Title

A multiancestry genome-wide association study of unexplained chronic ALT elevation as a proxy for nonalcoholic fatty liver disease with histological and radiological validation

Permalink

https://escholarship.org/uc/item/8611z9sd

Journal

Nature Genetics, 54(6)

ISSN

1061-4036

Authors

Vujkovic, Marijana
Ramdas, Shweta
Lorenz, Kim M
et al.

Publication Date

2022-06-01

DOI

10.1038/s41588-022-01078-z

Peer reviewed

Published in final edited form as:
Nat Genet. 2022 June ; 54(6): 761–771. doi:10.1038/s41588-022-01078-z.

Reprints and permissions information is available at www.nature.com/reprints.

✉**Correspondence and requests for materials** should be addressed to Benjamin F. Voight bvoight@pennmedicine.upenn.edu or Kyong-Mi Chang kmchang@pennmedicine.upenn.edu.

*Lists of authors and their affiliations appear at the end of the paper.

Author contributions

M.V., S.R., K.-M.C., D.J.R., B.F.V. and P.S.T. were responsible for the concept and design. M.V., S.R., K.-M.C., D.J.R., B.F.V. and P.S.T. performed data acquisition, analysis or interpretation. M.V., S.R., K.-M.C. and B.F.V. drafted the manuscript. Critical revision of the manuscript for important intellectual content was carried out by M.V., K.M. Lorenz, K.M. Lee, M.S., D.E.K., X.Z., C.T., R.L.K., H.R.K., A.V., A.G., D.M.K., Y.V.S., J.E.H., D.R.M., P.D.R., L.S.P., S.M., S.L.D., J.S.L., T.L.A., S.P., K.C., T.L.E., S.M.D., P.W.W., J.M.G., C.J.O., P.S.T., J.B.M., J.A.L., B.F.V., K.-M.C., Q.M.A., R.D., H.J.C., A.K.D., L.A.L., J.B.N., A.E.L., M.B.J., N.V., A.B., S.R., X.G., J. He, Y.G., C.C., R.P.M., C.V.S., J.P., R.M.C., M.E.H., M.T.M., W.R.W., J. Huang, K.T.C., N.J.H., C.-T.L., M.T.L., J.Y., M.B., J.T., X.L., H.J.L., Y.-D.I.C., K.D.T., R.-K.C., R.M.K., S.V., J.B., K.R.R., B.A.N.-T., J.B.S., A.J.S., N.C., K.A.R., B.D.M., D.G., A.D.W., E.M., Y.S., N.M., A.V.K., S.F.A.G., C.D.B., D.S., L.B., J.I.R. and D.J.R. Finally, K.-M.C., D.J.R. and B.F.V. provided administrative, technical or material support. VA MVP: M.V., K.M. Lorenz, K.M. Lee, M.S., D.E.K., X.Z., C.T., R.L.K., H.R.K., A.V., A.G., D.M.K., Y.V.S., J.E.H., D.R.M., P.D.R., L.S.P., S.M., S.L.D., J.S.L., R.M.C., T.L.A., S.P., K.C., T.L.E., S.M.D., P.W.W., J.M.G., C.J.O., P.S.T., J.B.M., J.A.L., B.F.V. and K.-M.C. Geisinger-Regeneron DiscovEHR Collaboration/Regeneron Genetics Center: L.A.L., J.B.N., A.E.L., M.B.J., N.V. and A.B. EPoS Consortium: Q.M.A., R.D., H.J.C. and A.K.D.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01078-z.

Reporting summary.

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

Imputation was performed using MiniMac4 and EAGLE v2. Association analysis was performed using PLINK2a. Post-GWAS processing software includes LD Hub v1.9.3, METAL v2011-03-25, DEPICT v140721, LDSC v1.0, GREGOR v4.0, HiCUP v0.8, STRING v11 and Ensembl Variant Effect Predictor with assembly GRCh37. p13 as outlined in Methods. Clear code for analysis is available at the associated website of each software package. Additional analyses were performed in R-4.1, Bioconductor v3.140 and R packages corrcoverage, CHiCAGO and OmnipathR, for which code can be found in their associated vignettes.

Competing interests

H.R.K. is a scientific advisory board member for Dicerna Pharmaceuticals, Sophrosyne Pharmaceuticals and Enthion Pharmaceuticals; a consultant for Sobrera Pharmaceuticals; the recipient of research funding and medication supplies for an investigator-initiated study from Alkermes; a member of the American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative, which during the past 3 years was supported by Alkermes, Amygdala Neurosciences, Arbor Pharmaceuticals, Dicerna, Ethypharm, Indivior, Lundbeck, Mitsubishi and Otsuka; and is named as an inventor on the Patent Cooperation Treaty patent application #15/878,640 entitled 'Genotype-guided dosing of opioid agonists,' filed 24 January 2018. D.G. is employed part-time by Novo Nordisk. A.V.K. is an employee and holds equity in Verve Therapeutics; has served as a scientific advisor to Amgen, Third Rock Ventures, Illumina, and Foresite Labs; received a sponsored research agreement from IBM Research; and is listed as a co-inventor on a patent application for use of imaging data in assessing body fat distribution and associated cardiometabolic risk. S.J.A. is President of Sanyal Bio; has stock options in Genfit, Galmed, Exhalenz, Durect, Tiziana, Algernon and Indalo; has served as a consultant to Intercept, Gilead, Bristol Myers Squibb, Novartis, Pfizer, Lilly, Novo Nordisk, AstraZeneca, Medimmune, Merck, Allergan, Albireo, Boehringer Ingelhiem, Celgene, NGM, Glympse, Conatus, Genentech, Tern, Takeda, Hemoshear, Immuron, Surrozen, Poxel, Path AI, Second Genome, Zydus, Chiasma, Surrozen, Poxel, Blade, Pliant, Liposcience, Cymabay, Salix, Ferring and Teva; and his institution has received grants from Intercept, Gilead, Novartis, Merck, AstraZeneca, Malinckrodt, Pfizer, Lilly, Salix and Bristol Myers Squibb. V.C.U. has ownership interests in Sanyal Bio. K.R.R. is on the NASH Advisory Board at Novo Nordisk and receives grant support from TARGET-NASH, Bristol Myers Squibb and Intercept Pharmaceuticals. J.B.N., A.E.L., M.B.J., N.V., A.B., M.E.H. and L.A.L. receive salary, stocks and/or stock options from Regeneron Pharmaceuticals. R.P.M. and C.C. are employees and shareholders of Gilead Sciences. Q.M.A. is coordinator of the EU IMI-2 LITMUS consortium, which is funded by the EU Horizon 2020 program and the European Federation of Pharmaceutical Industries and Associations. Q.M.A. reports research grant funding from Allergan/Tobira, AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Glympse Bio, Intercept, Novartis Pharma and Pfizer; Consultancy for 89Bio, Abbvie/Allergan, Akero, Altimentiv, Altimmune, AstraZeneca, Axcella, Blade, BMS, BNN Cardio, Boehringer Ingelheim, Cirius, CymaBay, EcoR1, E3Bio, Eli Lilly & Company, Galmed, Genentech, Genfit, Gilead, Grunthal, HistoIndex, Indalo, Intercept Pharma, Inventiva, IQVIA, Janssen, Johnson & Johnson, Madrigal, MedImmune, Medpace, Merck, Metacrine, NGMBio, North Sea Therapeutics, Novartis, Novo Nordisk, PathAI, Pfizer, Poxel, ProSciento, Raptor Pharma, Roche, Servier, Shionogi, Terns, The Medicines Company and Viking Therapeutics; speaker fees from Abbott Laboratories, Allergan/Tobira, BMS, Clinical Care Options, Falk, Fishawack, Genfit, Gilead, Integritas Communications, Kenes and Medscape; and royalties from Elsevier. S.M.D. receives research support from RenalytixAI and personal consulting fees from Calico Labs outside the scope of the current research. S.L.D. reports grants from Alnylam Pharmaceuticals, Astellas Pharma, AstraZeneca Pharmaceuticals, Biodesix, Boehringer Ingelheim International, Celgene Corporation, Eli Lilly and Company, Genentech, Gilead Sciences, GlaxoSmithKline, Innocrin Pharmaceuticals, IQVIA, Janssen Pharmaceuticals, Kantar Health, MDxHealth, Merck & Co, Myriad Genetic Laboratories, Novartis International and Parexel International Corporation through the University of Utah or Western Institute for Veteran Research outside the submitted

