Causal relationships between genetically predicted circulating levels of amino acids and NAFLD risk: a Mendelian randomisation study

Jian Zhao¹,²,³*, Jing Zeng⁴, Dong Liu¹, Jun Zhang¹,², Fei Li¹,²,⁵, Giovanni Targher⁶, Jian-Gao Fan⁴,⁷*

1. The Ministry of Education and Shanghai Key Laboratory of Children's Environmental Health, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

2. Department of Maternal and Child Health, School of Public Health, Shanghai Jiao Tong University, Shanghai, China

3. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

4. Department of Gastroenterology, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

5. Department of Developmental and Behavioral Pediatric & Child Primary Care, Brain and Behavioral Research Unit of Shanghai Institute for Pediatric Research, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

6. Section of Endocrinology, Diabetes and Metabolism, Department of Medicine, University and Azienda Ospedaliera Universitaria Integrata of Verona, Verona, Italy.

7. Shanghai Key Lab of Pediatric Gastroenterology and Nutrition, Shanghai, China

*Corresponding authors:

Jian-Gao Fan
No.1665, Kongjiang Road, Yangpu District, Shanghai 200092, China;
Tel: +86(0)21 25077340;
Email address: fanjiangao@xinhuamed.com.cn

Jian Zhao
No.1665, Kongjiang Road, Yangpu District, Shanghai 200092, China;
Tel: +86(0)21 25078864;
Email address: jzhao.epi@gmail.com

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Background and Aims:

Emerging metabolomics-based studies suggested links between amino acids metabolism and NAFLD risk, however, whether there exists an aetiological role of amino acid metabolism in NAFLD development remains unknown. The aim of the present study was to assess the causal relationship between circulating levels of amino acids and NAFLD risk.

Approach and Results:

We performed two-sample Mendelian randomisation (MR) analyses using summary level data from genome-wide association studies (GWAS) to assess causal relationships between genetically predicted circulating levels of amino acids and NAFLD risk. Data from the largest GWAS on NAFLD (8,434 cases and 770,180 controls) were used in discovery MR analysis, and from a GWAS on NAFLD (1,483 cases and 17,781 controls) where NAFLD cases were diagnosed using liver biopsy, were used in replication MR analysis. Wald ratios or multiplicative random-effect inverse variance weighted (IVW) methods were used in the main MR analysis, and weighted median and MR-Egger regression analysis were used in sensitivity analyses. We additionally performed an MR conservative analysis by restricting genetic instruments to those directly involved in amino acid metabolism pathways.

We found that genetically predicted higher alanine (OR=1.45, 95% CI 1.15-1.83) and lower glutamine (OR = 0.81, 95% CI 0.66-1.00) levels were associated with a higher risk of developing NAFLD. Results from MR sensitivity analyses and conservative analysis supported the main findings.

Conclusion:
Genetically predicted higher circulating levels of alanine was associated with an increased risk of NAFLD, whereas higher glutamine was associated with a decreased risk of NAFLD.

**Keywords**

Amino acid, NAFLD, GWAS, Causal effect, Mendelian randomisation
Introduction

Non-alcoholic fatty liver disease (NAFLD) is one of the most prevalent chronic liver diseases, affecting up to ~30% of the general population globally (1). NAFLD has been also predicted to become the most frequent indication for liver transplantation in Western countries by 2030 (2). NAFLD is a progressive disease characterized by the accumulation of lipid droplets within hepatocytes in the absence of excessive alcohol consumption and defined by the presence of at least 5% hepatic steatosis (3). This condition has been consistently reported to be associated with important cardiometabolic comorbidities, including obesity, type 2 diabetes mellitus, cardiovascular disease and stroke (4, 5). To date, although substantial efforts have been put forth to prevent or treat NAFLD, there are no effective preventions or therapeutic treatments for NAFLD.

