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
Émilie Gobeil
universitaire de cardiologie et de pneumologie de Québec
Nele Taba
University of Tartu
Alexis St-Amand
l'Institut universitaire de cardiologie et de pneumologie de Québec
Nicolas Perrot
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
Sébastien Thériault
Institut universitaire de cardiologie et de pneumologie de Québec-Université Laval, Quebec City https://orcid.org/0000-0003-1893-8307
André Tchernof
Université Laval
Tõnu Esko
University of Tartu
Benoit Arsenault (benoit.arsenault@criucpq.ulaval.ca)
Department of Medicine, Laval University, Quebec

Article

Keywords: Electronic Health Record, Liver Disease, blood biomarkers and chronic diseases

Posted Date: November 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-97977/v1

Abstract

Non-alcoholic fatty liver disease (NAFLD) has been associated with several blood biomarkers and chronic diseases. Whether these associations underlie causal effects remains to be determined. We aimed at identifying blood metabolites, blood proteins and human diseases that are causally impacted by the presence of NAFLD using Mendelian randomization. We created a NAFLD genetic instrument from NAFLD loci (*MTARC1*, *GCKR*, *LPL*, *TRIB1*, *LMO3*, *FTO*, *TM6SF2*, *APOE* and *PNPLA3*) identified in a new electronic health record based-GWAS meta-analysis (6715 cases and 682,748 controls). We found a potentially causal effect of NAFLD on tyrosine metabolism as well as on blood levels of eight proteins that could potentially represent new early biomarkers of NAFLD. Using results from the UK Biobank, FinnGen and the COVID-19 Host Genetics Initiative, we found that NAFLD was not causally associated with diseases outside the spectrum of liver diseases, suggesting that the resolution of NAFLD might not prevent other diseases.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is one of the most prevalent chronic liver diseases.[1,2] According to recent estimates, up to 25% of the adult population worldwide may have NAFLD.[3,4]. NAFLD covers a broad disease spectrum from liver steatosis, to steatohepatitis, cirrhosis and hepatocellular carcinoma.[5,6] It has been predicted to become the most frequent indication for liver transplantation in western countries by 2030.[7] 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),[8-11] type 2 diabetes (T2D), [12,13], dyslipidaemia[14] and other extrahepatic manifestations such as chronic kidney disease[15] and gastrointestinal neoplasms.[16] Recent studies also reported an association between NAFLD and COVID-19 complications. [17-19] However, it is unknown if the association between NAFLD and these diseases reflects a true causal association and, more importantly, if drugs targeting NAFLD could simultaneously decrease the long-term risk of these life-threatening illnesses.

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.[20] Affordable and easily obtainable tests are required to identify NAFLD patients who may benefit from therapies under investigation. Causally associated biomarkers, which 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, a modern epidemiology investigation technique, is increasingly used to explore whether risk factors associated with disease traits reflect true causal associations or not.[21] MR is also a valuable tool to anticipate outcomes of RCTs of chronic diseases prevention.[22] Akin to a RCT, MR takes advantage of the random allocation of genetic variation at conception to determine the phenotypic consequences of human traits 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.[23,24] Here, we performed a meta-analysis of electronic health record (EHR)-based genome-wide association studies (GWAS) to identify genetic variants robustly associated with NAFLD. We then used a MR study design to identify novel blood proteins/metabolites causally associated with NAFLD. We next explored the impact of NAFLD on the human disease-related phenome in the UK Biobank and FinnGen cohorts as well as the COVID-19 host genetics initiative.

Results

Identification of independent single-nucleotide polymorphisms associated with non-alcoholic fatty liver disease

The study design is presented in Supplementary Figure 1. In order to identify independent genetic variants robustly associated with NAFLD and suitable for MR analyses, we first performed a meta-analysis of four cohorts totalling 6715 NAFLD cases identified through electronic health records[25] and 682,748 controls to derive GWAS summary statistics. We identified six genetic loci that harboured at least one SNP that passed the genome-wide significance threshold of $p<5x10^{-8}$ (*MTARC1*, *GCKR*, *TRIB1*, *LMO3*, *SUGP1* [*TM6SF2*] and *PNPLA3*). After the LD-clumping procedure, 8 independent SNPs ($r^2<0.1$) were identified (one at each locus with the exception of the *PNPLA3* locus, which included 3 independent SNPs). Figure 1 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 2. Regional association plots are also presented in Supplementary Figure 3.

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[26] 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[27] (T2D, body mass index [BMI] and triglyceride levels) as priors, this analysis revealed new SNPs at previously identified loci but also at three loci that were not identified in the original GWAS meta-analysis (*LPL*, *FTO* and *APOE*). We identified four genome-wide significant SNPs acting through selected NAFLD risk factors on Bayes Factors, five SNPs acting through posterior effects and nine SNPs acting through direct effects (Supplementary Figure 4 and Supplementary Table 1). We selected SNPs from this list that were not in LD with SNPs identified in the conventional GWAS, but who nevertheless showed suggestive evidence of association with NAFLD ($p<5x10^{-5}$) in the conventional GWAS meta-analysis. To create a multilocus genetic instrument for NAFLD, these four SNPs identified by bGWAS were added to the eight independent SNPs found in the conventional GWAS meta-analysis. This brought the total of SNPs included in our NAFLD genetic instrument to 12. The association of these 12 SNPs with NAFLD in the conventional GWAS are presented in Supplementary Table 2, in the GWAS

meta-analysis and in the four cohorts separately. Because some of these SNPs showed evidence of heterogeneity, p-values are presented from fixed effects and random effects meta-analysis.

