMA-IR indicates homeostatic model of insulin resistance, adjBMI indicates adjusted for body mass index, VLDL indicates very-low-density lipoproteins and FEV1/FVC indicates forced expiratory volume 1/forced vital capacity.

Supplementary Files

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

- SupplTablesNAFLDreverseMRpaperBA20210304.xlsx
- FiguresNAFLDMRpaperNG20210304.pptx