Somatically mutated genes in fatty liver disease have minimal influence on germline risk

Jake P. Mann* (1,2), Matthew Hoare* (2,3)

Affiliations
1. Institute of Metabolic Science, University of Cambridge, Cambridge, UK
2. School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom;
3. CRUK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom;

Correspondence:
jm2032@cam.ac.uk; mwh20@cam.ac.uk

School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom, Cambridge, CB2 0QQ, United Kingdom

Keywords:
NAFLD, alcohol-related liver disease, GPAM, variation, genomic analysis, precision medicine

Number of main figures: 2; number of main tables: 4

Authors’ declaration of personal interests:
MH is a co-inventor on a patent detailing the finding of recurrent somatic mutations in chronic liver disease.

Funding:
JPM is supported by a Wellcome Trust fellowship (216329/Z/19/Z); MH is supported by a CRUK Advanced Clinician Scientist fellowship (C52489/A19924); CRUK-OHSU Project Award (C52489/A29681) and CRUK Accelerator award to the HUNTER consortium (C18873/A26813).

Author contributions:
JPM and MH conceived and designed the study. JPM curated and analysed the data. JPM and MH wrote and revised the manuscript.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.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:
Understanding the genetics of liver disease has the potential to facilitate clinical risk stratification. We recently identified six genes and one IncRNA enriched for acquired somatic mutations in patients with NAFLD and alcohol-related liver disease. We hypothesised that germline variation in these genes would be associated with risk of liver disease development and contribute to prognostication.

Methods:
Genome-wide association study (GWAS) summary statistics were extracted from seven studies (>1.7 million participants) for variants near ACVR2A, ALB, CIDEB, FOXO1, GPAM, NEAT1 and TNRC6B for: aminotransferases, liver fat, HbA1c, diagnosis of NAFLD, ARLD, and cirrhosis. Findings were replicated using GWAS data from multiple independent cohorts. A phenome-wide association study was performed to examine for related metabolic traits, using both common and rare variants, including gene-burden testing.

Results:
There was no evidence of association between rare germline variants or SNPs near five genes (ACVR2A, ALB, CIDEB, FOXO1, and TNRC6B) and risk or severity of liver disease. Variants in GPAM were associated with liver fat (p=3.6x10^{-13}), ALT (p=2.8x10^{-39}), and serum lipid concentrations. Variants in NEAT1 demonstrated borderline significant associations with ALT (p=1.9x10^{-11}) and HbA1c, but not with liver fat, as well as influencing waist-to-hip ratio, adjusted for BMI.

Conclusions:
Despite strong selective advantage to acquire somatic mutations at these loci, there was no evidence of an association between germline variation and markers of liver disease, except in
GPAM. Polygenic risk scores based on germline variation alone will not capture prognostic data from genes affected by somatic mutations.
Introduction

Non-alcoholic fatty liver disease affects around 25% of the worldwide population and is emerging as the fastest growing form of liver disease in developed countries [1,2]. It encompasses a spectrum of disease from simple hepatic steatosis, through non-alcoholic steatohepatitis (NASH) to cirrhosis, with the attendant risks of liver failure and hepatocellular carcinoma (HCC). Given the number of patients at risk of development of NAFLD and subsequent health problems, there is an urgent need to stratify patients based upon long-term risks, so that even countries with well-resourced healthcare schemes can cope with the patient numbers predicted to develop end-stage liver disease over the next few decades.

At present, there are a number of clinically-deployed non-invasive tests, including elastography and serum biomarkers that can predict levels of hepatic fibrosis and long-term risk in NAFLD. Significant interest has been generated in the potential of germline genetic variation in the pathogenesis and prognostication of NAFLD[3]. Single-nucleotide polymorphisms (SNPs) near several genes, including PNPLA3[4], HSD17B13[5], and TM6SF2[6] have been shown to be associated with long-term risk of several disease-related outcomes in NAFLD[7], as well as in alcohol-related liver disease (ARLD)[8], through genome-wide association studies (GWAS). These variants are common, but associated with small effect sizes, when studied in isolation. This has suggested the potential for combining multiple genotypes into a polygenic risk score: the analysis of several SNPs in patients at baseline as a predictive tool for long-term outcomes in both NAFLD[9] and ARLD[8].

Similarly, other studies have focussed on rare pathogenic germline variants, such as in APOB the gene that encodes apolipoprotein B, that lead to NAFLD with high penetrance[10]. Although these cases can inform us about underlying disease pathogenesis, their low frequency suggests that
they are unlikely to be informative in the stratification of most patients within the general NAFLD population.

We have recently identified recurrent somatic mutations within the liver of patients with NAFLD and ARLD[11]. Recurrent non-synonymous mutations were seen in 6 coding genes: ACVR2A, ALB, CIDEB, FOXO1, GPAM, TNRC6B, plus the long non-coding (lnc)RNA NEAT1. These mutations were found in multiple hepatocyte clones throughout the liver, with evidence of convergent evolution: the identification of independent clones with mutations in the same genes, suggesting strong selection pressure to acquire these variants in the context of fatty liver disease. Amongst these genes, 3 are involved in lipid metabolism suggesting a potential functional role in disease pathogenesis: CIDEB, mediates fusion and cargo transfer of cytoplasmic lipid droplets[12]; FOXO1, the main transcription factor downstream of insulin[13], but also a major regulator of lipid metabolism[14] and; GPAM, encoding GPAT1, the enzyme catalyzing the initial step in triglyceride synthesis[15]. Through functional validation of the hot-spot mutations in FOXO1, we identified that this impacted upon the response to insulin, glycolysis and lipid metabolism. Although promising for improving the mechanistic understanding of disease pathogenesis, the prognostic or therapeutic role that these acquired mutations could play in NAFLD and ARLD remains unknown.