Emerging metabolomics-based studies have provided insights into mechanisms underlying the development and progression of NAFLD (6, 7). Identifying pathogenic molecules of NAFLD development is essential for improving aetiological understanding and developing novel therapeutic targets for early intervention of this common and burdensome liver disease. It is known that abnormal lipid and glucose metabolism exert putative roles in the pathogenesis of NAFLD (8), whereas recent studies suggested that amino acid metabolism might also contribute to the pathogenesis of NAFLD (9, 10). For example, lower glycine was reported to be associated with higher prevalence of NAFLD (11). Increased levels of aromatic amino acids (AAAs) (e.g., tyrosine and phenylalanine) were found to be associated with increased risk of liver diseases (12). Increased levels of branched chain amino acids (BCAAs), including leucine, isoleucine and valine, have also been reported during the progression of NAFLD (13). In addition, a recent Mendelian randomisation study found a causal effect of NAFLD on blood tyrosine levels (14). Most previous studies have focused on the profiling of amino acids or altered amino acid metabolism in individuals with NAFLD, compared to those without NAFLD, for discovery of non-invasive diagnostic biomarkers. Metabolism of amino acids including BCAAs, alanine, glutamine and tyrosine has been reported to be impacted by NAFLD (13, 15). This implies that altered metabolism of amino acids might be a consequence of NAFLD rather than a causal risk factor for NAFLD. Thus, whether there exists an aetiological role of amino acid metabolism in NAFLD development (i.e., a causal effect of circulating levels of amino acids on NAFLD risk) remains currently unknown.
Mendelian randomisation (MR) is a causal inference approach using germline genetic variants as instrumental variables (IVs), which could largely minimize the risk of bias due to residual confounding or reverse causation (16, 17). In this study, we implement two-sample MR analyses to systematically assess causal effects of genetically predicted circulating levels of amino acids on risk of NAFLD using summary data from both the latest and largest genome-wide association studies (GWASs) of human metabolites and two independent GWASs of NAFLD (Figure 1 Panel A).

Methods

Data sources

Ethics approval has been obtained in original studies that contributed to GWASs on amino acids and NAFLD. All participants provided written informed consent. Declaration of Helsinki statement has been described in the original publications of these studies. The present study only used summary level data from relevant GWASs.

Exposure measure: Amino acids

Summary data for genetic associations with amino acids were retrieved from a recently conducted cross-platform GWAS of 174 metabolites that included up to 86,507 participants (for individual metabolites sample sizes varied from 8,569 to 86,507) (18). Genome-wide association results were meta-analysed in three cohort studies (i.e., the Fenland, EPIC-Norfolk and INTERVAL studies) followed by a further meta-analysis with publicly available GWAS summary data from two studies (19, 20). Of 174 plasma metabolites investigated in the GWAS, 20 circulating levels of amino acids (alanine, arginine, asparagine, aspartate, cysteine, glutamate, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine and valine) were included. We calculated the SNP specific genetic associations with amino acids and their corresponding standard errors based on the original information, including the z-score, sample size and minor allele frequency (MAF), reported in the GWAS of metabolites using the following formula

\[ \hat{b} = \frac{z}{\sqrt{2p(1-p)}} \frac{1}{n+z^2} \text{ and } SE = \frac{1}{\sqrt{2p(1-p)}} \frac{1}{n+z^2} \]

(21), where \( \hat{b} \) is the SNP specific genetic association, \( z \) is the z-score, \( p \) is the MAF of the SNP, \( n \) is the sample size, and \( SE \) is the standard error of the genetic association.

Outcome measure: NAFLD
Two independent datasets on the outcome measure (i.e., NAFLD) were namely retrieved from recently conducted GWASs of NAFLD (22, 23). Data for discovery analysis were obtained from the largest NAFLD GWAS meta-analysis conducted in four cohorts of European ancestry: the Electronic Medical Records and Genomics (eMERGE) network, the UK Biobank (UKB), the Estonian Biobank (EstBB) and the FinnGen (23). In the GWAS meta-analysis on NAFLD, two GWASs of NAFLD were firstly conducted in the UKB and EstBB cohorts, and then combined with results from two publicly available NAFLD GWASs (eMERGE and FinnGen). As a result, 8,434 NAFLD cases were identified by electronic health records (EHR) and 770,180 controls were included in the GWAS meta-analysis (Figure 1 Panel B). Furthermore, for the replication analysis, data were retrieved from a large GWAS of NAFLD in individuals of European ancestry (1,483 cases and 17,781 controls), where NAFLD cases were diagnosed using liver biopsy (i.e., the gold standard method for diagnosing NAFLD) (22) (Figure 1 Panel B).