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

We performed a two-sample MR analysis to determine the impact of NAFLD on the blood metabolome. For this purpose, we used GWAS summary statistics of 123 blood lipids and metabolites measured in 24,925 individuals from 10 European cohorts, as described by Kettunen et al.[28] Using IVW-MR, we did not find evidence that NAFLD was causally linked with lipoprotein lipids and subclasses, fatty acids, glycolysis precursors or most amino acids. However, NAFLD was robustly associated with increases in tyrosine levels after correction for false-discovery rate (FDR) with the Benjamini-Hochberg method (Figure 2A and Supplementary Table 3). We also found an association between NAFLD and the tyrosine precursor phenylalanine. The association between NAFLD and tyrosine and phenylalanine levels was consistent across MR methods and robust to outliers and pleiotropy (Table 1 and Supplementary Figure 5). 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] = 0.079 [0.016], p=8.78E-07]) and phenylalanine levels (beta [SE] = 0.052 [0.017], p=1.98E-03]) using IVW-MR.

We next investigated whether these results could be replicated observationally in the Estonian Biobank. Tyrosine and phenylalanine levels were measured in 10809 individuals including 359 patients with NAFLD (obtained from EHR). Supplementary Figure 6 presents the distribution of tyrosine and phenylalanine levels in cases and controls. Table 2 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).

Given the important association with tyrosine levels and the liver's significant contribution to tyrosine degradation, (e.g. produces intermediate precursors for gluconeogenesis and ketogenesis), we used a similar approach based on IVW-MR to determine the impact of NAFLD exposure on liver expression of genes encoding tyrosine catabolic pathway enzymes using the Genotype-Tissue Expression dataset (GTEx v8). This resource combines whole genome sequencing and bulk liver tissue RNA sequencing of 208 liver samples.[29] We performed IVW-MR, testing the impact of NAFLD on hepatic gene expression of tyrosine aminotransferase (*TAT*), 4-hydroxyphenylpyruvate dioxygenase (*HPD*), homogentisate 1,2-dioxygenase (*HGD*), glutathione S-transferase zeta 1 (*GSTZ1*) and fumaryl acetoacetate hydrolase (*FAH*). This analysis revealed that NAFLD might not have an important effect on the expression of genes involved in tyrosine metabolism, NAFLD, with the exception of *GSTZ1* expression, which appear to be positively associated with NAFLD presence (Supplementary Table 4). Altogether, results of this analysis show 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.[30] After FDR correction, we found that NAFLD was associated with higher levels of eight circulating proteins: fructose-bisphosphatase 1 (encoded by the *FBP1* gene), cathepsin Z (encoded by the *CTSZ* gene), hydroxymethylglutaryl-CoA synthase (encoded by the *HMGCS1* gene), argininosuccinate lyase (encoded by the *ASL* gene), alpha-L-iduronidase (encoded by the *IDUA* gene), glutathione S-transferase alpha 1 (encoded by the *GSTA1* gene), alcohol dehydrogenase 4 (encoded by the *ADH4* gene) and cytochrome p450 oxidoreductase (encoded by the *POR* gene)(Figure 2B and Supplementary Table 5). The association between NAFLD and plasma levels of these circulating proteins was consistent across MR methods and robust to outliers and pleiotropy (Table 1 and Supplementary Figure 7). We also performed IVW-MR testing the impact of NAFLD on hepatic gene expression of the genes encoding these proteins, again using the GTEx dataset and found no impact of NAFLD on the expression of genes encoding these proteins (Supplementary Table 6). Further, in order to gain insight into potential tissue specificity of the genes encoding these proteins, we obtained the tissue-specific gene expression metric (Tau) as described by Kryuchkova-Mostacci and Robinson-Rechavi.[31] 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 several of the genes encoding circulating proteins that may causally be influenced by NAFLD had tissue-specific expression (Tau ≥0.80), including the *ADH4, GSTA1* and *FBP1* (Tau =0.79) genes, which appeared to be liver-specific (Figure 3). Altogether, this analysis revealed additional proteins that are influenced by the presence of NAFLD and that may represent new biomarkers of NAFLD.

Phenotypic consequences of non-alcoholic fatty liver disease

In order to determine if the association between NAFLD and cardiometabolic diseases shows evidence of causality, and to explore whether drugs targeting NAFLD specifically could impact the risk of other human diseases such as cardiometabolic disease, we performed MR across the human disease-related phenome using our genetic instrument for NAFLD. We used IVW-MR to assess the potentially causal relationship between exposure to NAFLD and 853 disease-specific binary traits in the UK Biobank and for 1169 disease-specific binary traits in the FinnGen cohorts. Results for all diseases that passed correction for multiple testing using MR-pheWAS analyses are presented in Figure 4A and Figure 4B, respectively, for the UK Biobank and FinnGen cohorts. Disease associated with genetically predicted NAFLD include mostly digestive phenotypes such as portal hypertension, liver abscess, oesophageal bleeding, hepatitis, ascites and cirrhosis. Detailed results on all diseases are presented in Supplementary Table 7 and 8. Finally, using the same analytical strategy, we explored the relationship between genetic exposure to NAFLD and COVID-19 diagnosis and complications using GWAS summary statistics from the COVID-19 host genetic initiative.[32] We found no causal association between genetically predicted NAFLD and COVID-19-related hospitalizations or a positive COVID-19 test, both compared to the general population (Supplementary Figure 8 and Supplementary Table 9). Results of these MR analyses performed across the human disease-related phenome suggest that NAFLD was not causally associated with diseases outside the spectrum of liver diseases.

Discussion

Our GWAS meta-analysis combined with a risk factor-informed bGWAS identified 9 candidate genetic regions for NAFLD. This enabled us to establish a MR framework aimed at identifying novel early biomarkers of NAFLD that may be causally impacted by the presence of NAFLD as well as disease-related traits influenced by the presence of NAFLD. This analysis revealed an intriguing effect of NAFLD on tyrosine metabolism and on the presence of eight circulating blood proteins. Our analysis also revealed that NAFLD may not have a causal impact on human diseases outside the spectrum of liver diseases.