As there was strong selective pressure to acquire these mutations, we were interested to explore whether common or rare germline variation near these recurrently mutated genes might be associated with liver-related outcomes in NAFLD. We hypothesised that germline variation providing weaker modulation of disease phenotypes would exist and might improve the performance of polygenic risk scores being developed for prognostication in NAFLD and ARLD.
Amongst these genes, germline coding variants GPAM have been associated with serum ALT levels in an exome-wide association study of the UK Biobank cohort[16]. Further study found that these variants were also associated with hepatic fat content and histological markers of liver damage in independent NAFLD cohorts. Other studies using GWAS have replicated these findings in further cohorts[17–19], where rs10787429 in GPAM was associated with elevated ALT in both NAFLD and ARLD. Here we use data from multiple GWAS to investigate whether similar associations are observed for the other 5 recurrently mutated genes and one recurrently mutated locus.
Methods

Identification of genes enriched for somatic mutations in ARLD and NAFLD

As described in detail elsewhere[11], whole genome sequencing was performed on diseased, non-malignant hepatic tissue from patients with NAFLD and ARLD undergoing tumour resection or liver transplant. The dN/dScv method[20] was used to identify six coding genes with higher numbers of nonsynonymous mutations relative than expected.

Annotation of somatic mutants and genomic regions

After removal of duplicates, somatic mutations at ACVR2A, ALB, CIDEB, FOXO1, GPAM, and TNRC6B, plus one non-coding IncRNA NEAT1[11] were annotated with predicted consequence, impact on transcript, and prevalence within the 1000 Genomes dataset[21] (n=2,504), Exome Sequencing Project[22] (n=6,503), and gnomAD[23] (n=141,456) using the Ensembl Variant Effect Predictor[24]. For the variants that had previously been identified in any of the above population genomic sequencing datasets, we searched Phenoscanner[25] and the Common Metabolic Disease Portal[26] for any evidence of association with metabolic traits. No data was available on Phenoscanner for these ultra-rare variants. The Common Metabolic Disease Portal search yielded 96 variant-trait associations therefore, rather than apply genome-wide association significance cut-off, the critical p-value for significance for this analysis only was p<5.2x10^{-4} (i.e. 0.05 / 96).

In addition, we annotated each of the six genes with their expected and observed number of predicted loss of function (pLoF) and missense mutations from gnomAD[23]. In brief, the ratio between observed and expected pLoF mutants is an indicator of each gene’s tolerance to haploinsufficiency. Such that a gene with a low observed / expected ratio (defined as the upper
95% confidence interval <0.35) has fewer pLoF mutants than other genes, suggesting that humans are relatively intolerant of haploinsufficiency.

Association with markers of liver disease from genome-wide association studies

Summary statistics from genome-wide association studies (GWAS) were searched for all variants within the genomic regions for the seven regions of interest (SupTable 1). Regions included were (GRCh37): ACVR2A: chr2:148,602,086-148,688,393; ALB: chr4:74,262,831-74,287,129; CIDEB: chr14:24,774,302-24,780,636; FOXO1: chr13:41,129,804-41,240,734; GPAM: chr10:113,909,624-113,975,135; NEAT1: chr11:65,190,245-65,213,011; TNRC6B: chr22:40,440,821-40,731,812.

In addition, for comparison, we extracted summary statistics for four well-validated genome-wide significance risk variants for fatty liver disease: rs738409C>G in PNPLA3[4], rs58542926C>T in TM6SF2[6], rs2642438A>G in MTARC1[27], and rs72613567TA>T in HSD17B13[5]. However, for some studies data on rs72613567TA>T was not available therefore we used rs13125522A>G as a proxy, which is in strong linkage disequilibrium (R²=0.97) with rs72613567[28].

Summary statistics were obtained from the Pan-UK BioBank analysis[29] for alanine aminotransferase (ALT), aspartate aminotransferase (AST), glycosylated haemoglobin (HbA1c), diagnosis of NAFLD (phecode-571.5), diagnosis of ARLD (phecode-317.11), liver fibrosis (icd10-K74), cirrhosis (phecode-571), and other liver diseases (phecode-571.5).

We also obtained summary statistics for diagnosis of NAFLD from Anstee et al.[30] and MRI liver fat from Liu et al.[31], which utilises UK BioBank data. These data were accessed through GWAS
Catalogue[32]. For replication of findings, we first obtained categorical data on liver-related diagnoses from the FinnGen study: NAFLD, ARLD, hepatocellular carcinoma and intrahepatic cholangiocarcinoma, and cirrhosis. For replication of observations for ALT, we obtained data from BioBank Japan[33] and Pazoki et al.[34]. For replication of findings for HbA1c, we obtained summary statistics on HbA1c from the MAGIC consortium (trans-ancestry meta-analysis by Chen et al.[35]) and BioBank Japan, plus diagnosis of type 2 diabetes mellitus (T2DM) in East Asian individuals from Spracklen et al.[36]. The total number of unique participants included from these GWAS summary statistics was 1,628,945.

All variants from the above regions were extracted, and coordinates from FinnGen were carried over from GRCh38 to GRCh37 using the Ensembl Assembly Converter. Manhattan plots were produced for each trait, illustrating only variants within the regions of interest. Significance was defined as $p<5\times10^{-8}$. Within regions that had variants with significant association, we used FIVEx to look for expression quantitative trait loci (eQTL), which extracts data from the European Bioinformatic Institute eQTL Catalogue[37].

Gene-based phenome-wide association study for common variants

We sought to explore associations between germline variation in the coding genes and lncRNA of interest and metabolic traits using a phenome-wide association study approach. Phenoscanner[25] and the Common Metabolic Diseases Knowledge Portal[26] were searched for each of the six coding genes plus NEAT1. Rare variants (mean allele frequency (MAF) <0.01) were excluded and results were filtered for traits relevant to ARLD and NAFLD. Significance was defined as $p<5\times10^{-8}$. Data on eQTLs was obtained using the Qtlizer package for R[38] for all significant associations.
We also searched for any metabolite-wide associations within the regions of interest using data from Lotta et al.[39] (https://omicscience.org/apps/crossplatform/); however, we did not identify any significant (p<4.9x10^{-10}, as defined by the authors) associations.