**Figure 1 here**

**Genetic instruments selection and data harmonization**

Based on the GWAS summary data on cross-platform measured metabolites, 112 genetic variants, which were associated with at least one of 20 circulating amino acids at a metabolome-adjusted genome-wide significance level ($p < 5 \times 10^{-8} / 102 = 4.9 \times 10^{-10}$), were selected as candidate instrument variables (IV). In the present study, a stringent linkage disequilibrium (LD) clumping threshold ($r^2 < 0.01$ and window = 10 Mb) for genetic instruments selection was applied using the “clump_data” function in the TwoSampleMR R package. A total of 111 SNPs (excluding rs61937878) were retained after LD clumping.

Each genetic instrument was looked up in two NAFLD GWASs (for discovery and replication analysis respectively) for SNP-NAFLD associations. SNPs that are in high LD with genetic instruments ($r^2 > 0.8$ and window = 500 Mb) were identified to proxy the absent variants in the discovery and replication NAFLD datasets (14 and 6 proxies were identified respectively). Detailed information on the proxy SNPs can be found in the Supplementary Table S1. Two SNPs (rs142714816, the unique instrument for cysteine, and rs3970551), both of which were absent from the summary data of NAFLD GWAS with no proxy SNPs available, were excluded from the discovery analysis, and one SNP (rs142714816) was removed from the replication analysis.
A data harmonization procedure was performed to merge SNP-amino acid and SNP-NAFLD associations using the “harmonise_data” function in the TwoSampleMR R package (24). Two palindromic SNPs (rs2422358 and rs1935) were removed from further analysis. As a result, a total of 107 and 108 eligible SNPs used as instrumental variables for 19 circulating amino acids were included in the discovery and replication MR analysis, respectively.

Statistical analysis

In both discovery and replication stages, we used Wald ratios (for glutamate and methionine because only one SNP was available for each of these two amino acids), and multiplicative random-effect inverse variance weighted (IVW) approach for all other amino acids as the main MR analysis method, to estimate the causal effect of genetically predicted circulating levels of amino acids on risk of having NAFLD. To increase the statistical power and precision of the causal estimates, a fixed-effect meta-analysis was performed to combine the causal estimates in both discovery and replication stages using the meta R package (25).

Additionally, for certain amino acids that have five or more genetic instruments, we performed several sensitivity analyses, including weighted median and MR-Egger regression analysis to test the consistency of the causal estimates under the different assumptions and to detect possible pleiotropy. Unlike the IVW method that assumes all the SNPs are valid IVs (26), the MR-Egger regression could generate a consistent estimate in presence of invalid genetic instruments, as long as the Instrument Strength Independent of Direct Effect (InSIDE) assumption holds (27). The weighted median method assumes that more than half of the genetic instruments are valid and is a robust approach to outliers (28).

To assess the strength of the selected genetic instruments in MR analysis, we calculated F statistic for each genetic instrument, which are generally considered strong when greater than 10 (29). We used Cochrane’s Q statistic to examine the heterogeneity between SNP-specific causal estimates. Substantial heterogeneity between SNP-specific causal estimates could be indicative of horizontal pleiotropy.

Furthermore, to minimize the risk of bias due to horizontal pleiotropy, we also performed a conservative MR analysis by restricting genetic instruments to those directly involved in amino acid metabolism pathways, as described elsewhere (30). Two sets of genetic variants, namely biologically and genetically prioritised conservative SNPs, were used as instrumental
variables in conservative MR analysis to estimate causal effects using the Wald ratios or fixed-effect IVW method as appropriate.

Given that a total of 20 amino acids were investigated in the present study, after a multiple testing Bonferroni correction, an estimate with a p-value <0.0025 (p=0.05/20) was considered as strong evidence for causal effects, whereas a p-value between 0.0025 and 0.05 indicated a suggestive causal effect. All statistical analyses were undertaken with R version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Characteristics of the included studies and the selected SNPs

Genetic variants instrumenting for amino acids in our study were obtained from a meta-analysis of metabolites GWAS using data from up to 23 cohorts included in previous three GWASs (20, 31, 32) and three independent studies (the Fenland cohort (33), EPIC-Norfolk Study (34), and INTERVAL trial (35)) (Table 1). Average participant age of included studies ranged from 43.5 to 59.8 years old (36). Approximately 50.4% to 53.9% of the study participants were women, except for the GWAS conducted by Shin et al.,(20) where only 16.5% of participants were women. A large-scale meta-analysis of NAFLD GWASs in four studies of European ancestry (the eMERGE, FinnGen, UKB, and EstBB cohorts) included 8,434 NAFLD cases and 770,180 controls and were used for discovery MR analysis (23). Another independent NAFLD GWAS used for the replication analysis included 1,483 NAFLD cases diagnosed with liver biopsy and 47.3% of these participants were women (22).