A multiancestry genome-wide association study of unexplained chronic ALT elevation as a proxy for nonalcoholic fatty liver disease with histological and radiological validation

A full list of authors and affiliations appears at the end of the article.

Abstract

Nonalcoholic fatty liver disease (NAFLD) is a growing cause of chronic liver disease. Using a proxy NAFLD definition of chronic elevation of alanine aminotransferase (cALT) levels without other liver diseases, we performed a multiancestry genome-wide association study (GWAS) in the Million Veteran Program (MVP) including 90,408 cALT cases and 128,187 controls. Seventy-seven loci exceeded genome-wide significance, including 25 without prior NAFLD or alanine aminotransferase associations, with one additional locus identified in European American-only and two in African American-only analyses ($P < 5 \times 10^{-8}$). External replication in histology-defined NAFLD cohorts (7,397 cases and 56,785 controls) or radiologic imaging cohorts (n = 44,289) replicated 17 single-nucleotide polymorphisms (SNPs) ($P < 6.5 \times 10^{-4}$), of which 9 were new (*TRIB1*, *PPARG*, *MTTP*, *SERPINA1*, *FTO*, *IL1RN*, *COBLL1*, *APOH* and *IFI30*). Pleiotropy analysis showed that 61 of 77 multiancestry and all 17 replicated SNPs were jointly associated with metabolic and/or inflammatory traits, revealing a complex model of genetic architecture. Our approach integrating cALT, histology and imaging reveals new insights into genetic liability to NAFLD.

Chronic liver disease with progression to cirrhosis and hepatocellular carcinoma is a global health issue[1]. In particular, non-alcoholic fatty liver disease (NAFLD) is an increasingly common cause of chronic liver disease, with an estimated world prevalence of 25% among adults[2]. In the United States, NAFLD prevalence is projected to reach 33.5% among adults by 2030 (ref. [3]). NAFLD is defined by ≥5% fat accumulation in the liver in the absence of other known causes for liver disease, based on liver biopsy and/or radiologic imaging[4,5].

Individual susceptibility to NAFLD involves both genetic and environmental risk factors. Current estimates of NAFLD heritability range from 20% to 50%[6], and risk factors include obesity, insulin resistance and several features of metabolic syndrome[2]. A limited number of genetic variants that promote NAFLD have been identified in GWASs using liver biopsy, imaging and/or isolated liver enzyme values, such as *PNPLA3*, *TM6SF2* and

work. C.J.O. is an employee of Novartis Institute for Biomedical Research. S.F.A.G. is the Daniel B. Burke Endowed Chair for Diabetes Research. The remaining authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01078-z.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

MTARC1[7–11], which highlights the need for expanded discovery with larger samples and greater population diversity.

The MVP is among the world's largest and ancestrally diverse biobanks[12] and a promising resource for precision medicine. NAFLD is markedly clinically underdiagnosed due to the invasive nature of the liver biopsy procedure, variable use of imaging modalities and poor sensitivity of diagnostic codes[5]. The use of chronic elevation of alanine aminotransferase (cALT) as a proxy for NAFLD has good specificity and positive predictive value in the NAFLD diagnostic algorithm within the US Department of Veterans Affairs (VA)[13]. Accordingly, we recently adapted and validated this cALT phenotype to facilitate NAFLD case identification in MVP[14], applying a rigorous exclusion of other conditions that are known to increase liver enzymes. Our aims are to (1) perform a multiancestry genetic susceptibility analysis of cALT in MVP, (2) replicate the lead SNPs in external NAFLD cohorts defined by liver histology or radiologic imaging, (3) predict putative effector genes at the lead loci and (4) characterize the genetic architecture using cross-phenotypic associations.

Results

An ancestry-diverse study population enriched for metabolic conditions.

Our study consisted of 90,408 cALT cases and 128,187 controls comprising four ancestral groups: European American (EA, 75.1%), African American (AA, 17.1%), Hispanic American (HISP, 6.9%) and Asian American (ASN, 0.9%; Supplementary Table 1 and Supplementary Fig. 1). Consistent with the US veteran population, MVP cases and controls were predominantly male (92.3%) and enriched for metabolic conditions (Supplementary Table 1).

Multiancestry and ancestry-specific cALT-associated loci.

To identify cALT-associated loci, we performed a multiancestry GWAS meta-analysis (Methods and Fig. 1). Seventy-seven independent SNPs met conventional genome-wide significance ($P < 5 \times 10^{-8}$), of which 60 exceeded multiancestry genome-wide significance ($P < 5 \times 10^{-9}$). Fifty-two SNPs were previously reported to be associated with alanine aminotransferase (ALT), including nine that were also associated with NAFLD (i.e., *PNPLA3*, *TM6SF2*, *HSD17B13*, *PPP1R3B*, *MTARC1*, *ERLIN1*, *APOE*, *GPAM* and *SLC30A10-LYPLAL1*; Fig. 2 and Supplementary Table 2)[7–11,14–22]. Of the 25 remaining loci, 11 have been associated with gamma-glutamyl transferase and/or alkaline phosphatase levels before[22], and 14 were new.