Association between rare coding variants and traits related to NAFLD or ARLD

We next investigated whether rare coding variants individually, or combined using gene-burden analyses, were associated with markers of liver disease or related metabolic traits. We used data from https://genebass.org/ [40] and https://azphewas.com/ [41], which derive data from the UK BioBank 300k Exomes. These analyses were not available for the lncRNA NEAT1. Extracted associations for: individual variants and gene-burden tests for predicted loss of function (pLoF), missense, and synonymous variants. Significance, as defined by the original studies, was p<2.5x10^{-8} for GeneBass (using SKAT-O test) and p<2.0x10^{-9} (-log10(8.7)) for AZPheWAS.

Analyses

Data were analysed using R 4.0.2[42] and the code used in analyses is available from https://doi.org/10.5281/zenodo.4656979.
Results

We have recently identified recurrent somatic mutations in non-malignant liver tissue from individuals with ARLD and NAFLD through laser capture microdissection and whole genome sequencing[11]. Through this approach we identified six coding genes (ACVR2A, ALB, CIDEB, FOXO1, GPAM and TNRC6B) and one lncRNA (NEAT1) significantly enriched for acquired somatic variants. We hypothesised that given the selective advantage these variants must endow, rare pathogenic variants of these regions, associated with features of liver disease, might be identified.

Specific somatic mutations in recurrently mutated genes are not found in the germline

We previously identified 129 unique variants across these six coding genes and one lncRNA (Supplementary Table 2), the majority of which were either missense (49/129, 38%) or non-coding exonic variants in NEAT1 (49/129, 38%), and predicted to have a moderate or high impact upon protein structure (61/129, 47%, Table 1). These variants are extremely rare in the germline, with 118/129 (92%) having never been identified across 150,463 individuals from gnomAD, 1000G, or the Exome Sequencing Project. The most common of these eleven previously reported variants was rs368997599 G>A (p.Arg45Trp) in CIDEB, which was identified in 10 individuals from gnomAD in the heterozygous state, 7 of whom are of Ashkenazi Jewish ancestry. Even in this genetic ancestry p.Arg45Trp in CIDEB remained an ultra-rare mutation with allele frequency = 7.0x10^-4. There was no evidence of associations between any of the previously identified variants and metabolic traits in the Common Metabolic Disease Portal (Supplementary Table 3).
Whilst all of the 6 coding genes had the expected number of missense mutations, four of the six (ACVR2A, ALB, FOXO1, and TNRC6B) are under selection pressure to prevent against haploinsufficiency; in gnomAD there were significantly fewer predicted loss of function (pLoF) variants observed than expected (Table 1). This implies that there are a reduced number of germline variants in these genes that will cause pLoF, compared to other coding genes and potential haploinsufficiency in these genes is deleterious. The exception to this was CIDEB, where there was an expected number of missense and pLoF mutations, but no reports of these rare variants associated with liver phenotypes. Overall, the specific somatic mutations that we had previously identified are very rare in the germline, due to negative selection pressure and not known to be associated with the development of liver disease.

Among recurrently mutated genes only germline exonic variation in GPAM is associated with liver phenotypes

Given the rarity (or absence) of these specific 129 variants in the germline, we investigated whether other rare variants at these loci were associated with liver disease (or related metabolic traits). We used data from https://genebass.org/ and https://azphewas.com/ (n=281,852 from UK Biobank), which tests whether exonic germline variants either individually, or cumulatively using a gene-based burden method, demonstrated an association with phenotypes.

Analysis of individual coding variants found that p.Ile43Val in GPAM was associated with differences in serum lipids and the synonymous mutation 10-112157327-T-A (p.Pro681Pro) in GPAM was associated with ALT (Table 2). Exonic variants in other genes were associated with related metabolic traits, but not with liver disease phenotypes (Table 2 & Supplementary Table 4). Using gene-based burden testing, which adds together all variants within a single category (e.g. pLoF or missense), variants in ALB demonstrated associations with serum lipids, but no other markers of liver disease. Burden testing for pLoF variants in TNRC6B demonstrated a
significant association with alcohol consumption habits, but no other markers of alcohol-related
disease (Supplementary Table 4). Therefore, amongst the 6 recurrently mutated genes only rare
germline coding variants in GPAM have been identified to be associated with serum lipid levels,
rather than liver phenotypes.

Germline variation at GPAM, but not other somatically mutated genes, is
associated with liver phenotypes

We next explored whether more common germline variants might be associated with liver
phenotypes in previously published datasets. We utilised the summary statistics from the UK
BioBank cohort of 420,000 subjects[29] and searched for variants within these seven regions
(Figures 1-2, Table 3, & Supplementary Figure 1). As previously described[16,17], we observed
genome-wide significant variants within GPAM, associated with elevated serum levels of ALT
(Table 2, e.g. rs10787429 C>T beta=.006, p=2.8x10⁻³⁹), AST, and liver fat by MR imaging (e.g.
rs11446981 T>TA beta=-.003, p=3.6x10⁻¹³). There were no significant associations between
SNPs near any of the other recurrently mutated coding genes and disease correlates in NAFLD
or ARLD.