Table 1 here

The characteristics of the selected SNPs instrumenting for amino acids are presented in Suppportmental Table S2. A total of 133 and 134 genetic variants were used as IVs, ranging from 1 IV (for glutamate and methionine) to 20 IVs (for alanine), to estimate the causal effects of 19 amino acids on NAFLD in discovery and replication MR analysis, respectively. The F-statistics of genetic variants instrumenting for 19 amino acids ranged from 38.7 to 7504.1 (Supplementary Table S2), suggesting a low risk of weak instrument bias. Proportions of variation in amino acids explained by genetic instruments ranged from 0.13% (glutamate) to 10.38% (glycine) (Supplementary Table S3). Cochrane’s Q statistics indicated that there was no significant heterogeneity between SNP-specific causal estimates.
for arginine, aspartate, phenylalanine, proline, and tryptophan in the discovery MR analysis (Supplementary Table S4).

MR main analysis results

Figure 2 here

Of 19 amino acids examined, genetically predicted higher circulating alanine levels were causally associated with an increased risk of NAFLD in both discovery and replication analyses. The odds ratio (OR) of NAFLD was 1.45 (95% CI 1.15-1.83; p = 0.002) per 1-SD increment in alanine levels, after combining causal effect estimates from discovery (OR = 1.37, 95% CI 1.07-1.76; p = 0.012) and replication (OR = 2.06, 95% CI 1.08-3.94; p = 0.029) MR analyses (Figure 2). Additionally, genetically predicted higher circulating glutamine levels appeared to be suggestively associated with a lower risk of NAFLD (OR = 0.81, 95% CI 0.66-1.00; p = 0.048) after meta-analysing estimates from discovery (OR = 0.80, 95% CI 0.64-1.02; p = 0.068) and replication (OR = 0.84, 95% CI 0.55-1.29; p = 0.436) MR analyses (Figure 2). There was little evidence for a causal association between circulating levels of the remaining amino acids and NAFLD risk. Causal effect estimates from the replication MR analysis were broadly consistent with that from the discovery analysis, except for methionine, which had discrepant directions of effect but low precisions.

MR sensitivity analyses results

MR sensitivity analyses, including weighted median and MR-Egger regression analyses, were conducted in 14 amino acids that had at least 5 SNPs as genetic instruments (Supplementary Table S5). A broad consistence was observed when comparing results from sensitivity analyses with those from main analysis presented above, except for several amino acids which had very low precisions in replication MR-Egger regression analysis. Reasons for the low precisions of estimates included smaller sample size of data used in replication analysis and homogeneous SNP-amino acid associations (37). Of note, meta-analysed causal effect estimates derived from weighted median analysis, which is statistically more robust compared to MR-Egger regression, supported potential causal effects of both alanine (OR = 1.58, 95% CI 1.22, 2.04; p < 0.001) and glutamine (OR = 0.81, 95% CI 0.70, 0.93; p = 0.004) on NAFLD risk.

Conservative MR analysis results
By restricting genetic instruments for amino acids to SNPs that were biologically or genetically prioritized in previous published GWAS of metabolites (18), we performed a conservative MR analysis to achieve more reliable causal inference. We were unable to investigate histidine, threonine, methionine and glutamate as genetic variants instrumenting for these amino acids were not on the list of biologically or genetically prioritized genes nor directly involved in relevant metabolism pathway. Results from the conservative MR analysis confirmed a causal role of alanine (OR = 1.80, 95% CI 1.09, 2.97, p = 0.022 for biologically prioritised IVs; OR = 1.93, 95% CI 1.26, 2.96, p = 0.003 for genetically prioritised IVs) and glutamine (OR = 0.83, 95% CI 0.73, 0.96, p = 0.009 for both biologically and genetically prioritised IVs) on the risk of NAFLD (Figure 3).

Figure 3 here

Discussion

In this MR study, we provided novel evidence for a causal role of genetically predicted circulating levels of alanine and glutamine in the development of NAFLD. Specifically, genetically predicted higher alanine and lower glutamine were associated with a higher risk of developing NAFLD. To the best of our knowledge, it is the first study systematically assessing the causal relationships between levels of plasma amino acids and the development of NAFLD using multi-omics (i.e., genomic and metabolomic) data from large-scale human studies (in up to 778,614 individuals).