In the ancestry-specific analyses, 55 loci in EAs, eight loci in AAs and three in HISPs were genome-wide significant ($P < 5 \times 10^{-8}$; Supplementary Tables 3–5 and Supplementary Figs. 2–5), of which one EA and two AA SNPs were not captured in the multiancestry analysis (Supplementary Fig. 5). Notably, the top two SNPs in the AA-only scan (near/in *GPT* and *ABCB4*) are polymorphic among AAs but nearly monomorphic in other populations.

Replication in liver biopsy and radiologic imaging data.

To assess whether our cALT-associated SNPs capture genetic susceptibility to NAFLD, we performed a multiancestry lookup of the 77 SNPs in two external NAFLD cohorts, namely (1) a liver biopsy cohort consisting of 7,397 histologically characterized NAFLD cases and 56,785 population controls (Methods and Supplementary Tables 6 and 7) and (2) a liver imaging cohort consisting of 44,289 participants with available radiologic liver imaging-based quantitative hepatic fat (qHF) measurements (Methods and Supplementary Table 8). In the liver biopsy cohort, there was directional concordance between effect estimates of NAFLD in 66 of 77 SNPs (86%), including 15 SNPs with a significant association (adjusted Bonferroni $P < 6.5 \times 10^{-4}$), of which eight have not been reported previously (e.g., *TRIB1*, *MTTP*, *APOH*, *IFI30*, *COBLL1*, *SERPINA1*, *IL1RN* and *FTO*; Supplementary Table 7)[9]. In the liver imaging cohort, there was directional concordance between effect estimates of cALT and qHF in 49 of 77 SNPs (64%). Among these, 11 were significantly associated (Bonferroni-adjusted $P < 6.5 \times 10^{-4}$; Supplementary Table 8), of which six were new (e.g., *TRIB1*, *MTTP*, *APOH*, *IFI30*, *COBLL1* and *PPARG*). The *PPP1R3B* locus was significantly associated with qHF, but in the opposite direction from cALT, and was not associated with biopsy-proven NAFLD. Collectively, 17 of 77 SNPs were replicated in external histologic and/or radiologic NAFLD cohorts, of which nine were previously unreported (Supplementary Table 2).

We performed a SNP-specific statistical power analysis to investigate the SNP-specific type I error (α) in the liver biopsy cohort (Bonferroni-adjusted $P < 6.5 \times 10^{-4}$; Supplementary Tables 2 and 7). Twenty-two loci showed sufficient statistical power (>80%) for replication, of which ten replicated (Supplementary Fig. 6). Twelve sufficiently powered SNPs did not replicate, which included *GPT*, *PPP1R3B* and *PANX1*. Remarkably, of the 55 SNPs without sufficient statistical power for replication, six SNPs did in fact replicate (namely *PPARG*, *MTTP*, *FTO*, *IL1RN*, *IFI30* and *COBLL1*). The effect size of cALT SNPs that replicated in the histological dataset was on average 92.4% higher than their respective cALT effect estimates (Supplementary Figs. 7 and 8). Of the remaining 49 SNPs without sufficient statistical power, 22 SNPs showed higher effect sizes for histological NAFLD than cALT and were labeled as candidate NAFLD loci (Supplementary Table 2).

GRSs and histologically characterized NAFLD.

We next constructed genetic risk scores (GRSs) in four independent liver biopsy cohorts to quantify the cumulative predictive power of our 77 sentinel variants (Supplementary Table 9 and Supplementary Fig. 9). A 77 candidate SNP-based GRS was predictive of NAFLD in the liver biopsy cohort (GRS-77, $P = 3.7 \times 10^{-28}$). A partitioned GRS consisting of nine established NAFLD SNPs (GRS-9) and a GRS consisting of 68 new SNPs (GRS-68) revealed that both scores independently predicted biopsy-defined NAFLD (GRS-9, $P = 2.2 \times 10^{-11}$; GRS-68, $P = 2.4 \times 10^{-7}$). A GRS that consisted of the 17 externally replicated SNPs strongly predicted NAFLD ($P = 2.53 \times 10^{-10}$) (Supplementary Fig. 10 and Supplementary Table 9).

Heritability and genetic correlations with other phenotypes.

To further characterize the genetic architecture of our cALT phenotype, we estimated heritability and genetic correlations with other traits using linkage disequilibrium (LD) score regression (Methods). The SNP-based liability-scaled heritability was estimated at 16% (95% confidence interval (CI), 12–19, $P < 1 \times 10^{-6}$) in EAs. Genetic correlation analysis between cALT and 774 complex traits from LD Hub (Methods) identified a total of 116 significant associations (Bonferroni-adjusted $P < 6.5 \times 10^{-5}$; Supplementary Table 10). These associations encompassed 78 cardiometabolic risk factors (67.2%), which is consistent with reports from observational studies correlating these traits to NAFLD[23]. Additional genetically correlated traits represented general health conditions (11.2%), educational attainment and/or socioeconomic status (12.0%) and other conditions (9.5%), such as gastroesophageal reflux, smoking and osteoarthritis.

Conditionally independent variants.

To discover additional conditionally independent cALT signals within the 77 genomic regions, we performed exact conditional analysis using stepwise regression on individual-level data for all single-ancestry sentinel variants. We detected a total of 29 conditionally independent SNPs ($P < 1 \times 10^{-5}$) (Supplementary Table 11). Specifically, the *GPT* locus showed the highest degree of genomic regional complexity with four conditionally independent SNPs, followed by *AKNA* with three conditionally independent SNPs. We observed a total of six conditionally independent variants at three genomic regions for AAs. For one new locus located on chromosome 12 between 121 and 122 Mb, the multiancestry lead variant (rs1626329) was located in *P2RX7*, whereas the lead variant for EA mapped to *HNF1A* (rs1169292; Supplementary Fig. 11). Both variants are linked to compelling candidate genes for metabolic liver disease.

Fine mapping to define potential causal variants.

To leverage the increased sample size and population diversity to improve fine-mapping resolution, we computed statistically derived 95% credible sets using Wakefield's approximate Bayes' factors[24] (Supplementary Tables 12–15 and Methods). Multiancestry fine mapping reduced the median 95% credible set size from 9 in EA (interquartile range, 3–17) to 7.5 variants (interquartile range, 2–13). A total of 11 distinct cALT associations were resolved to a single SNP in the multiancestry metaregression, with four additional loci suggesting single SNP sets from EA ($n = 2$) and AA ($n = 2$) ancestry-specific scans.