We validated these associations across a European-ancestry GWAS meta-analysis
(n=753,010)[34] and in the BioBank Japan cohort (n=134,182[33]; Figure 2, Supplementary
Figure 1 & Supplementary Table 5). The magnitude of effect size for the lead variants within
GPAM was similar to that for well-established loci in HSD17B13 and MTARC1, but smaller than
observed for risk variants in PNPLA3 and TM6SF2 (Supplementary Table 6). For example, for
change in hepatic fat[31]: 10-113950257-T-TA in GPAM beta = -.06, p=3.60x10⁻¹³; compared to
rs738409C>G in PNPLA3: beta = 0.22, p=1.5x10⁻¹³³; and rs58542926T>C in TM6SF2 beta =
0.33, p=2.5x10⁻¹¹⁶. Therefore, the only somatically-mutated gene for which germline variation was
associated with liver disease was GPAM.
Germline variation at the long non-coding RNA NEAT1 is associated with liver phenotypes and glycaemic control

In the UK BioBank cohort we observed genome-wide significant variants within the lncRNA NEAT1 for elevated serum ALT (Figure 1). This was replicated in the European ancestry ALT meta-analysis, though not in the BioBank Japan ALT results (Figure 2). The lead variant for NEAT1 (rs595366 T>A) was also found to have an eQTL for NEAT1 in both adipose and liver tissue.

However, no associations were identified between variants at NEAT1 and categorical definitions of liver disease (e.g. diagnosis of NAFLD or ARLD, Figures 1-2) or severity of liver disease (e.g. cirrhosis). This was consistent across data from the UK BioBank and FinnGen (Supplementary Figure 1 and Supplementary Table 5) datasets.

Given the causal associations between insulin resistance and hepatic steatosis, we investigated whether variants were associated with HbA1c or a diagnosis of type 2 diabetes mellitus (T2DM). Variants in NEAT1 were associated with HbA1c levels in the UK BioBank (e.g. rs34743766 C>CA, beta=0.1, p=8.5x10⁻⁹) and had an associated eQTL for NEAT1 in adipose tissue, but this was not replicated for diagnosis of T2DM or other analyses of HbA1c (Supplementary Figure 1). These associations are interesting as expression levels of NEAT1 have previously been associated with both the presence of diabetes[43] and the progression of diabetic complications[43,44].

Germline variation at recurrently mutated genes is associated with metabolic phenotypes

We then explored whether common variants in or near these regions were associated with other related metabolic traits that affect patients with NAFLD, using a gene- (or region-) based
phenome-WAS from two sources (Phenoscanner[25] and Common Metabolic Disease Portal[26], n=7,637-1,546,260, Supplementary Table 7). In addition to its association with ALT, variants in or near GPAM influenced the concentration of many serum lipids (Table 4), including LDL cholesterol and triglyceride levels. We also identified variants near NEAT1 that were associated with a diagnosis of T2DM, waist-to-hip ratio and had a significant eQTL for NEAT1 in adipose tissue. In addition, variants in or near TNRC6B were associated with BMI and creatinine, as were those in ACVR2A. Therefore, germline variation in these genes may impact on related phenotypes that accompany the metabolic syndrome.
Discussion

Investigating the genetics of fatty liver disease has the potential to inform our understanding of disease biology and facilitate clinical risk stratification for affected patients[3]. We recently identified six protein-coding genes and one IncRNA (NEAT1) that are enriched for somatic mutations in patients with NAFLD or ARLD[11]. In this study we find that germline variation in only one, GPAM, was robustly associated with markers of liver disease. Therefore, genetic risk scores capturing only germline variation will not include the pathogenicity conveyed by these regions and additional strategies may be needed to understand the prognostic implications of these somatic mutations.

For five (ACVR2A, ALB, CIDEB, FOXO1, TNRC6B) of the six protein-coding genes under investigation in this study, we found no evidence for association between germline variation and liver disease. These observations were consistent across multiple data sources, traits, genetic ancestries, and analysis methodologies, for example both common variant analyses and rare-variant gene burden testing. This was an unexpected observation, given the genetic evidence for their selective advantage in hepatocyte clones in diseased non-malignant NAFLD and ARLD[11]. Moreover, CIDEB and FOXO1 have well-established functions as a lipid droplet-associated protein[12] and a component of the insulin signalling cascade[14], respectively. Conversely, our previous study did not identify acquired somatic mutations in genes with strong germline associations with liver disease (e.g. PNPLA3, TM6SF2). Collectively, this suggests that the influence of germline and acquired variants on parenchymal liver disease occurs through independent mechanisms (and genes).

It should be noted that the methodology employed in this study does not exclude the possibility that individuals with rare loss of function mutations in these genes may have liver-related phenotypes. Identification and studying human knock-outs for these genes is an alternative
strategy for investigating whether germline variation plays a role in liver disease, as has been illustrated for other conditions\cite{45,46}. However, our data suggest that these individuals will be very rare.

Our analysis has replicated the known associations between variants in or near \textit{GPAM} and liver fat, as well as ALT and AST. Whilst we did not show these variants to influence diagnosis of NAFLD, this has been demonstrated by others\cite{18} with larger sample sizes, variants in \textit{GPAM} (particularly p.Ile43Val) are associated with radiological and histological diagnosis of NAFLD\cite{16–19}. However, It is important to note that variants in \textit{GPAM} have a comparatively small effect size compared to variants in \textit{PNPLA3} and \textit{TM6SF2}, but similar to those in \textit{HSD17B13} and \textit{MTARC1}.

We identified borderline associations between variants in \textit{NEAT1}, a long non-coding RNA (lncRNA), with ALT and HbA1c. Our more broad analyses implicated \textit{NEAT1} in influencing multiple metabolic traits (e.g. serum triglycerides, diagnosis of T2DM, and coronary artery disease), that are of potential relevance to patients with NAFLD. These results point towards a primary role on insulin resistance, potentially through modulation of adipose tissue biology, as several variants in this lncRNA also had significant eQTLs in adipose tissue. The biology of \textit{NEAT1} is poorly understood, but there is some \textit{in vitro} evidence for its role in adipogenesis\cite{47}.

We suggest that the subtle effects of germline variants in \textit{NEAT1} on ALT are likely indirect, via perturbation of insulin resistance and/or development of T2DM\cite{48}, however further work is required to establish this. This data also underlines the principle that the genomic regions enriched for somatic mutations in our original analysis are principally those involved in metabolism.