Although finding new genetic loci for NAFLD was not a primary objective of this work, we believe that our GWAS meta-analysis revealed important information on the genetic architecture of NAFLD. Our analysis supports the notion that variation at the *MTARC1*, *GCKR*, *TRIB1*, *LMO3*, *SUGP1*, and *PNPLA3* loci may be linked with NAFLD. While genetic variants at most of these loci such as *MTARC1*, *GCKR*, *SUGP1* (where the lead variant was in linkage disequilibrium with a variant at *TM6SF2*), and *PNPLA3* have been associated with some form of liver diseases.[33-36] *TRIB1*, encoding tribbles pseudokinase 1 and *LMO3*, encoding LIM domain only 3 may be new NAFLD loci. However, additional validation and fine-mapping studies will be required, especially for the genetic signal at *LMO3*, which encodes an oncogene that, to our knowledge, has not been previously associated with disease or metabolic traits. One study however suggested that LMO3 might have a role in the development of hepatocellular carcinoma.[37] Variation at *TRIB1* has been associated with cholesterol, triglyceride and liver enzymes levels as well as CAD risk.[38,39] Using bGWAS, our study identified three potentially new loci for NAFLD (*LPL*, *FTO* and *APOE)* that may be associated with NAFLD through their effects on NAFLD risk factors (BMI, T2D and triglycerides). Genetic variation at *APOE* has been linked with NAFLD in another study.[36] 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.[40] Lipoprotein lipase (LPL) on the other hand is a key enzyme that regulates the catabolism of triglycerides-rich lipoproteins in adipose tissue, skeletal muscle and heart. Gain-of-function mutations in LPL were associated with lower triglyceride levels and lower risk for coronary artery diseases.[41]

The majority of previous studies that have linked NAFLD with metabolic or phenotypic traits have used variation at one or two loci to create a NAFLD genetic instrument. This is the case of Lauridsen et al.[42] who have shown that genetically higher liver fat (estimated by *PNPLA3* I148M and the *TM6SF2* E167K genotypes) was not associated with increased risk of ischemic heart disease (IHD) in the general population, despite being strongly associated with the presence of NAFLD. These alleles were also associated with lower plasma LDL cholesterol levels. In another study, however, these variants were associated with a lower risk of coronary artery disease[43] in a large genetic consortium.[41] Both these studies are in contrast with observational studies that provided a positive association between NAFLD and CVD risk.[44,45] These variants were however associated with a higher T2D risk in the ExTexT2D Consortium.[46] Our study also identified a variant at the *MTARC1* locus associated with NAFLD. A previous study by Emdin et al.[34] had already described associations of this gene (then known as *MARC1*) with protection against liver diseases and lower lipid levels. The investigation of isolated variants on metabolic and disease-related traits is prone to pleiotropic association as each variant may cause perturbation in one specific pathway that may cause NAFLD. We therefore used MR to create a multilocus genetic instrument strongly associated with NAFLD in an effort to investigate dose-response associations of NAFLD with metabolic and phenotypic traits while at the same time evaluating and correcting for potential pleiotropic associations. This also enabled us to delineate the direct impact of NAFLD from other causes of NAFLD such as dyslipidemia, insulin resistance or T2D.

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.[47-51] Sliz et al.[52] also documented the individual impact of 4 variants (at the *PNPLA3*, *TM6SF2*, *GCKR* and *LYPLAL1* loci) on the blood metabolome and found inconsistent associations. We investigated whether the presence of NAFLD impacted lipoprotein levels and metabolites of these pathways to identify early biomarkers of NAFLD and to determine whether the results of observational studies could reflect a causal association. Surprisingly, we did not find evidence of a causal association of NAFLD with triglyceride-rich lipoprotein metabolism, glucose-insulin homeostasis or branched-chain amino acids. We did however find an important impact of NAFLD on tyrosine and its metabolic precursor phenylalanine. Although the impact of NAFLD on tyrosine metabolism has been reported decades ago[53], 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.

MR analysis identified eight proteins that may be causally impacted by NAFLD. With the exception of glutathione S-transferase alpha 1 (encoded by the *GSTA1* gene), which has been shown to be a sensitive biomarker of hepatocellular damage[54], few of these proteins have been linked with liver-related disease. However, variation at the *CTSZ* locus, the gene encoding cathepsin Z, has been associated with jaundice-stage progression in primary biliary cholangitis in the Japanese population.[55] Fructose-bisphosphatase 1 (encoded by the *FBP1* gene) is the protein that showed the strongest associated with NAFLD. It is expressed in the liver and lung. It is a gluconeogenesis regulatory enzyme elevated in obesity potentially influenced by dietary fat intake.[56] Among the other liver-expressed proteins identified in this analysis is cytochrome p450 oxidoreductase.[57] POR is a microsomal electron transport protein essential to cytochrome P450-mediated drug metabolism and sterol and bile acid synthesis. ADH4 is also a liver expressed enzyme that mediates oxidative pathways involved in alcohol metabolism.[58] Other, non-liver-specific circulating proteins that appeared to be influenced by the presence of NAFLD included HMGCS1, ASL and IDUA. Interestingly, only one of the eight proteins potentially influenced by the presence of NAFLD has a signal peptide, suggesting that these proteins might not be destined for hepatic secretion and that they may be leaked into the bloodstream following liver damage. Our MR analysis does not suggest a role of NAFLD on the regulation of the genes' expression.

Previous studies have shown that NAFLD could be associated with, or predict the future risk of chronic diseases like CVD, T2D, dyslipidemia and even infectious diseases such as COVID-19. Our MR study design enabled us to explore whether these associations underlie a causal association. Results of our phenome-wide MR analyses in both UK Biobank and FinnGen indicated that NAFLD was not causally associated with diseases outside the spectrum of liver diseases. Although this remains to be demonstrated experimentally, results of this study suggest that the impact of drugs aiming at decreasing NAFLD consequences may improve some liver-associated outcomes, but may not influence the risk of diseases previously associated observationally with NAFLD such as CVD, hypertension, dyslipidemia, chronic kidney disease or COVID-19 complications. Altogether, these results suggest that many biomarkers and diseases previously thought to be caused by NAFLD might be due to secondary causes of NAFLD including abdominal obesity and its associated metabolic

dysfunction. This hypothesis is supported by the study of Liu et al.[27] who have reported a potential causal effect of both T2D and central obesity with NAFLD risk. Along those lines, a cohort study on a Copenhagen population showed that adiposity amplifies the genetic risk of NAFLD.[59] The previously reported association of liver fat accumulation with COVID-19 associated complications may also be confounded by the presence of obesity in patients with NAFLD since a previous MR has suggested a causal effect of a high body mass index on COVID-19 complications.[60]