Previous observational studies mainly focused on the profiling of amino acids or altered amino acid metabolism in individuals with NAFLD compared with those without NAFLD. Metabolism of amino acids including BCAAs (i.e., leucine, isoleucine and valine), alanine, glutamine and tyrosine has been reported to be impacted by NAFLD (13, 15). These findings were beneficial to identifying diagnostic biomarkers of NAFLD, whereas they were not capable to provide causal evidence for aetiological biomarkers of NAFLD development. Thus, our findings provide novel insights into the causal mechanism between altered amino acid metabolism and NAFLD development.

Glutamate is one of the major substrates for the synthesis of glutathione, which is a tripeptide consisting of glutamate, cysteine and glycine and protects tissues from free radical injury via detoxification of active species and/or repair of injury. Since glutamate is poorly transported into cells and glutamine can be efficiently transported across the cell membrane and
deaminated in the mitochondria to produce glutamate and NH3, plasma glutamine is thus important for the generation of intracellular glutamate and consequently glutathione. Experimentally, it has been reported a potential causal role of glutamine administration in decreasing liver injury and mortality in animal studies (38-40). However, there are sparse human studies on the effect of glutamine administration on liver function and its related biomarkers. The present study, from a genetic perspective, provides causal evidence for a protective causal effect of higher circulating glutamine levels on the development of NAFLD. Further, in our MR conservative analysis, we found that only the \textit{GLS-2} (rs2657879) genetic variant predicted glutamine exerting a causal effect on NAFLD risk, compared with another variant (\textit{GLS}, rs7587672) instrumenting for glutamine. Our results were partly supported by findings from a previous study, where the authors found that reducing glutamine metabolism (loss-of-function of \textit{GLS2}) in the liver resulted in decreased severity of hyperglycaemia (increased plasma levels of glutamine and reduced levels of fasting glucose) (41).

Alanine is the predominant amino acid contributing to hepatic gluconeogenesis, therefore, abnormal levels of which generally indicate dysregulation of the alanine-glucose cycle, even further consequent tricarboxylic acid (TCA) and urea cycles (42, 43). Elevated alanine concentrations in NAFLD have been observed in multiple latest metabolomics studies (44), which was consistent with an emerging hypothesis of dysregulated TCA and urea cycles in NAFLD (45, 46). Recently, in the Young Finns Study that examined prospective associations between baseline metabolite levels and the future risk of NAFLD, plasma alanine levels were also found to be positively associated with future onset of NAFLD (47). In our study, results from MR analysis confirmed the positive causal effect of circulating levels of alanine on NAFLD risk.

Among other amino acids, higher levels of BCAAs (including leucine, isoleucine and valine) and AAAs (including phenylalanine, tryptophan, and tyrosine) have been reportedly linked with NAFLD (48-50), however, to our knowledge, we identified only one study in which the prospective associations between baseline concentrations of amino acids and the risk of developing NAFLD during 10-year follow-up were examined (47). Interestingly, in our study, on contrary to the positive associations revealed in the above-mentioned studies, we find little evidence to support a causal effect of both BCAAs and AAAs on NAFLD development. One reason for these discordant results might be reverse causation or confounding bias that cannot be ruled out in previous observational studies. For example, in the prospective Young Finns...
Study, increased plasma tyrosine levels were associated with higher 10-year risk for fatty liver when first adjusted for sex and age, whereas after adjusting for additional baseline confounders, such as waist circumference, alcohol intake, smoking and leisure-time physical activity, this association attenuated and became statistically non-significant (47). Further, in a recent MR study investigating the causal effect of NAFLD on consequent blood metabolites, NAFLD was found to have a positive impact on plasma tyrosine levels (14). Taken together with our results, it seems more plausible to consider altered tyrosine metabolism as a response to the presence of NAFLD rather than an aetiological factor for NAFLD development.

Our study has several strengths. Firstly, this is the first and largest study systematically investigating the causal effects of human circulating amino acids on NAFLD risk, utilizing multi-omics data. Secondly, we leveraged data from an independent GWAS of NAFLD to validate our findings in the discovery population, and combined causal effect estimates from both datasets using meta-analysis to increase statistical power and estimate precision. Thirdly, the conservative MR analysis that was less susceptible to horizontal pleiotropy using genetically and biologically prioritized SNPs as instrumental variables confirmed findings from our MR main analysis. Finally, our results can be generalized to European ancestry as samples span the entire Europe.