Clinically, one aim of human genetics is to stratify patients into high- and low- risk groups for disease progression using polygenic gene scores. Such an approach can identify individuals with...
a five-fold increased risk of coronary artery disease[49]. This would be of particular use for NAFLD and ARLD, both common conditions where only a minority of individuals progress to liver-related clinical events. To date there have been four PGS published for liver disease[8,9,50,51], all derived using genome-wide significant hits, and therefore none of our seven genomic regions of interest were included. If a genome-wide PGS were derived[52], which included weighting from sub-genome wide-significant variants, then variants in or near GPAM would contribute. However, they would still receive comparatively minimal weighting compared to variants in PNPLA3 and TM6SF2. More broadly, it is not clear how the magnitude of prognostic implication would compare for germline variation risk scores compared to somatic mutations, as the prognostic implication of these remains unknown. Our results illustrate that integration of somatic mutants into prognostic tools will be a complex process and separate from existing methods for polygenic gene scores.

To understand the role that somatically mutated genes could play in germline PGS there is a need for more independent cohorts with liver-directed exomic or targeted sequencing. This will facilitate PGS derivation and validation against hard clinical outcomes (e.g. liver transplantation) to formally test whether variants in or near ACVR2A, ALB, CIDEB, FOXO1, and TCNR6B are prognostic.

**Conclusion**
Out of seven genomic regions with strong evidence for playing a role in NAFLD and ARLD through somatic mutations, only germline variation in GPAM is associated with liver disease. Other genes analysed, including the IncRNA NEAT1, have a role in metabolic pathways Therefore, genes with pathogenic somatic mutations in NAFLD and ARLD are distinct to those that confer germline risk and would not be captured by polygenic risk scores. Novel approaches will be required to integrate somatic and germline variation with clinical variables for risk prediction algorithms.
Acknowledgements

We want to acknowledge the participants and investigators of FinnGen, UK BioBank, BioBank Japan, EPoS, Million Veterans, and MAGIC Consortium studies.
References


Figure and table legends

Table 1: Characteristics of variants identified as somatic mutants in non-malignant NAFLD and ARLD liver tissue. Features of 129 unique somatic mutants across six protein-coding genes and one IncRNA (NEAT1). Variants were annotated using Variant Effect Predictor to derive a predicted consequence and severity of impact. Germline frequency of variants was assessed across any of gnomAD, 1000G, and ESP (n=150,463). Data on predicted loss of function (pLoF), observed and expected (O/E) and missense variants was obtained from gnomAD v2.

Table 2: Top associations from rare variant and gene-burden analyses. Summary statistics were obtained from analyses of individual rare variants (using UK BioBank 300k Exomes) or gene-burden testing for pLoF or missense variants.

Table 3: Genome-wide significant associations of common variants within regions of interest with eQTLs. Lead variants from the six protein-coding genes and one IncRNA (NEAT1) of interest associated with markers of liver disease or glycaemic control. GWAS summary statistics were obtained from the UK BioBank or Pazoki et al. (2021). Expression quantitative trait locus (eQTL) data were obtained from FIVEx.

Table 4: Associations between common variants in or near regions of interest and related metabolic traits. Gene-based PheWAS performed using PhenoScanner and Common Metabolic Disease Portal to identify common variants significantly associated with the six protein-coding genes and one IncRNA (NEAT1) of interest. eQTL data were obtained using the Qtlizer package for R.
**SupTable 1:** Summary of data sources for genome-wide association study summary statistics used, including number of participants.

**SupTable 2:** List of all included somatic variants with full annotation from Variant Effect Predictor.

**SupTable 3:** Associations between somatic variants previously identified in the germline and metabolic traits from the Common Metabolic Disease Portal. Critical p-value for significance p<5.2x10^{-4}.

**SupTable 4:** Full results from rare exonic variant and gene-burden analyses.

**SupTable 5:** Full results from single common variant analysis for markers of liver disease from all studies (discovery and replication) across all regions of interest.

**SupTable 6:** Comparison of lead variants within GPAM region with well-established genome-wide significant variants across all studied traits.

**SupTable 7:** Full results from gene-based PheWAS for common variants for associated metabolic traits from Phenoscanner and Common Metabolic Disease Knowledge Portal.

**Figure 1:** Association between common variants at recurrently mutated regions and markers of liver disease or glycaemic control. Manhattan plots focusing on the six protein-coding genes and one lncRNA (NEAT1) of interest, illustrating all variants within their genomic coordinates. -log10 p-value was obtained from summary statistics from the UK BioBank for
alanine aminotransferase (ALT), liver fat (from Liu et al., 2021), glycosylated haemoglobin (HbA1c), and alcohol-related liver disease (ARLD). Significance was defined as $p < 5 \times 10^{-8}$.

**Figure 2:** Association between common variants at recurrently mutated regions and markers of liver disease or glycaemic control. Manhattan plots focusing on the six protein-coding genes and one lncRNA ($NEAT1$) of interest, illustrating all variants within their genomic coordinates. $-\log_{10} p$-value was obtained from summary statistics from the UK Biobank for aspartate aminotransferase (AST), hepatic fibrosis, and cirrhosis. Data on diagnosis of NAFLD was obtained from Anstee et al. (2021).