Our study has limitations. For instance, an EHR-based diagnosis of complex diseases such as NAFLD might be prone to misclassification of cases and controls. We also did not have access to individual patient data to further study gene-environment interactions such as what was reported in the Copenhagen study. Some of the study samples that were used to determine the physiological effects of NAFLD, such as GTEx had a limited number of participants (208 liver samples obtained post-mortem). Studies with a higher number of liver eQTLs will be required to fully appreciate the metabolic effects of NAFLD and its impact on hepatic gene expression. 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 metabolome and 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. 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 in the phenome-wide MR analyses. Finally, EHR-based diagnoses of complex diseases such as NAFLD might be prone to misclassification of cases and controls as NAFLD diagnosis was not confirmed using imaging.

In conclusion, our study identified new NAFLD genetic loci a potentially causal impact of the presence of NAFLD on tyrosine metabolism as well as on blood levels of eight circulating proteins. 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. By exploring the impact of NAFLD on the human disease-related phenome, we found that NAFLD was not associated with diseases outside those of the liver diseases spectrum. Overall, our findings should optimize patient recruitment for NAFLD trials and help predict the outcomes of these trials.

Methods

Genome-wide association study summary statistics NAFLD

To obtain a comprehensive set of NAFLD GWAS summary statistics, we identified three datasets with GWAS summary statistics available for NAFLD identified through electronic health records [25]: The Electronic Medical Records and Genomics (eMERGE) network, the UK Biobank and FinnGen. The NAFLD GWAS eMERGE network has previously been published.[61] 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. 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. A recent study performed in the UK Biobank generated GWAS summary statistics on 1403 disease-specific binary traits in 408,961 white British participants.[62] In this study, a new method called SAIGE (Scalable and Accurate Implementation of Generalized Mixed Models), which is based on generalized mixed models was developed to control for case-control imbalance, sample relatedness and population structure. 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. The EHR code for “Other chronic non-alcoholic fatty liver diseases” (NAFLD) were grouped under phecode 571.5. 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. GWAS was performed using over 28 million genetic markers directly genotyped or imputed by the Haplotype Reference Consortium (HRC) panel with SAIGE, adjusting for sex and birth year. This UK Biobank analysis included 1664 NAFLD cases and 400,055 controls. SAIGE was also used to obtain GWAS summary statistics of the FinnGen cohort. 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 3 (June 16, 2020), 485 patients had a NAFLD diagnosis (EHR code K76.0). They were compared to 135,153 controls. Finally, we performed a GWAS for NAFLD using SAIGE in 142,429 participants of the Estonian Biobank. This study and the use of data from 3460 cases and 138,989 controls was approved by the Research Ethics Committee of the University of Tartu (Approval number 288/M-18). We used the same EHR codes as the UK Biobank to identify NAFLD cases. Age, sex and the 10-main ancestry-based PCs were used as covariates. We performed a fixed-effect GWAS meta-analysis of the eMERGE, UK Biobank, FinnGen and Estonian Biobank cohorts using the METAL package.[63] When variants showed evidence of pleiotropy, we performed a random effect meta-analysis. To identify independent SNPs from this list, SNPs were clumped with plink 1.9 using the 1000 genome population and a $R^2 < 0.1$, a p-value < $<5x10^{-8}$ and a physical distance threshold of 250kb. Regional association plots were obtained from the *gassocplot* R package and the 1000G phase 3 LD reference panel (European ancestry).

Risk-factor informed Bayesian genome-wide association study

We used bGWAS to identify more SNPs associated with NAFLD.[26] 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[27] (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.

Impact of NAFLD variants on blood markers in the UK Biobank

Age, sex and ancestry-based principal components-adjusted GWAS summary statistics on 34 serum biomarker 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. Association of genetically-determined NAFLD and the blood metabolome was assessed using the IVW-MR with the *mr* function from *TwoSampleMR* package in R.[21] The IVW-MR is comparable to performing a meta-analysis of each Wald ratio (the effect of the genetic instrument on eGenes divided by its effect on outcomes).

Impact of NAFLD on the blood metabolome

We used GWAS summary statistics from the study of Kettunen et al.[28] 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.). Additional MR analysis were performed to evaluate heterogeneity (intercept p-value from MR Egger[64]) and the presence of outliers. We used MR-PRESSO[65], 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 10809 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.[30]

Impact of NAFLD on liver gene expression of genes involved in the tyrosine catabolic pathway and proteins influenced by NAFLD

We used data from the Genotype-Tissue Expression Project (GTEx) resource (version 8) to obtain the normalized expression of genes of interest in the liver samples (genes encoding proteins found to be causally influenced by NAFLD or genes encoding enzymes involved in the tyrosine catabolism pathway). GTEx is a large-scale multi-omic dataset where DNA and RNA were collected postmortem from 49 tissue samples from 838 donors. 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.[66] 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.[67] For each gene, expression values were normalized across samples using an inverse normal transformation. Association of genetically predicted NAFLD and the genes of interest was assessed using the IVW-MR.

Tissue-specificity of gene expression and analysis of single-cell sequencing data of human livers

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.[68] 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 $\geq$0.80.

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

We used the same versions of the datasets (UK biobank and FinnGen) as those obtained to derive our NAFLD genetic instrument. In the UK Biobank, outcomes with a case:control ratio <1:1000 were excluded leaving 853 traits for PheWAS. We considered associations that had a p-value $<5.9\times10^{-5}$ (0.05/853 traits) to be statistically significant. In FinnGen, outcomes with <400 cases were excluded leaving 1169 traits for PheWAS. Since several of the phenotypes that were investigated were however genetically correlated, accounting for all phenotypes as they were independent may be too conservative. We therefore used the PhenoSpD tool,[69] to estimate the number of independent tests that are performed. PhenoSpD applies GWAS summary statistics to LD score regression to estimate the phenotypic correlation matrix of the traits and estimates the number of independent variables among the traits. In the UK Biobank, we considered associations that had a p-value $<7.1\times10^{-5}$ (0.05/706 traits) (instead of 853) to be statistically significant. In FinnGen, we considered

associations that had a p-value <$6.5x10^{-5}$ (0.05/773 traits) (instead of 1169) to be statistically significant. In both datasets, we used IVW-MR to determine the association between genetic instruments for NAFLD and disease-specific binary traits.