Liver-specific enrichment of cALT heritability.

To ascertain the tissues contributing to the disease-association underlying cALT heritability, we performed tissue-specific heritability analysis (Methods). The strongest associations were observed for genomic annotations surveyed in liver, hepatocytes, adipose and immune cell types, among others (Supplementary Table 16). Medical subject heading-based analysis showed enrichment mainly in hepatocytes and liver (Supplementary Table 17). Gene set analysis showed strongest associations for liver and lipid-related traits (Supplementary Table 18). Enrichment analyses using publicly available epigenomic data (Methods) showed that most significant enrichments were observed for active enhancer chromatin state in liver, epigenetic modification of histone H3 in hepatocytes or liver-derived HepG2 cells (Bonferroni-adjusted $P < 1.8 \times 10^{-5}$; Supplementary Tables 19 and 20). DEPICT-based predicted gene function nominated 28 gene candidates, including *PNPLA3*, *PPARG* and *ERLIN1* (false discovery rate (FDR) < 5%; Supplementary Table 21). These analyses support the hypothesis that our cALT GWAS captures multiple physiological mechanisms that contribute to NAFLD heritability.

Coding variants in putative causal genes.

Six new multiancestry loci are missense variants (Supplementary Table 22) in *CPS1*, *GPT*, *TRIM5*, *DNAJC22*, *SERPINA1* and *APOH*. To identify additional coding variants, we investigated predicted loss-of-function and missense variants in high LD to the identified cALT lead variants (ancestry-specific $r^2 > 0.7$; Supplementary Table 22). Four previously described missense variants were replicated in the current study (in *MTARC1*, *ERLIN1*, *TM6SF2* and *PNPLA3*). Among new loci, missense variants in high LD with lead variants included variants in *CCDC18, MERTK, APOL3, PPARG, MTTP, MLXIPL, ABCB4, GPAM, SH2B3, P2RX7, ANPEP, IFI30* and *MPV17L2*. An AA-specific locus (rs115038698) was in high LD to the nearby missense variant Ala934Thr in *ABCB4* (rs61730509, AFR $r^2 = 0.92$) with a predicted deleterious effect using sorting intolerant from tolerant (SIFT) and polymorphism phenotyping 2 (PolyPhen-2) tools. In summary, 24 loci prioritized a candidate gene based on a missense variant in high LD with the lead SNP.

Additional approaches to nominate putative causal genes.

Colocalization analyses.—To prioritize putative causal genes at the cALT-associated loci, we performed colocalization analyses with gene and splicing expression quantitative trait loci (QTLs) in 48 tissues, and histone QTL data from primary liver (hQTLs; Methods). A total of 123 genes were prioritized, including 20 genes expressed in liver tissue (Supplementary Table 23). In liver tissue, eight variant-gene pairs were identified where the direction of association between the effect allele was concordant between cALT risk and transcription levels. Splicing QTL analysis prioritized two genes in the liver (*HSD17B13* and *ANPEP*) (Supplementary Table 24). Finally, two lead SNPs were in high LD ($r^2 > 0.8$) with variants that regulated H3K27ac levels in liver tissue (hQTLs), namely *EFHD1* (rs2140773 and rs7604422) and *FADS2* (rs174566).

Assay for chromatin accessibility using liver-derived cells.—To decipher regulatory mechanisms involved in the pathophysiology of NAFLD, we mapped our cALT loci to regions of open chromatin using assay for transposase-accessible chromatin with high-throughput sequencing in three biologically relevant liver-derived tissues (human liver, liver cancer cell line (HepG2) and hepatocyte-like cells (HLCs) derived from pluripotent stem cells)[25]. Additionally, we used promoter-focused Capture-C data to identify those credible sets that physically interact with genes in two relevant cell types (HepG2 and liver). For each credible set, we identified genes with significant interactions (CHiCAGO score > 5; Methods) that overlap with at least one lead variant (Supplementary Table 25). Based on DEPICT gene prediction, coding variant linkage analysis and QTL colocalization (Supplementary Tables 18–25), we identified 215 potentially relevant genes for the 77 loci. A protein–protein interaction (PPI) analysis revealed that among the 192 available proteins, 86 nodes were observed, with strong PPI enrichment ($P < 9.0 \times 10^{-8}$), indicating that the protein network shows substantially more interactions than expected by chance (Supplementary Table 26 and Supplementary Fig. 12).

Cumulating evidence to nominate putative causal genes.

We present an ensemble method for predicting the likely causal effector gene at 77 loci based on eight distinct gene-mapping analyses (Methods). For each available gene within a sentinel locus, we counted the number of times it was identified in these eight analyses as a measure of cumulative evidence that the respective gene is the actual causal effector gene in the region. This ensemble method resulted in the nomination of a single gene as the causal effector gene at 53 of 77 genomic loci. At the remaining 24 loci, two loci lacked any data to support the nomination of a causal gene, and at 22 loci, two or more causal genes were nominated (Supplementary Table 27). We highlighted 35 loci for which a causal gene was prioritized by at least three sources of evidence (or four sources of evidence for coding variants) in Tables 1 and 2. To confirm that the nominated genes are involved in liver biology, we performed a gene expression lookup in single-cell RNA-sequencing data from the Liver Single Cell Atlas[26] and found that at 76 of 77 loci, a gene was nominated that was expressed in at least one liver cell type (Supplementary Table 27).

Transcription factor analysis.

We observed that 14 nominated genes are transcription factors (TFs) (Supplementary Table 28). Using the DoRothEA data in OmniPath, we identified that two of these TFs have several downstream target genes that were also identified in our GWAS scan (Methods). Notably, CEBPA targets the downstream genes *PPARG*, *TRIB1*, *GPAM*, *FTO*, *IRS1*, *CRIM1*, *HP*, *TBC1D8* and *CPS1*. Similarly, HNF1A, encoded by the lead gene in the EA scan, targets *SLC2A2*, *MTTP* and *APOH*.

Cross-phenotypic associations.