**SupFig1:** Replication analyses for association between common variants and markers of liver disease or glycaemic control. Manhattan plot focusing on the six protein-coding genes and one lncRNA ($NEAT1$) of interest, illustrating all variants within their genomic coordinates. $-\log_{10} p$-value was obtained from summary statistics from Paroki et al. (2021) and BioBank Japan for ALT, Spracklen et al. (2020) for diagnosis of type 2 diabetes mellitus (T2DM), and from FinnGen for diagnosis of alcohol-related liver disease and cirrhosis.
**Table 1**: Characteristics of variants identified as somatic mutants in non-malignant liver tissue. Features of 129 unique somatic mutants across six protein-coding genes and one lncRNA (*NEAT1*). Variants were annotated using Variant Effect Predictor to derive a predicted consequence and severity of impact. Germline frequency of variants was assessed across any of gnomAD, 1000G, and ESP (n=150,463). Data on observed and expected (O/E) predicted loss of function (pLoF) and missense variants was obtained from gnomAD v2.
### Variant-level analyses

<table>
<thead>
<tr>
<th>Gene</th>
<th>Phenotype</th>
<th>Variant</th>
<th>AA change</th>
<th>Allele frequency</th>
<th>Model</th>
<th>Beta</th>
<th>P value</th>
<th>Total</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACVR2A</td>
<td>Creatinine</td>
<td>2-147899548-G-A</td>
<td>p.Pro118Pro</td>
<td>0.30</td>
<td>genotypic</td>
<td>0.03</td>
<td>6.84E-25</td>
<td>254,544</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>Alanine aminotransferase</td>
<td>10-112157327-T-A</td>
<td>p.Pro681Pro</td>
<td>0.28</td>
<td>genotypic</td>
<td>0.03</td>
<td>9.11E-19</td>
<td>255,248</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>Alkaline phosphatase</td>
<td>10-112157327-T-A</td>
<td>p.Pro681Pro</td>
<td>0.28</td>
<td>dominant</td>
<td>0.04</td>
<td>2.61E-24</td>
<td>255,341</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>Apolipoprotein A</td>
<td>10-112180571-T-C</td>
<td>p.Ile43Val</td>
<td>0.27</td>
<td>genotypic</td>
<td>0.04</td>
<td>1.16E-42</td>
<td>231,005</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>Cholesterol</td>
<td>10-112180571-T-C</td>
<td>p.Ile43Val</td>
<td>0.27</td>
<td>genotypic</td>
<td>0.03</td>
<td>2.12E-17</td>
<td>253,523</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>Direct bilirubin</td>
<td>10-112157327-T-A</td>
<td>p.Pro681Pro</td>
<td>0.28</td>
<td>genotypic</td>
<td>0.02</td>
<td>2.61E-25</td>
<td>254,544</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>HDL cholesterol</td>
<td>10-112180571-T-C</td>
<td>p.Ile43Val</td>
<td>0.27</td>
<td>genotypic</td>
<td>0.04</td>
<td>3.95E-43</td>
<td>232,376</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>Hip circumference</td>
<td>10-112180571-T-C</td>
<td>p.Ile43Val</td>
<td>0.27</td>
<td>genotypic</td>
<td>-0.02</td>
<td>1.59E-08</td>
<td>265,294</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>LDL direct</td>
<td>10-112180571-T-C</td>
<td>p.Ile43Val</td>
<td>0.27</td>
<td>genotypic</td>
<td>0.02</td>
<td>1.03E-09</td>
<td>253,059</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>Leg fat mass (right)</td>
<td>10-112180571-T-C</td>
<td>p.Ile43Val</td>
<td>0.27</td>
<td>genotypic</td>
<td>-0.01</td>
<td>3.47E-08</td>
<td>261,190</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>Total bilirubin</td>
<td>10-112157327-T-A</td>
<td>p.Pro681Pro</td>
<td>0.28</td>
<td>genotypic</td>
<td>0.02</td>
<td>1.34E-12</td>
<td>254,316</td>
<td>AZ</td>
</tr>
<tr>
<td>GPAM</td>
<td>Triglycerides</td>
<td>10-112180571-T-C</td>
<td>p.Ile43Val</td>
<td>0.27</td>
<td>genotypic</td>
<td>-0.02</td>
<td>2.07E-13</td>
<td>253,317</td>
<td>AZ</td>
</tr>
<tr>
<td>TNRC6B</td>
<td>Body mass index (BMI)</td>
<td>22-40301373-A-G</td>
<td>-</td>
<td>0.34</td>
<td>genotypic</td>
<td>-0.02</td>
<td>2.19E-10</td>
<td>267,287</td>
<td>AZ</td>
</tr>
<tr>
<td>TNRC6B</td>
<td>Creatinine</td>
<td>22-40301172-TGCA-T</td>
<td>p.Gln524del</td>
<td>0.30</td>
<td>genotypic</td>
<td>0.02</td>
<td>2.58E-09</td>
<td>255,188</td>
<td>AZ</td>
</tr>
<tr>
<td>TNRC6B</td>
<td>Impedance of whole body</td>
<td>22-40301373-A-G</td>
<td>-</td>
<td>0.34</td>
<td>genotypic</td>
<td>0.02</td>
<td>1.93E-16</td>
<td>263,528</td>
<td>AZ</td>
</tr>
<tr>
<td>TNRC6B</td>
<td>Total bilirubin</td>
<td>22-40301172-TGCA-T</td>
<td>p.Gln524del</td>
<td>0.30</td>
<td>genotypic</td>
<td>-0.02</td>
<td>3.87E-08</td>
<td>254,308</td>
<td>AZ</td>
</tr>
</tbody>
</table>