We acknowledge some important limitations of our study. Firstly, our study was limited to individuals of European ancestry due to data availability, thus generalizability to other ethnic populations needs to be further investigated. Secondly, although summary data from the largest histology based NAFLD GWAS was used to replicate our findings, results derived from discovery analysis were based on electronic health record (EHR) data where diagnosis of NAFLD may be biased by misclassification of cases and controls due to using hospital records (i.e., ICD-9 and ICD-10 codes). Therefore, future replications in larger cohorts of participants with NAFLD diagnosed with gold standard (i.e., liver biopsy) are warranted.

In conclusion, novel causal biomarkers including alanine and glutamine of NAFLD development were revealed in our study with integrating genomic and metabolomic data. Although further studies are needed, these findings suggest the potential for the glutamine supplementation or alanine depletion for personalized nutrition in NAFLD prevention and treatment.
Author contributions

Jian Zhao: Study conception and design; data curation and analysis; drafting and reviewing the article. Jing Zeng: Data curation; drafting and reviewing the article. Dong Liu: Data curation and analysis; reviewing the article. Jun Zhang: Reviewing the article. Fei Li: Reviewing the article. Giovanni Targher: Drafting and reviewing the article. Jian-Gao Fan: Study conception and design; data curation; drafting and reviewing the article.

Acknowledgements

This study was supported by Xinhua Hospital, Shanghai Jiao Tong University School of Medicine (2021YJRC02). Genetic association estimates for amino acids were obtained by accessing to ‘omicscience’ web (https://omicscience.org). The summary statistics on NAFLD were obtained from GWASs conducted by Ghodsian et al. and Anstee et al., which had been deposited at the GWAS catalog (https://www.ebi.ac.uk/gwas/). The authors would like to thank all investigators for sharing these valuable data.

Conflict of Interest

The authors declare there are no conflicts of interest.

References

46. Sunny NE, Parks EJ, Browning JD, Burgess SC. Excessive hepatic mitochondrial TCA cycle and gluconeogenesis in humans with nonalcoholic fatty liver disease. Cell Metab 2011;14:804-810.
Table 1. Characteristics of included studies.

<table>
<thead>
<tr>
<th>Amino acids (metabolites GWAS)</th>
<th>Studies</th>
<th>Sample size (n)</th>
<th>Female %</th>
<th>Mean age (SD)</th>
<th>Metabolomics platform</th>
<th>Ancestry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fenland study</td>
<td>9,736</td>
<td>53.5%</td>
<td>48.4 (7.4)</td>
<td>Biocrates p180 Kit</td>
<td>European</td>
<td></td>
</tr>
<tr>
<td>EPIC-Norfolk study</td>
<td>5,841</td>
<td>53.3%</td>
<td>59.8 (9.0)</td>
<td>Metabolon HD4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERVAL trial</td>
<td>40,818</td>
<td>50.4%</td>
<td>43.5 (14.2)</td>
<td>Serum NMR platform (Nightingale) and Metabolon HD4</td>
<td>European</td>
<td></td>
</tr>
<tr>
<td>GWAS by Shin et al. 2014</td>
<td>7,824</td>
<td>16.5%</td>
<td>57.1 (11.4)</td>
<td>Metabolon HD1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWAS by Draisma et al. 2015</td>
<td>7,478</td>
<td>53.5%</td>
<td>48.7 (12.7)</td>
<td>Biocrates p150 Kit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GWAS by Kettunen et al. 2016</td>
<td>24,925</td>
<td>53.9%</td>
<td>46.3 (9.7)</td>
<td>Serum NMR platform (Nightingale)</td>
<td>European</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NAFLD (studies for discovery analysis)</th>
<th>Studies</th>
<th>Cases</th>
<th>Controls</th>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>eMERGE</td>
<td>Adults: 710 / pediatrics: 7,725</td>
<td>54.8% / (16.9)</td>
<td>ICD9: 571.5, 571.8, and 571.9,</td>
<td>Alcohol dependence,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>adults: 3,96</td>
<td>44.2%</td>
<td>ICD10: K75.81, K76.0, and K76.9</td>
<td>alcoholic liver disease,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>396</td>
<td>13.1</td>
<td>K76.9</td>
<td>alpha-1 antitrypsin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>846</td>
<td>13.1</td>
<td>K76.9</td>
<td>deficiency, Alagille</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ICD10: K74.0, K74.2, K75.8, K76.0, and</td>
<td>52.1</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K76.9</td>
<td>55.5%*</td>
<td>52.1*</td>
<td>ICD10: K76.0</td>
<td></td>
</tr>
<tr>
<td>UK Biobank</td>
<td>2,558</td>
<td>395,241</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Estonian Biobank</td>
<td>4,119</td>
<td>190,120</td>
<td>NA</td>
<td>ICD10: K74.0, K74.2, K75.8, K76.0, and K76.9</td>
<td></td>
</tr>
<tr>
<td>FinnGen study (data freeze 4)</td>
<td>651</td>
<td>176,248</td>
<td>55.5%*</td>
<td>ICD10: K76.0</td>
<td></td>
</tr>
<tr>
<td>NAFLD (studies for replication analysis)</td>
<td>GWAS by Amstee et al. 2020</td>
<td>1,483</td>
<td>17,781</td>
<td>Liver biopsy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,483</td>
<td>17,781</td>
<td>47.3%*</td>
<td>50.1 (13.0)</td>
<td></td>
</tr>
</tbody>
</table>