Impact of NAFLD on COVID-19 diagnosis and hospitalizations

We used GWAS summary statistics from the COVID-19 host genetics initiative that were released on September 30th, 2020. GWAS were performed in each cohort using SAIGE and IVW meta-analysis were performed. We investigated the association of genetically predicted NAFLD and 1) very severe respiratory confirmed COVID-19 versus population (in 9 studies including 2072 cases and 284,472 controls), 2) hospitalized COVID-19 versus population (in 17 studies including 6492 cases and 1,012,809 controls), 3) COVID-19 diagnosis versus lab/self-reported negative (in 22 studies including 11,181 cases and 116,456 controls) and 4) COVID-19 diagnosis versus population (in 32 studies including 17,607 cases and 1,345,334 controls) using IVW-MR.

Data availability

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 the UK Biobank are available here: https://www.leelabsg.org/resources

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

GWAS summary statistics on the 34 blood biomarkers 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.

Gassocplot R package is available at https://github.com/jrs95/gassocplot. 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 >1100 binary phenotypes in the FinnGen cohorts by SAIGE are available to download at https://www.finngen.fi/en/access_results.

The GWAS summary statistics for COVID-19-related diagnoses using SAIGE are available to download at https://www.covid19hg.org/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, 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). 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 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. The work of Estonian Genome Center, Univ. of Tartu has been supported by the European Regional Development Fund and grants SP1GI20181T, SP1GI18045T, No. 2014-2020.4.01.15-0012 GENTRANSMED and 2014-2020.4.01.16-0125. This study was also funded by EU H2020 grant 692145, MP1GI18418R and Estonian Research Council Grant PUT1660. Data analyzes with Estonian datasets were carried out in part in the High-Performance Computing Center of University of Tartu.

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 Ahmed, A., Wong, R. J. & Harrison, S. A. Nonalcoholic fatty liver disease review: diagnosis, treatment, and outcomes. *Clinical Gastroenterology and Hepatology***13**, 2062-2070 (2015).

6 Castera, L., Vilgrain, V. & Angulo, P. Noninvasive evaluation of NAFLD. *Nature reviews Gastroenterology & hepatology***10**, 666-675 (2013).

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

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

9 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).

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

11 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).

12 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).

13 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).

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

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

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

17 Ji, D. *et al.* Non-alcoholic fatty liver diseases in patients with COVID-19: A retrospective study. *Journal of Hepatology* (2020).

18 Roca-Fernandez, A. *et al.* HIGH LIVER FAT ASSOCIATES WITH HIGHER RISK OF DEVELOPING SYMPTOMATIC COVID-19 INFECTION-INITIAL UK BIOBANK OBSERVATIONS. *medRxiv* (2020).

19 Chen, V. L. *et al.* Hepatic Steatosis Is Associated with Increased Disease Severity and Liver Injury in Coronavirus Disease-19. *Digestive diseases and sciences*, doi:10.1007/s10620-020-06618-3 (2020).

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

21 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).

22 Mokry, L. E., Ahmad, O., Forgetta, V., Thanassoulis, G. & Richards, J. B. Mendelian randomisation applied to drug development in cardiovascular disease: a review. *Journal of medical genetics***52**, 71-79 (2015).

23 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).

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

25 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).

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

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

28 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).

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

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

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

32 Initiative, C.-H. G. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European Journal of Human Genetics*, 1 (2020).

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

34 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).

35 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).

36 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).

37 Wu, J., Yang, Y., Wang, X., Zhou, X. & Zhang, C. Modified triple pelvic osteotomy for adult symptomatic acetabular dysplasia: clinical and radiographic results at midterm follow-up. *Journal of orthopaedic surgery and research***13**, 1-7 (2018).

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

39 Waterworth, D. M. *et al.* Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arteriosclerosis, thrombosis, and vascular biology***30**, 2264-2276 (2010).

40 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).

41 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).

42 Lauridsen, B. K. *et al.* Liver fat content, non-alcoholic fatty liver disease, and ischaemic heart disease: Mendelian randomization and meta-analysis of 279 013 individuals. *European Heart Journal***39**, 385-393 (2018).

43 Liu, D. J. *et al.* Exome-wide association study of plasma lipids in> 300,000 individuals. *Nature genetics***49**, 1758-1766 (2017).

44 Ling, S. & Shu-zheng, L. Association between non-alcoholic fatty liver disease and coronary artery disease severity. *Chinese medical journal***124**, 867-872 (2011).

45 Friedrich-Rust, M. *et al.* Severity of coronary artery disease is associated with non-alcoholic fatty liver dis-ease: A single-blinded prospective mono-center study. *PloS one***12**, e0186720 (2017).

46 Mahajan, A. *et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature genetics***50**, 559-571 (2018).

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

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

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

50 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).

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

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

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

54 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).

55 Shimazawa, R. & Ikeda, M. Safety information in drug labeling: a comparison of the USA, the UK, and Japan. *Pharmacoepidemiology and drug safety***22**, 306-318 (2013).

56 Visinoni, S. *et al.* The role of liver fructose-1, 6-bisphosphatase in regulating appetite and adiposity. *Diabetes***61**, 1122-1132 (2012).

57 Kroschwald, P. *et al.* Occurrence of the erythroid cell specific arachidonate 15-lipoxygenase in human reticulocytes. **160**, 954-960 (1989).

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

59 Stender, S. *et al.* Adiposity amplifies the genetic risk of fatty liver disease conferred by multiple loci. *Nature genetics***49**, 842-847 (2017).