### Gene burden analyses

<table>
<thead>
<tr>
<th>Gene</th>
<th>Phenotype</th>
<th>Burden test</th>
<th>Beta</th>
<th>pval</th>
<th>Total</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALB</td>
<td>Apolipoprotein B</td>
<td>ptv</td>
<td>0.75</td>
<td>7.29E-13</td>
<td>255,235</td>
<td>AZ</td>
</tr>
<tr>
<td>ALB</td>
<td>Cholesterol</td>
<td>pLoF</td>
<td>0.03</td>
<td>3.45E-10</td>
<td>268,558</td>
<td>Genebass</td>
</tr>
<tr>
<td>ALB</td>
<td>Direct bilirubin</td>
<td>mis</td>
<td>0.003</td>
<td>1.86E-08</td>
<td>228,253</td>
<td>Genebass</td>
</tr>
<tr>
<td>ALB</td>
<td>LDL direct</td>
<td>ptv</td>
<td>0.56</td>
<td>5.65E-08</td>
<td>256,046</td>
<td>AZ</td>
</tr>
<tr>
<td>ALB</td>
<td>Total bilirubin</td>
<td>mis</td>
<td>0.003</td>
<td>1.60E-11</td>
<td>267,473</td>
<td>Genebass</td>
</tr>
<tr>
<td><strong>TNRC6B</strong></td>
<td>Frequency of failure to fulfil normal expectations due to drinking alcohol in last year</td>
<td><strong>pLoF</strong></td>
<td><strong>1.67E-24</strong></td>
<td><strong>51,479</strong></td>
<td><strong>Genebass</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2: Top associations from rare variant and gene-burden analyses.** Summary statistics were obtained from analyses of individual rare variants (using UK BioBank 300k Exomes) or gene-burden testing for pLoF or missense variants. AA, amino acid; AZ, Astrazeneca PheWAS; mis, missense variants; pLoF, predicted loss of function; ptv, protein truncating variant; snv, single nucleotide variant.
<table>
<thead>
<tr>
<th>Variant</th>
<th>Gene</th>
<th>Trait</th>
<th>Beta</th>
<th>P value</th>
<th>EAF</th>
<th>Source</th>
<th>n</th>
<th>eQTL gene (tissue)</th>
<th>eQTL beta (p-value)</th>
<th>eQTL_study</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10787429 C &gt; T</td>
<td>GPAM</td>
<td>ALT</td>
<td>0.006</td>
<td>2.80E-39</td>
<td>0.27</td>
<td>European GWAS meta-analysis</td>
<td>753,010</td>
<td>Nil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7096937 T &gt; C</td>
<td>GPAM</td>
<td>AST</td>
<td>-0.02</td>
<td>2.77E-09</td>
<td>0.73</td>
<td>UK BioBank</td>
<td>419,034</td>
<td>Nil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs11446981 T &gt; TA</td>
<td>GPAM</td>
<td>Liver fat</td>
<td>-0.06</td>
<td>3.60E-13</td>
<td>0.70</td>
<td>UK BioBank</td>
<td>32,858</td>
<td>Nil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs595366 T &gt; A</td>
<td>NEAT1</td>
<td>ALT</td>
<td>-0.003</td>
<td>1.90E-11</td>
<td>0.27</td>
<td>European GWAS meta-analysis</td>
<td>753,010</td>
<td>NEAT1 (adipose)</td>
<td>-0.4 (p=6.5x10^-17)</td>
<td>FUSION</td>
</tr>
<tr>
<td>rs34743766 C &gt; CA</td>
<td>NEAT1</td>
<td>HbA1c</td>
<td>0.01</td>
<td>8.49E-09</td>
<td>0.17</td>
<td>UK BioBank</td>
<td>419,446</td>
<td>NEAT1 (adipose)</td>
<td>0.27 (p=9.3x10^-17)</td>
<td>TwinsUK</td>
</tr>
</tbody>
</table>