We next sought to identify additional traits that were associated with our 77 multiancestry lead SNPs using (1) a LabWAS of clinical laboratory test results[27] in MVP (Methods, Supplementary Table 29 and Supplementary Fig. 13), (2) PheWAS analyses in the UK Biobank using SAIGE (Methods and Supplementary Table 30), (3) candidate SNP lookups in the Medical Research Council (MRC) Integrative Epidemiology Unit (IEU) OpenGWAS project (Supplementary Table 32) and (4) ancestry-specific cross-trait colocalization analyses with 36 GWAS statistics of cardiometabolic and blood cell-related traits (Methods and Supplementary Table 33). Specifically, we examined all associations for PheCode 571.5, ‘Other chronic nonalcoholic liver disease’ in the UK Biobank, which comprised 1,664 cases and 400,055 controls, which with a disease prevalence of 0.4% seems to be under-reported. Still, of the 73 available lead variants, 14 were nominally associated and directionally consistent with our scan (signed binomial test $P = 3.4 \times 10^{-9}$) (Supplementary Table 31). Of the 17 replicated SNPs in liver biopsy and qHF, 14 were consistent with the UK Biobank NAFLD phenotype.

Based on the four analyses described above, we selected all SNP–trait associations with relevant phenotypes to NAFLD biology and classified them as liver (for example ALT, aspartate aminotransferase (AST)), metabolic (for example cholesterol, triglycerides, body mass index (BMI) and HbA1c) or inflammatory traits (e.g., C-reactive protein, white blood cell count) (Fig. 3 and Supplementary Tables 29–33). Seventeen multiancestry SNPs and one EA-specific SNP showed associations with liver traits only (Fig. 3). In contrast, 17 multiancestry and three ancestry-specific loci showed associations with both liver and metabolic traits, whereas four multiancestry loci showed associations with both liver and inflammatory traits. Finally, 39 multiancestry loci showed association with all three traits (liver enzymes, cardiometabolic traits and inflammation), including 15 of 17 externally replicated loci. Our findings confirm that NAFLD-associated loci are highly pleiotropic.

Pleiotropy-stratified GRS and histological NAFLD.

The above analyses raised the possibility that SNPs with greater pleiotropy relative to metabolic and/or inflammatory traits beyond liver-related traits may have greater contributions to NAFLD. To this end, we compared the GRS between four subgroups of multiancestry SNPs as defined in Fig. 3, including (1) 17 SNPs only associated with liver traits, (2) 5 SNPs associated with liver and inflammatory traits, (3) 17 SNPs associated with liver and metabolic traits and (4) 38 SNPs associated with liver, cardiometabolic and inflammatory traits. All four subgroups showed significant capacity to predict NAFLD (Supplementary Table 9). However, the strongest effect was observed for the GRSs in which SNPs were associated with all three traits ($P = 2.8 \times 10^{-9}$) (Supplementary Fig. 14). Collectively, these findings show highest discriminative accuracy for NAFLD of SNPs associated with liver, metabolic and inflammatory traits.

Directional pleiotropy and gene clusters.

Finally, we visualized the direction and strength of the associations between 77 multiancestry loci and 7 inflammatory biomarkers and 13 cardiometabolic traits in a heatmap (Fig. 4). The loci were grouped into seven gene clusters (Methods). Gene cluster 1 consisted of five multiancestry loci (including *APOE*) for which cALT risk alleles were associated with increased low-density lipoprotein (LDL) and total cholesterol, apolipoprotein B1 and markers of inflammation. Gene cluster 2 composed of genes (such as *IL1RN*, *MTARC1*, *GPAM* and *TRIB1*) for which the cALT risk alleles were associated with increased LDL, total cholesterol and apolipoprotein B1 but decreased levels of inflammatory markers. Gene cluster 3 (including *MTTP*) included genes that showed predominantly positive associations with apolipoprotein B1, LDL and total cholesterol. Gene cluster 4 was characterized by a lack of distinctive biomarker coassociation profiles. Genes in cluster 5 (including *PNPLA3*, *ERLIN1* and *PPP1R3B*) were characterized by higher rates of type 2 diabetes but decreased levels of triglycerides, LDL cholesterol, high-density lipoprotein (HDL) cholesterol, apolipoprotein A1 and B1 and white blood cell count. Genes in cluster 6 (e.g., *PPARG* and *SLC30A10*) were associated with higher triglycerides and type 2 diabetes but decreased sex hormone binding globulin, HDL cholesterol and apolipoprotein A1. Finally, genes in cluster 7 (including *TM6SF2* and *FTO*) were associated with increased inflammatory markers but lower apolipoprotein B1 and total and LDL cholesterol. Interestingly, for a total of nine SNPs (*TRIB1*, *PPARG*, *SLC30A10* (formerly *LYPLAL1*), *MLXIP*, *CEBPA*, *COBLL1*, *C6orf223*, *MIR5702* and *SH2B3*), the cALT risk allele was associated with lower BMI, consistent with a ‘lean NAFLD’ phenotype. Similarly, the cALT risk alleles of *SERPINA1* and *OSGIN1* loci seemed to be associated with lower rates of type 2 diabetes and *SH2B3* and *SLC2A2* with lower glucose and HbA1c. Overall, these directional associations define distinct gene cluster characteristics with potential biological implications.

Discussion

In this study, we describe a multiancestry GWAS of cALT as a proxy for NAFLD, which resulted in a total of 77 multiancestry loci, of which 25 have not been associated with NAFLD or ALT before. We additionally identified three ancestry-specific loci and 35 conditionally independent SNPs. We assembled two external replication cohorts with histologically confirmed NAFLD and hepatic fat defined by imaging and replicated the association of 17 SNPs with NAFLD, of which nine are new (*TRIB1*, *PPARG*, *MTTP*, *SERPINA1*, *FTO*, *IL1RN*, *COBLL1*, *APOH* and *IFI30*).

All 17 replicated SNPs showed significant associations with metabolic risk factors and/or inflammatory traits, and a GRS based on the subset of SNPs that are associated with liver, cardiometabolic and inflammatory markers showed the highest discriminative accuracy to predict histological NAFLD. Our directional pleiotropy analysis for metabolic risk factors are overall concordant with the results from Sliz et al.[28], which investigated four NAFLD SNPs (*LYPLAL1*, *PNPLA3*, *GCKR* and *TM6SF2*). Collectively, our findings offer a comprehensive, expanded and refined view of the genetic contribution to cALT with potential clinical, pathogenic and therapeutic relevance.

Our proxy NAFLD phenotype was based on chronic elevation of ALT levels with the exclusion of other known diagnoses of liver disease or causes of ALT elevation (e.g., viral hepatitis, alcoholic liver disease and hemochromatosis), based on previous validation within VA population[13,14]. In this regard, several GWASs of liver enzyme levels have been reported, particularly of serum ALT[8,17,21,22], but not all studies systematically excluded non-cardiometabolic causes of ALT elevation. Pazoki et al. recently reported 230 loci related to ALT, of which 52 were also included in our panel of 77 (67.5%) lead cALT loci[21,22]. We recognize that some cALT loci such as *GPT* may be involved more directly in ALT biology rather than NAFLD.