60 Ponsford, M. J. *et al.* Cardiometabolic Traits, Sepsis and Severe COVID-19: A Mendelian Randomization Investigation. *Circulation*, doi:10.1161/CIRCULATIONAHA.120.050753 (2020).

61 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).

62 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).

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

64 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).

65 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).

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

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

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

69 Zheng, J. *et al.* PhenoSpD: an integrated toolkit for phenotypic correlation estimation and multiple testing correction using GWAS summary statistics. *GigaScience***7**, doi:10.1093/gigascience/giy090 (2018).

Tables

Table 1. Association of NAFLD with blood metabolites and proteins across multiple Mendelian randomization methods.

Metabolites/proteins	N SNPs	Inverse-variance weighted			Simple median			Weighted median			Mr-Egger		MR_PRESSO outlier test
		Beta	SE	P-value	Beta	SE	P-value	Beta	SE	P-value	Intercept	P-value intercept	P-value
Tyrosine	12	0.109	0.022	6.84E-07	0.116	0.030	9.38E-05	0.111	0.027	5.57E-05	0.009	0.259	0.343
Phenylalanine	12	0.096	0.023	3.79E-05	0.106	0.032	0.001	0.088	0.029	0.003	0.012	0.124	0.310
POR	12	0.205	0.050	4.13E-05	0.204	0.070	0.004	0.204	0.069	0.003	0.000	0.996	0.898
ADH4	12	0.225	0.055	4.01E-05	0.201	0.071	0.005	0.206	0.071	0.004	0.003	0.890	0.399
FBP1	12	0.236	0.050	2.37E-06	0.220	0.075	0.003	0.204	0.071	0.004	0.010	0.561	0.542
IDUA	12	0.239	0.057	2.47E-05	0.198	0.077	0.010	0.299	0.069	1.34E-05	-0.030	0.126	0.284
GSTA1	12	0.208	0.050	3.24E-05	0.238	0.073	0.001	0.222	0.068	0.001	0.000	0.990	0.587
ASL	12	0.218	0.050	1.39E-05	0.186	0.072	0.010	0.176	0.073	0.015	-0.004	0.821	0.531
CTSZ	12	0.234	0.052	5.52E-06	0.214	0.077	0.006	0.209	0.072	0.004	-0.017	0.356	0.422
HMGCS1	12	0.217	0.050	1.41E-05	0.167	0.080	0.037	0.242	0.066	0.001	-0.029	0.127	0.760

Table 2. Impact of tyrosine and phenylalanine levels on non-alcoholic fatty liver disease presence in the Estonian Biobank.

	Odds ratio for NAFLD	P-value
Tyrosine		
Model 1	1.29 (1.18-1.42)	2.09E-08
Model 2	1.23 (1.12-1.36)	2.19E-05
Phenylalanine		
Model 1	1.09 (1.00-1.18)	0.040
Model 2	1.05 (0.95-1.16)	0.347

Model 1 is adjusted for age and sex. Model 2 is adjusted for age, sex, smoking, education and body-mass index. NAFLD indicates on non-alcoholic fatty liver disease.