MedRxiv preprint doi: https://doi.org/10.1101/2023.02.03.23285451; this version posted February 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

* indicates that the percentage of women or mean age was calculated in the NAFLD case group only.
Figure legends

Figure 1. Schematic overview of the study design and MR analysis. a. rs3970551 was absent from the IV set in the replication NAFLD GWAS by Anstee et al. due to non-available proxy SNPs being identified.

Figure 2. MR main analysis results of the causal effects of genetically predicted circulating levels of amino acids on NAFLD risk.

Figure 3. MR conservative analysis results using genetically and biologically prioritised variants as instrumental variables.
Exposure (19 amino acids)

Primary exposure

- Alanine (20 IVs, N = 86,411)
- Arginine (10 IVs, N = 30,724)
- Asparagine (9 IVs, N = 80,832)
- Aspartate (3 IVs, N = 30,775)
- Glutamate (1 IV, N = 30,977)
- Glutamine (13 IVs, N = 85,819)
- Glycine (15 IVs, N = 80,001)
- Histidine (8 IVs, N = 84,387)
- Isoleucine (1 IV, N = 30,977)
- Leucine (1 IV, N = 30,985)
- Lysine (9 IVs, N = 30,982)
- Methionine (1 IV, N = 30,967)
- Phenylalanine (8 IVs, N = 84,297)
- Proline (2 IVs, N = 30,974)
- Serine (5 IVs, N = 30,955)
- Threonine (7 IVs, N = 85,639)
- Tryptophan (2 IVs, N = 30,974)
- Tyrosine (2 IVs, N = 30,983)
- Valine (6 IVs, N = 86,402)

Genetically conservative set

- Alanine (1 IV)
- Arginine (3 IVs)
- Asparagine (2 IVs)
- Aspartate (1 IV)
- Glutamine (2 IVs)
- Glycine (3 IVs)
- Isoleucine (2 IVs)
- Leucine (1 IV)
- Proline (1 IV)
- Serine (2 IVs)
- Tryptophan (1 IV)
- Tyrosine (1 IV)
- Valine (2 IVs)
- Lysine (1 IV)

Biologically conservative set

- Alanine (2 IVs)
- Arginine (1 IV)
- Asparagine (3 IVs)
- Aspartate (1 IV)
- Glutamine (2 IVs)
- Glycine (4 IVs)
- Isoleucine (1 IV)
- Leucine (1 IV)
- Proline (2 IVs)
- Serine (2 IVs)
- Tryptophan (1 IV)
- Tyrosine (1 IV)
- Valine (1 IV)

Outcome (NAFLD)

Discovery analysis
- NAFLD GWAS by Ghodsian et al 2021 (8,434 cases and 770,180 controls)

Replication analysis
- NAFLD GWAS by Anstee et al 2020 (1,483 cases and 17,781 controls)

Statistical analysis
- IVW
- Weighted median
- MR-Egger

A

B

U

(confounders)

G

(genetic instruments)

X

(amino acids)

X

(NAFLD)

Exposure (19 amino acids)

Genetically conservative set

Biologically conservative set

Outcome (NAFLD)

Discovery analysis
- NAFLD GWAS by Ghodsian et al 2021 (8,434 cases and 770,180 controls)

Replication analysis
- NAFLD GWAS by Anstee et al 2020 (1,483 cases and 17,781 controls)