The MVP is one of the world's largest and most ancestrally diverse biobanks, and 25% of the participants are of non-white ancestry. Using data from multiple ancestries allowed us to narrow down putative causal variants for NAFLD through multiancestry fine mapping. Additionally, the affirmative external replication of 17 SNPs in two large biopsy- and imaging-based NAFLD cohorts supports the relevance of our proxy phenotype for NAFLD, including not only loci previously associated with NAFLD or all-cause cirrhosis but also several of the new loci reported here (e.g., *TRIB1*, *SERPINA1*, *MTTP*, *IL1RN*, *IFI30*, *COBLL1*, *APOH*, *FTO*, *PPP1R3B* and *PPARG*). For all loci except *PPP1R3B*, we observed concordant directionality of effects between cALT and hepatic fat. The apparent discrepancy in the *PPP1R3B* locus has been reported before[9] and may represent diffuse attenuation on radiologic images due to hepatic accumulation of glycogen[29] rather than triglycerides[30]. In addition, our study failed to replicate the *GCKR* locus, where a common missense variant (rs780094) has been repeatedly shown to confer susceptibility to NAFLD[11]. The SNP is a risk factor for increased triglycerides, C-reactive protein and LDL cholesterol but seems to be protective for T2D, fasting glucose, alcohol intake, alcohol use disorder, BMI and monocyte percentage. It is hypothesized that the variant GCKR protein loses interaction efficiency with glucokinase, which promotes hepatic glucose metabolism, decreases plasma glucose levels and increases NAFLD risk[31]. Our phenotype might not be a suitable proxy for NAFLD for SNPs that act through multiple pathways with opposing effects on ALT.

A substantial fraction of our cALT loci showed a shared genetic coarchitecture with metabolic traits (Fig. 4). Of interest is that for nine SNPs, the cALT risk allele was associated with lower BMI, including *PPARG*. These SNPs seem to exhibit mild lipodystrophic effects, characterized by reduced adipose tissue and increased hepatic steatosis. Further study is required to clarify whether and which loci are working primarily in adipose tissue with a secondary effect on liver steatosis. Several genes and liver-enriched TFs involved in LDL and triglyceride pathways have been identified, such as *TRIB1*, *FTO*, *COBLL1*, *MTTP*, *TM6SF2*, *PPARG*, *APOE* and *GPAM*[32–35]. TRIB1 presumably regulates very low density lipoprotein (VLDL) secretion by promoting the degradation carbohydrate-response element binding protein (ChREBP, encoded by *MLXIPL*), reducing hepatic lipogenesis and limiting triglyceride availability for apolipoprotein B lipidation. Furthermore, TRIB1 coactivates the transcription of MTTP, a microsomal triglyceride transfer protein that loads lipids onto assembling VLDL particles to facilitate their secretion. Lomitapide, a small-molecule inhibitor of MTTP, is an LDL cholesterol-lowering treatment in homozygous familial hypercholesterolemia[36]. TRIB1 is also involved in the degradation of the key hepatocyte TF CEBPA[37], which together with HNF1A, RORα and MIR-122 is involved in a feedback loop of the liver-enriched TF network to control hepatocyte differentiation[38]. RORα is also a suppressor of transcriptional activity of peroxisome proliferator-activated receptor γ (PPARγ)[39]. PPARγ facilitates the hepatic uptake of triglyceride-rich lipoproteins via interaction with apolipoprotein E[40]. Large randomized controlled clinical trials have reported that the PPARγ agonists rosiglitazone and pioglitazone improve NAFLD-related hepatic steatosis, inflammation and fibrosis[41–44]. However, treatment is frequently accompanied with weight gain and fluid retention, limiting its application. RORα competes with PPARγ for binding to PPARγ target promoters, and therapeutic strategies designed to modulate RORα activity in conjunction with PPARγ may be beneficial for the treatment of NAFLD.

More than half of our cALT loci were associated with inflammatory traits (Fig. 4), consistent with the multiple-hit hypothesis of NAFLD[45]. For example, the TF MafB regulates macrophage differentiation[46], and genetic variation in *MAFB* has been associated with hyperlipidemia and hypercholesterolemia[47]. *FADS1* and *FADS2* are markedly induced during monocyte to macrophage differentiation, and it is hypothesized that they impact metabolic disease by balancing inflammatory and lipid mediators[48]. Another interesting locus is *IL1RN*. *IL1RN* encodes the anti-inflammatory cytokine interleukin-1 receptor antagonist (IL-1Ra) and is a natural inhibitor of IL-1 activity. It remains to be investigated whether remodeling of the adipose tissue inflammasome via IL-1 signaling blockade in obesity-associated NAFLD offers potential therapeutic benefit[49]. We note during proofing that a recent report described NAFLD associations using an imputed phenotype[50], some of which overlap with loci we report here (Supplementary Table 2).

In conclusion, we discovered 77 genomic loci associated with cALT in a large, ancestrally diverse cohort. We replicated our findings in external cohorts with hepatic fat defined by liver biopsy or radiologic imaging. The genetic architecture of the lead loci indicate a predominant involvement of metabolic and inflammatory pathways. This study constitutes a much-needed large-scale, multiancestry genetic resource that can be used to build genetic prediction models, identify causal mechanisms and understand biological pathways contributing to NAFLD initiation and disease progression.

Methods

Discovery cohort in the MVP.

The MVP is a megabiobank that was launched in 2011 and supported entirely by the Veterans Health Administration Office of Research and Development in the United States. The MVP received ethical and study protocol approval from the VA Central Institutional Review Board (IRB) in accordance with the principles outlined in the Declaration of Helsinki. The specific design, initial demographics and quality-control procedures of the MVP have been detailed previously[12].

Proxy NAFLD phenotype.

MVP NAFLD phenotype definitions were adapted from a previously published VA corporate data warehouse-derived approach using noninvasive clinical parameters[13,14]. The primary cALT phenotype was defined by: (1) elevated ALT > 40 U liter^{-1} for men or >30 U liter^{-1} for women during at least two time points at least 6 months apart within a 2-year window at any point prior to enrollment and (2) exclusion of other causes of liver disease, chronic liver diseases or systemic conditions and/or alcohol use disorders. The control group was defined by having a normal ALT (≤30 U liter^{-1} for men, ≤20 U liter^{-1} for women) and no apparent causes of liver disease or alcohol use disorder or related conditions[14]. Habitual alcohol consumption was assessed with the age-adjusted Alcohol Use Disorders Identification Test score[51,52]. Demographics of the proxy NAFLD cohort are shown in Supplementary Table 1. The prevalence of cirrhosis and advanced fibrosis was based on ICD-9 codes (456.2, 456.21, 571.5, 572.2 and 572.3) and ICD-10 codes (K72.9, K72.91, K74.0, K74.02, K74.1, K74.2, K74.6 and K74.69).