Figures

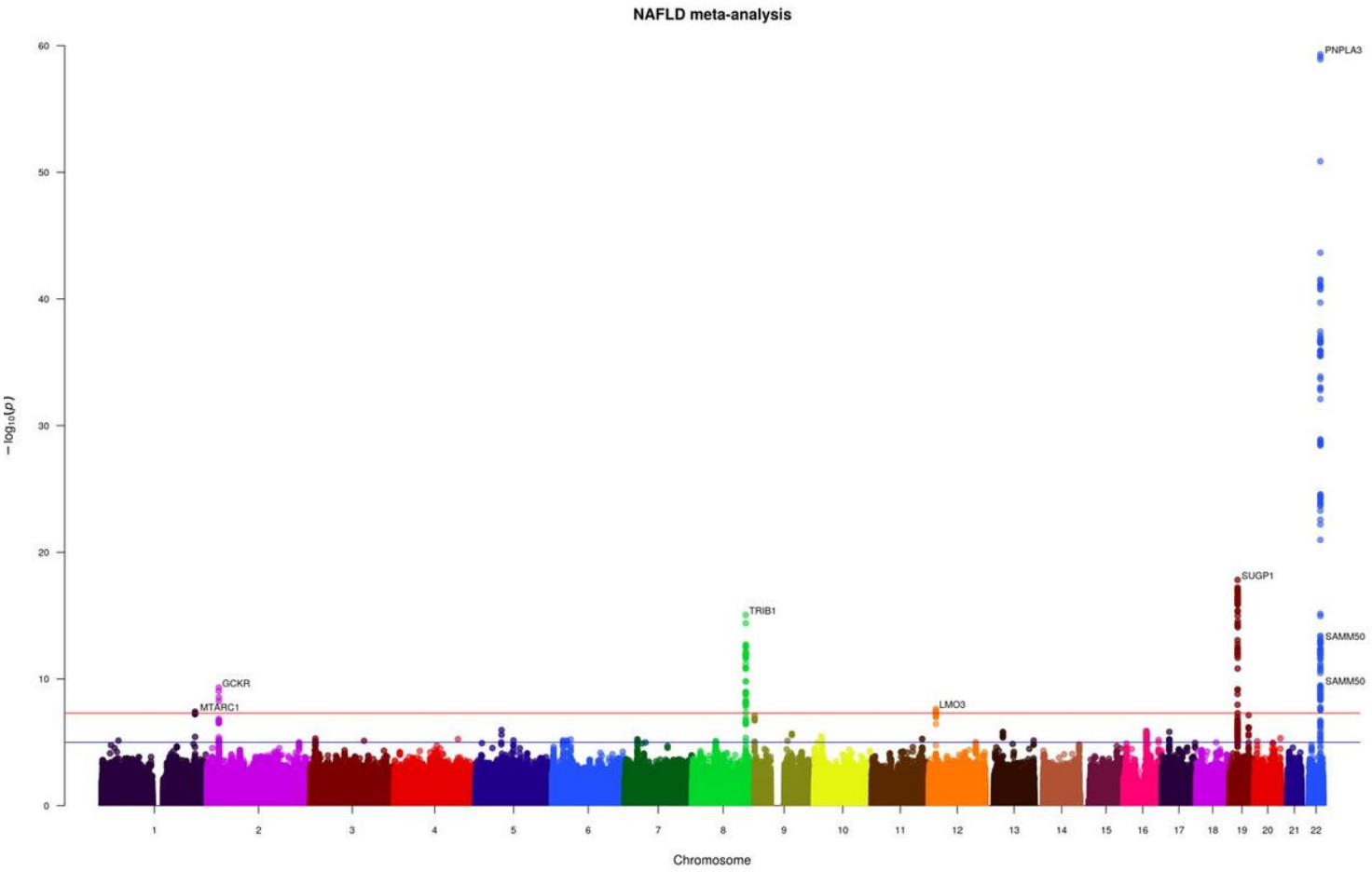

Figure 1

Main results of the meta-analysis of genome-wide association studies (GWAS). 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. Genetic loci harboring SNPs associated with NAFLD ($p<5.0x10-8$) are shown.

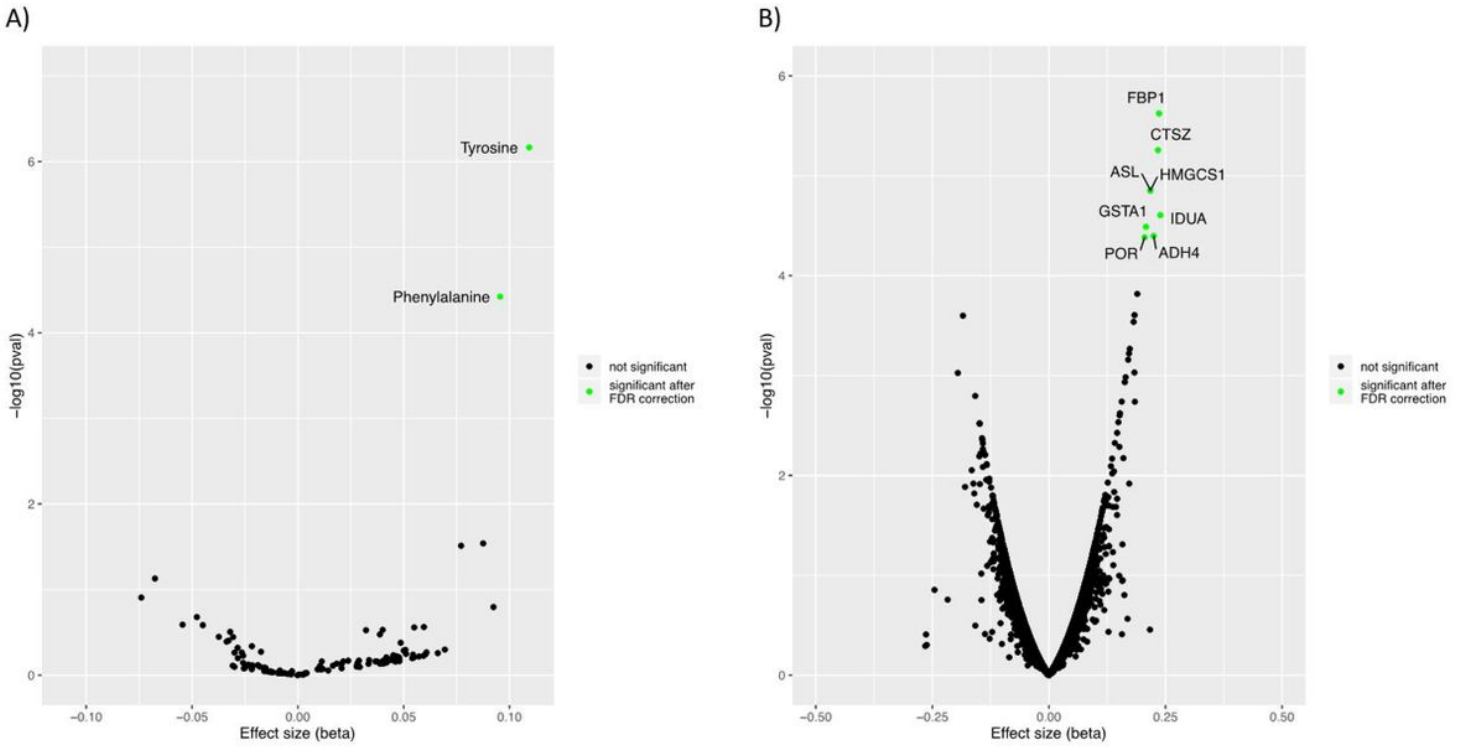

Figure 2

Causal impact of non-alcoholic fatty liver disease (NAFLD) on the blood metabolome and proteome. Volcano plot depicting blood metabolites (A) and blood proteins (B) influenced by the presence of NAFLD. Green dots represent metabolites and proteins significant influenced by the presence of NAFLD following correction for false discovery rate (FDR).