Statistical analysis
- IVW
- Weighted median
- MR-Egger
Discovery analysis

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>OR of NAFLD per SD change in amino acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>alanine</td>
<td></td>
</tr>
<tr>
<td>isoleucine</td>
<td></td>
</tr>
<tr>
<td>tryptophan</td>
<td></td>
</tr>
<tr>
<td>leucine</td>
<td></td>
</tr>
<tr>
<td>aspartate</td>
<td></td>
</tr>
<tr>
<td>tyrosine</td>
<td></td>
</tr>
<tr>
<td>methionine</td>
<td></td>
</tr>
<tr>
<td>arginine</td>
<td></td>
</tr>
<tr>
<td>proline</td>
<td></td>
</tr>
<tr>
<td>glycine</td>
<td></td>
</tr>
<tr>
<td>phenylalanine</td>
<td></td>
</tr>
<tr>
<td>lysine</td>
<td></td>
</tr>
<tr>
<td>histidine</td>
<td></td>
</tr>
<tr>
<td>asparagine</td>
<td></td>
</tr>
<tr>
<td>valine</td>
<td></td>
</tr>
<tr>
<td>threonine</td>
<td></td>
</tr>
<tr>
<td>serine</td>
<td></td>
</tr>
<tr>
<td>glutamine</td>
<td></td>
</tr>
<tr>
<td>glutamate</td>
<td></td>
</tr>
</tbody>
</table>

Replication analysis

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>OR of NAFLD per SD change in amino acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>alanine</td>
<td></td>
</tr>
<tr>
<td>isoleucine</td>
<td></td>
</tr>
<tr>
<td>tryptophan</td>
<td></td>
</tr>
<tr>
<td>leucine</td>
<td></td>
</tr>
<tr>
<td>aspartate</td>
<td></td>
</tr>
<tr>
<td>tyrosine</td>
<td></td>
</tr>
<tr>
<td>methionine</td>
<td></td>
</tr>
<tr>
<td>arginine</td>
<td></td>
</tr>
<tr>
<td>proline</td>
<td></td>
</tr>
<tr>
<td>glycine</td>
<td></td>
</tr>
<tr>
<td>phenylalanine</td>
<td></td>
</tr>
<tr>
<td>lysine</td>
<td></td>
</tr>
<tr>
<td>histidine</td>
<td></td>
</tr>
<tr>
<td>asparagine</td>
<td></td>
</tr>
<tr>
<td>valine</td>
<td></td>
</tr>
<tr>
<td>threonine</td>
<td></td>
</tr>
<tr>
<td>serine</td>
<td></td>
</tr>
<tr>
<td>glutamine</td>
<td></td>
</tr>
<tr>
<td>glutamate</td>
<td></td>
</tr>
</tbody>
</table>

Meta-analysis

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>OR of NAFLD per SD change in amino acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>alanine</td>
<td></td>
</tr>
<tr>
<td>isoleucine</td>
<td></td>
</tr>
<tr>
<td>tryptophan</td>
<td></td>
</tr>
<tr>
<td>leucine</td>
<td></td>
</tr>
<tr>
<td>aspartate</td>
<td></td>
</tr>
<tr>
<td>tyrosine</td>
<td></td>
</tr>
<tr>
<td>methionine</td>
<td></td>
</tr>
<tr>
<td>arginine</td>
<td></td>
</tr>
<tr>
<td>proline</td>
<td></td>
</tr>
<tr>
<td>glycine</td>
<td></td>
</tr>
<tr>
<td>phenylalanine</td>
<td></td>
</tr>
<tr>
<td>lysine</td>
<td></td>
</tr>
<tr>
<td>histidine</td>
<td></td>
</tr>
<tr>
<td>asparagine</td>
<td></td>
</tr>
<tr>
<td>valine</td>
<td></td>
</tr>
<tr>
<td>threonine</td>
<td></td>
</tr>
<tr>
<td>serine</td>
<td></td>
</tr>
<tr>
<td>glutamine</td>
<td></td>
</tr>
<tr>
<td>glutamate</td>
<td></td>
</tr>
</tbody>
</table>

Method

- Random-effects IVW
- Fixed-effects IVW
- Wald ratio
- Meta-analysis estimates

CC-BY-NC 4.0 International license

The copyright holder for this preprint this version posted February 7, 2023.

https://doi.org/10.1101/2023.02.03.23285451 doi: medRxiv preprint