Table 3: Genome-wide significant associations of common variants within regions of interest with eQTLs. Lead variants from the six protein-coding genes and one IncRNA (NEAT1) of interest associated with markers of liver disease or glycaemic control. GWAS summary statistics were obtained from the UK Biobank or Pazoki et al. (2021). Expression quantitative trait locus (eQTL) data were obtained from FIVEx. EAF, effect allele frequency.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Variant</th>
<th>Trait</th>
<th>Beta (SE)</th>
<th>P value</th>
<th>n</th>
<th>eQTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACVR2A</td>
<td>rs13008838 A&gt;G</td>
<td>eGFR-creat (serum creatinine)</td>
<td>-0.003 (0.0004)</td>
<td>4.97E-15</td>
<td>754,661</td>
<td>ACVR2A: Adipose - Subcutaneous (+), Muscle skeletal (+), Artery - Tibial (+), Artery - Aorta (+), Pancreas (+), Artery - Coronary (+), Adipose - Visceral (Omentum) (+)</td>
</tr>
<tr>
<td>ACVR2A</td>
<td>rs3764955 G&gt;C</td>
<td>Serum creatinine</td>
<td>0.020 (0.003)</td>
<td>2.01E-08</td>
<td>182,901</td>
<td>ACVR2A: Artery - Coronary (+), Adipose - Visceral (Omentum) (+), Adipose - Subcutaneous (+), Pancreas (+), Artery - Aorta (+), Artery - Tibial (+), Muscle skeletal (+)</td>
</tr>
<tr>
<td>CIDEB</td>
<td>rs12590407 G&gt;A</td>
<td>Diastolic blood pressure</td>
<td>-0.010 (0.002)</td>
<td>9.11E-12</td>
<td>552,754</td>
<td></td>
</tr>
<tr>
<td>FOXO1</td>
<td>rs2253001 A&gt;T</td>
<td>Basal metabolic rate</td>
<td>0.010 (0.002)</td>
<td>1.03E-08</td>
<td>331,307</td>
<td></td>
</tr>
<tr>
<td>GPAM</td>
<td>rs10787429 T&gt;C</td>
<td>Alanine transaminase</td>
<td>-0.030 (0.005)</td>
<td>3.44E-09</td>
<td>141,341</td>
<td>GPAM: Artery - Tibial (-)</td>
</tr>
<tr>
<td>GPAM</td>
<td>rs7898213 T&gt;C</td>
<td>Alkaline phosphatase</td>
<td>-0.030 (0.004)</td>
<td>5.00E-11</td>
<td>112,189</td>
<td></td>
</tr>
<tr>
<td>GPAM</td>
<td>rs2792759 C&gt;T</td>
<td>Bilirubin</td>
<td>-0.010 (0.001)</td>
<td>4.10E-09</td>
<td>467,109</td>
<td>GPAM: Artery - Tibial (-)</td>
</tr>
<tr>
<td>GPAM</td>
<td>rs2297991 T&gt;C</td>
<td>HDL cholesterol</td>
<td>-0.030 (0.003)</td>
<td>5.21E-20</td>
<td>191,159</td>
<td>GPAM: Artery - Tibial (-)</td>
</tr>
<tr>
<td>GPAM</td>
<td>rs1129555 A&gt;G</td>
<td>LDL cholesterol</td>
<td>0.030 (0.004)</td>
<td>1.00E-15</td>
<td>-</td>
<td>GPAM: Artery - Tibial (-)</td>
</tr>
<tr>
<td>GPAM</td>
<td>rs1129555 A&gt;G</td>
<td>Total cholesterol</td>
<td>0.030 (0.004)</td>
<td>2.00E-18</td>
<td>-</td>
<td>GPAM: Artery - Tibial (-)</td>
</tr>
<tr>
<td>GPAM</td>
<td>rs2254537 T&gt;A</td>
<td>Triglycerides</td>
<td>0.020 (0.002)</td>
<td>8.23E-17</td>
<td>482,392</td>
<td>GPAM: Artery - Tibial (-)</td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs10896037 A&gt;G</td>
<td>Chronic kidney disease</td>
<td>0.060 (0.009)</td>
<td>5.98E-10</td>
<td>140,966</td>
<td>NEAT1: Adipose - Visceral (Omentum) (-), Muscle skeletal (-), Adipose - Subcutaneous (-), Artery - Aorta (-)</td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs12801636 G&gt;A</td>
<td>Coronary artery disease</td>
<td>-0.040 (0.004)</td>
<td>6.74E-17</td>
<td>1,546,260</td>
<td></td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs2306363 G&gt;T</td>
<td>Diastolic blood pressure</td>
<td>-0.030 (0.002)</td>
<td>1.39E-28</td>
<td>552,754</td>
<td></td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs4930319 G&gt;C</td>
<td>eGFR-creat (serum creatinine)</td>
<td>0.003 (0.0004)</td>
<td>8.69E-27</td>
<td>757,454</td>
<td>NEAT1: Adipose - Subcutaneous (-), Adipose - Visceral (Omentum) (-), Artery - Aorta (-), Muscle skeletal (-)</td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs2236682 G&gt;T</td>
<td>Serum creatinine</td>
<td>0.020 (0.004)</td>
<td>2.66E-10</td>
<td>182,901</td>
<td>NEAT1: Muscle skeletal (-), Artery - Aorta (-), Adipose - Visceral (Omentum) (-), Adipose - Subcutaneous (-)</td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs2306363 G&gt;T</td>
<td>Systolic blood pressure</td>
<td>-0.020 (0.002)</td>
<td>3.90E-20</td>
<td>550,853</td>
<td></td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs10750766 C&gt;A</td>
<td>Triglycerides</td>
<td>0.020 (0.002)</td>
<td>9.96E-15</td>
<td>459,761</td>
<td></td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs947791 G&gt;A</td>
<td>Type 2 diabetes</td>
<td>0.050 (0.004)</td>
<td>2.71E-16</td>
<td>1,016,100</td>
<td>NEAT1: Artery - Aorta (+), Pancreas (+), Muscle skeletal (+), Artery - Tibial (+), Adipose - Visceral (Omentum) (+), Adipose - Subcutaneous (+)</td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs11227217 C&gt;T</td>
<td>Waist-hip ratio</td>
<td>0.020 (0.002)</td>
<td>2.60E-15</td>
<td>997,776</td>
<td>NEAT1: Adipose - Visceral (Omentum) (+), Muscle skeletal (+), Adipose - Subcutaneous (+), Artery - Aorta (+), Artery - Tibial (+), Pancreas (+)</td>
</tr>
<tr>
<td>NEAT1</td>
<td>rs4645917 G&gt;A</td>
<td>Waist-hip ratio adj BMI</td>
<td>-0.030 (0.002)</td>
<td>5.05E-18</td>
<td>969,126</td>
<td>NEAT1: Muscle skeletal (-), Artery - Tibial (-)</td>
</tr>
<tr>
<td>TNRC6B</td>
<td>rs4820410 A&gt;G</td>
<td>BMI</td>
<td>-0.020 (0.001)</td>
<td>1.72E-26</td>
<td>1,326,100</td>
<td>TNRC6B: Artery - Tibial (+)</td>
</tr>
<tr>
<td>Gene</td>
<td>SNP</td>
<td>Trait</td>
<td>p-value</td>
<td>Log10 p-value</td>
<td>Frequency</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------------</td>
<td>--------------------------------</td>
<td>----------</td>
<td>--------------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>TNCR6B</td>
<td>rs2294352 G&gt;A</td>
<td>eGFR-creat (serum creatinine)</td>
<td>0.010</td>
<td>2.90E-32</td>
<td>772,751</td>
<td></td>
</tr>
<tr>
<td>TNCR6B</td>
<td>rs470113 A&gt;G</td>
<td>Pulse pressure</td>
<td>0.020</td>
<td>4.33E-08</td>
<td>317,539</td>
<td></td>
</tr>
<tr>
<td>TNCR6B</td>
<td>rs2294352 G&gt;A</td>
<td>Serum creatinine</td>
<td>-0.030</td>
<td>1.06E-15</td>
<td>191,380</td>
<td></td>
</tr>
<tr>
<td>TNRC6B</td>
<td>rs5995840 T&gt;C</td>
<td>Body mass index</td>
<td>0.030</td>
<td>1.02E-11</td>
<td>173,430</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4: Associations between common variants in or near regions of interest and related metabolic traits.** Gene-based PheWAS performed using Phenoscanner and Common Metabolic Disease Portal to identify common variants significantly associated with the six protein-coding genes and one lncRNA (NEAT1) of interest. eQTL column provides the gene that have significant eQTLs for that variant, the tissues, and the direction (+, positive eQTL; -, negative eQTL). eQTL data were obtained using the Qtlizer package for R.