Single-variant autosomal analyses.

We tested imputed SNPs that passed quality control (i.e., Hardy-Weinberg equilibrium P-value > 1×10^{-10}, INFO imputation accuracy score > 0.3, and genotyping call rate > 0.975) for association with proxy NAFLD through logistic regression assuming an additive model of variants with minor allele frequency > 1% in EAs, AAs, HISPs and ASNs using PLINK2a software[53]. The regression coefficients from these analyses are the effect estimates and represent the log-odds change in the outcome for each unit of increase in effect alleles while holding other independents (e.g., covariates) in the model constant. Covariates included age, gender, age-adjusted Alcohol Use Disorders Identification Test score and first ten principal components (PCs) of genetic ancestry. Indels were excluded from analysis. We aggregated association summary statistics from the ancestry-specific analyses and performed a multiancestry meta-analysis. The association summary statistics for each analysis were meta-analyzed in a fixed-effects model using METAL with inverse-variance weighting of log odds ratios[54]. Variants were clumped using a range of 500 kb and/or LD $r^2 > 0.05$ in people of North European ancestry and were considered genome-wide significant if they passed the conventional P-value threshold of 5×10^{-8}. Multiancestry and ancestry-specific summary statistics are displayed in Supplementary Tables 2–5, and their corresponding genome-wide summary statistics are available through dbGAP accession code phs001672.v7.p1.

Secondary signal analysis.

The PLINK–condition and–condition-list parameters were used to conduct stepwise conditional analyses on individual-level data in MVP to detect ancestry-specific distinct association signals nearby lead SNPs. Regional SNPs were eligible if they were located within 500 kb of lead SNP with a minor allele frequency > 1%. Logistic regression was performed in a stepwise fashion, starting with a regional association analysis with the following set of covariates: lead SNP imputed allele dosage, age, gender, and 10 PCs of genetic ancestry. If the corresponding output file contained SNPs that reached locus-wide significance ($P < 1.0 \times 10^{-5}$), then the most significant SNP was selected and added to the covariate set. The regression was repeated until no locus-wide significant SNPs remained. The effect estimates (regression coefficients) for the secondary signals from logistic regression are shown in Supplementary Table 11.

Credible sets.

We calculated Wakefield's approximate Bayes' factors[24] based on the marginal summary statistics of the multiancestry and ancestry-specific summary statistics using the CRAN R package corrcoverage[55]. For each locus, the posterior probabilities of each variant being causal were calculated, and a 95% credible set was generated that contains the minimum set of variants that jointly have at least 95% posterior probability of including the causal variant (Supplementary Tables 12–15).

External replication in a liver imaging cohort.

A replication lookup of lead loci was performed to evaluate the extent to which genetic predictors of hepatocellular injury (cALT) correspond with qHF derived from computed tomography/magnetic resonance imaging-measured hepatic fat in the Penn Medicine Biobank[56], UK Biobank[57], Multi-Ethnic Study of Atherosclerosis, Framingham Heart Study[9] and University of Maryland Old Order Amish study (Supplementary Table 8). A detailed description is available in the Supplementary Note. All participating cohorts have ethical approval from their local institutions, and all relevant ethical regulations were followed. Liver fat was measured as attenuation in Hounsfield units in all computed tomography studies and as proton density fat fraction in the UK Biobank magnetic resonance imaging study. All cohorts underwent individual-level linear regression analysis on hepatic fat, adjusted for the covariates of age, gender, first ten PCs of genetic ancestry and alcohol intake if available. If the lead SNP was not available in any of the studies, then a proxy SNP in high LD with the lead variant was used ($r^2 > 0.7$); if no such variant was identified, then the SNP was set to missing for that respective study. The study-specific ancestry-stratified effect estimates were first standardized to generate standard scores or normal deviates (*z*-scores) and then meta-analyzed using METAL in a fixed-effects model with inverse-variance weighting of regression coefficients[54]. In a first round of meta-analysis, ancestry-specific effect estimates were generated, which then served as input for a subsequent round of meta-analysis that represents the multiancestry effects of our lead SNPs on qHF.

External replication in a liver biopsy cohort.

Available data from the following groups contributed to the liver biopsy cohort: (1) Non-Alcoholic Steatohepatitis Clinical Research Network (NASH-CRN) studies[20,41,58–62], (2) EPoS Consortium[11,63], (3) Geisinger Health System bariatric surgery cohort[64,65], (4) STELLAR-3 and ATLAS studies[66], (5) BioVU Biorepository[67] and (6) Penn Medicine Biobank. Results from liver biopsy data are shown in Supplementary Tables 6 and 7, and a detailed description of each of the individual studies is available in the Supplementary Note. All cohorts underwent individual-level logistic regression analysis on histologically defined NAFLD, adjusted for the covariates of age, gender, first ten PCs of genetic ancestry and alcohol intake if available. The study-specific effect estimates were meta-analyzed using METAL in a fixed-effects model with inverse-variance weighting of regression coefficients[54].

Heritability estimates and genetic correlations analysis.

LD score regression was used to estimate the heritability coefficient, and population and sample prevalence estimates were subsequently applied to estimate heritability on the liability scale[68]. A genome-wide genetic correlation analysis was performed to investigate possible coregulation or a shared genetic basis between cALT and other complex traits and diseases (Supplementary Table 9). Pairwise genetic correlation coefficients were estimated between the meta-analyzed cALT GWAS summary output in EA and each of 774 precomputed and publicly available GWAS summary statistics for complex traits and diseases by using LD score regression through LD Hub v1.9.3 (http://ldsc.broadinstitute.org). Statistical significance was set to a Bonferroni-corrected level of $P < 6.5 \times 10^{-5}$.

Tissue- and epigenetic-specific enrichment of cALT heritability.

We analyzed cell type-specific annotations to identify enrichments of cALT heritability as shown in Supplementary Table 16. First, a baseline gene model was generated consisting of 53 functional categories, including ENCODE functional annotations[69], Roadmap epigenomic annotations[70] and FANTOM5 enhancers[71]. Gene expression and chromatin data were also analyzed to identify disease-relevant tissues, cell types and tissue-specific epigenetic annotations. We used LDSC[68,72,73] to test for enriched heritability in regions surrounding genes with the highest tissue-specific expression. Sources of data that were analyzed included human tissue or cell type RNA-sequencing data from GTEx[74]; human, mouse or rat tissue or cell type array data from the Franke lab[75]; mouse brain cell type array data from Cahoy et al.[76]; mouse immune cell type array data from ImmGen[77]; and human epigenetic annotations from the Roadmap Epigenomics Consortium[70]. Expression profiles were considered statistically significantly enriched for cALT susceptibility if they passed the nominal *P*-value threshold of 0.003.