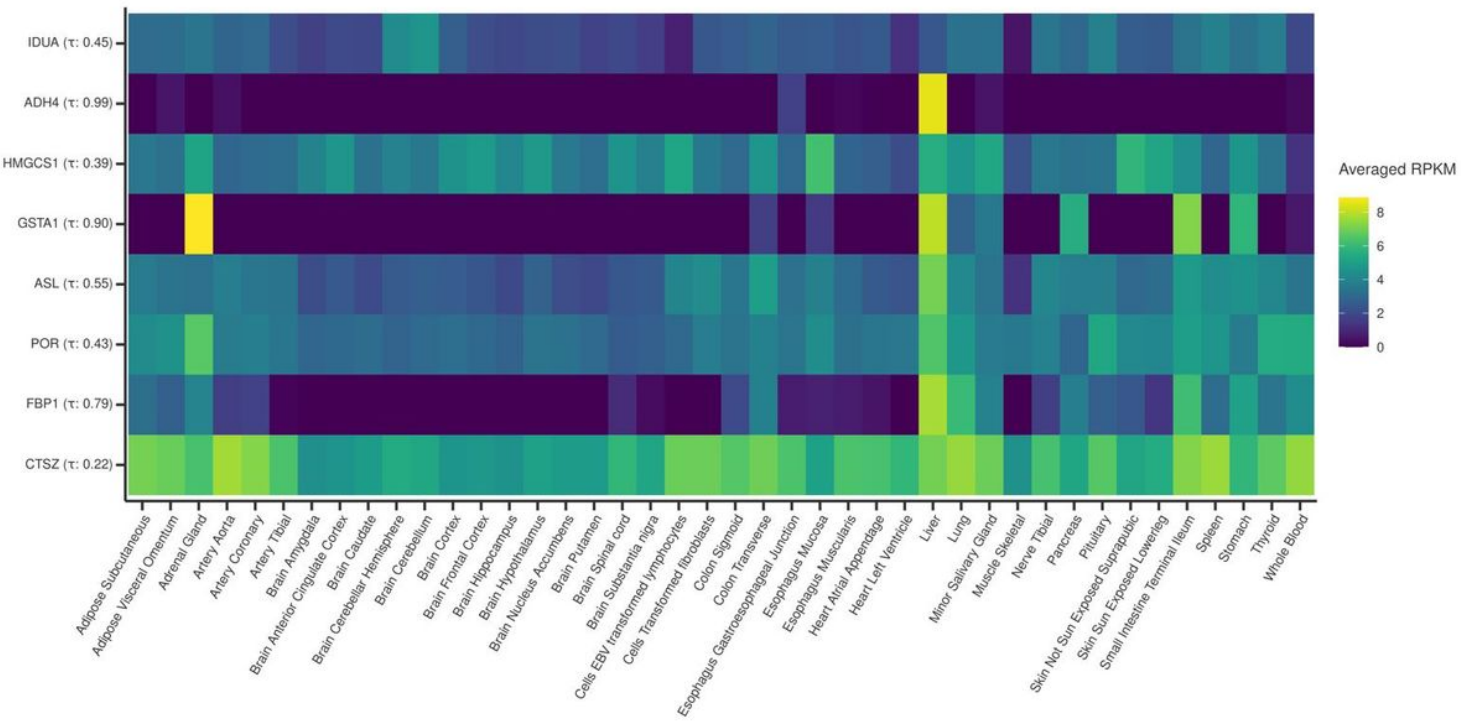

Figure 3

Tissue-specificity of genes encoding proteins influenced by the presence of non-alcoholic fatty liver disease (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. RPKM indicates reads per kilobase per million mapped reads.

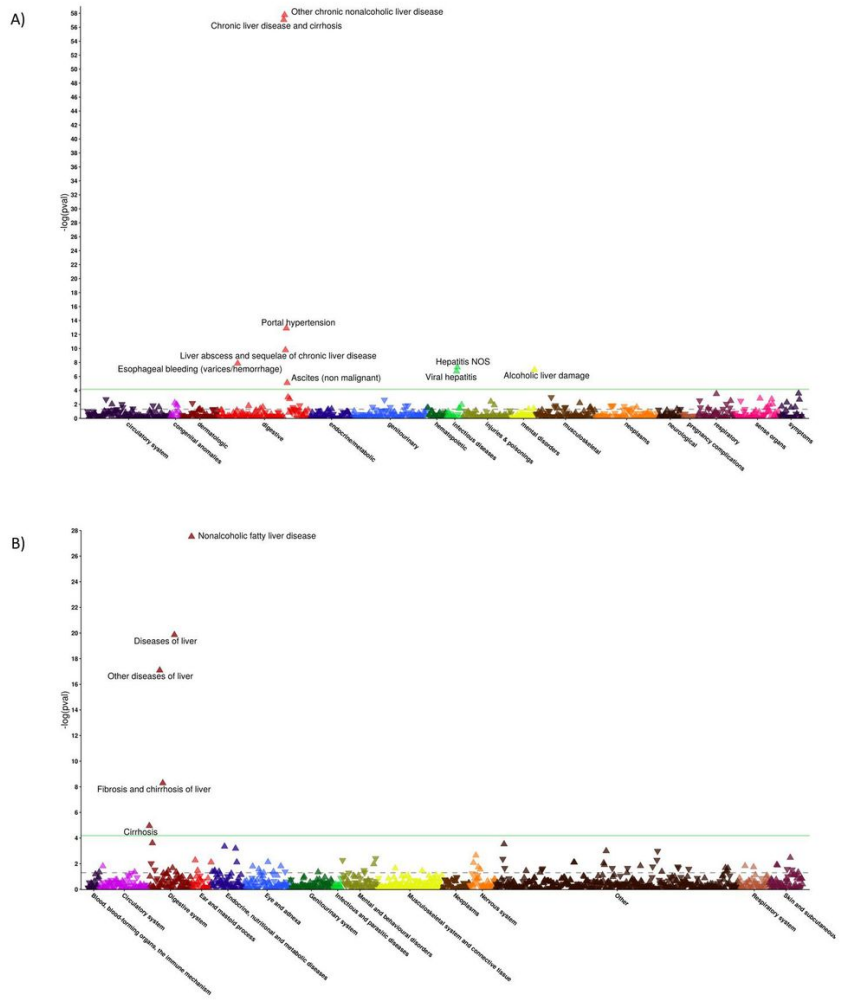

Figure 4

Impact of genetically predicted non-alcoholic fatty liver disease (NAFLD) on the human disease-related phenome. Phenome-wide inverse-variance weighted Mendelian randomization study depicting the association between NAFLD variants (weighted for their impact on NAFLD) and 853 binary disease-related traits in the UK Biobank (A) and 1169 binary disease-related traits in FinnGen (B). Arrows pointing up represent higher disease presence and arrows pointing down represent lower disease presence. The dotted line represents the nominal p-value of 0.05 and the green line represents the p-value after correction for multiple testing.

Supplementary Files

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

- SupplementaryTables.pptx
- SupplFiguresNAFLDMRpaperNG20201006.pptx