Pathway annotation enrichment.

Enrichment analyses in DEPICT[78] were conducted using genome-wide significant ($P < 5 \times 10^{-8}$) cALT GWAS lead SNPs (Supplementary Table 18) and considered an FDR threshold of 0.05 as significant. Tissue and gene set enrichment features are considered. We tested for epigenomic enrichment of genetic variants using GREGOR software (Supplementary Table 19)[79]. We selected EA-specific cALT lead variants with $P < 5 \times 10^{-8}$. We tested for enrichment of the resulting GWAS lead variants or their LD proxies (r^2 threshold of 0.8 within 1 Mb of the lead SNP) in genomic features including ENCODE, Epigenome Roadmap and manually curated data (Supplementary Table 20). Enrichment was considered significant if the enrichment *P*-value was less than the Bonferroni-corrected threshold of $P = 1.8 \times 10^{-5}$ (0.05/2,725 tested features).

Coding variant mapping.

All imputed variants in MVP were evaluated with Ensembl variant effect predictor[80], and all predicted loss-of-function and missense variants were extracted. The LD was calculated with established variants for multiancestry, EA, AA and HISP lead SNPs based on 1000 Genomes reference panel[81]. For SNPs with low allele frequencies, the MVP dataset was used for LD calculation for the respective underlying population. For the multiancestry coding variants, the EA panel was used for LD calculation. Coding variants that were in strong LD ($r^2 > 0.7$) with lead SNPs and had a strong statistical association ($P < 1 \times 10^{-5}$) were considered the putative causal drivers of the observed association at the respective locus (Supplementary Table 22).

Colocalization with gene expression.

Colocalization analysis was run separately for eQTLs and sQTLs for each of the 49 tissues in GTEx v8 (Supplementary Tables 23 and 24) (ref. [82]). For each tissue, we obtained an LD block for the genome with a sentinel SNP at $P < 5 \times 10^{-8}$, and then we restricted analysis to the LD blocks. For each LD block with a sentinel SNP, all genes within 1 Mb of the sentinel SNP (*cis*-Genes) were identified and then restricted to those that were identified as eGenes in GTEx v8 at an FDR threshold of 0.05 (*cis*-eGenes). For each *cis*-eGene, we performed colocalization using all variants within 1 Mb of the gene using the default prior probabilities in the ‘coloc’ function for the coloc package in R. We first assessed each coloc result for whether there was sufficient power to test for colocalization (PP3 + PP4 > 0.8), and for the colocalization pairs that passed the power threshold, we defined the significant colocalization threshold as PP4/(PP3 + PP4) > 0.9.

Overlap with open chromatin.

At each of the 77 NAFLD-associated loci from the multiancestry meta-analysis, we looked for overlaps between any variant in the credible set and regions of open chromatin previously identified using assay for transposase-accessible chromatin with high-throughput sequencing experiments in two cell types (three biological replicates of HepG2 (ref. [83]) and three biological replicates of HLCs[84] produced by differentiating three biological replicates of induced pluripotent stem cells), which in turn were generated from peripheral blood mononuclear cells using a previously published protocol[85]. Results are shown in Supplementary Table 25.

Overlap with promoter Capture-C data.

We used two promoter Capture-C datasets from two cell/tissue types to capture physical interactions between gene promoters and their regulatory elements and genes: three biological replicates of HepG2 liver carcinoma cells and HLCs[83]. The detailed protocol has been previously described[85]. Briefly, for each dataset, 10 million cells were used for promoter Capture-C library generation. Custom capture baits were designed using an Agilent SureSelect library design targeting both ends of DpnII restriction fragments encompassing promoters of all human coding genes, noncoding RNA, antisense RNA, small nuclear RNA, microRNA, small nucleolar RNA and long intergenic non-coding RNA transcripts, totaling 36,691 RNA baited fragments. Each library was then sequenced on an Illumina NovoSeq (HLCs) or Illumina HiSeq 4000 (HLCs), generating 1.6 billion read pairs per sample (50-bp read length). HiCUP[86] was used to process the raw FastQ files into loop calls; we then used CHiCAGO[87] to define significant looping interactions; a default score of 5 was defined as significant. We identified those NAFLD loci at which at least one variant in the credible set interacted with an annotated bait in the Capture-C data (Supplementary Table 25).

PPI network analysis.

We used the search tool for retrieval of interacting genes (STRING) v11 (ref. [88]) to seek potential interactions between nominated genes. STRING integrates both known and predicted PPIs and can be applied to predict functional interactions of proteins. In our study, the sources for interaction were restricted to the *Homo sapiens* species and limited to experimentally validated and curated databases. An interaction score >0.4 was applied to construct the PPI networks, in which the nodes correspond to the proteins and the edges represent the interactions (Fig. 4 and Supplementary Table 26).

Ensemble variant-to-gene mapping to identify putative causal genes.

Based on DEPICT gene prediction, coding variant linkage analysis, QTL analysis, annotation enrichment and PPI networks (Supplementary Tables 18–26), a total of 215 potentially relevant genes for NAFLD were mapped to multiancestry 77 loci. For each locus, we counted how many times each gene in that region was identified in the eight analyses. We then divided this number by the total number of experiments (i.e., eight) to calculate an evidence burden (called nomination score) that ranges from 0% to 100%. For each genomic locus, the gene that was most frequently identified as a causal gene was selected as the putative causal gene for that locus. In the case of a tie break, and if the respective genes had identical nomination profiles, the gene with eQTLs in multiple tissues was selected as the putative causal gene. Similarly, gene nomination was preferred for loci that strongly tagged ($r^2 > 0.8$) a coding variant. Loci that scored with three distinct sources of evidence or greater are listed for the coding variant (Table 1) and noncoding variants (Table 2), respectively.

MVP LabWAS.

A total of 21 continuous traits in the discovery MVP dataset (e.g., AST, alkaline phosphatase, fasting triglycerides, HDL, LDL, total cholesterol, random glucose, HbA1c, albumin, bilirubin, platelet count, BMI, blood urea nitrogen, creatinine, eGFR, systolic blood pressure, diastolic blood pressure, erythrocyte sedimentation rate, international normalized ratio and C-reactive protein) were tested in 186,681 EAs with association of 77 SNPs using linear regression of log-linear values. Covariates included age, gender and the first 10 PCs of EA ancestry (Supplementary Table 29). The Bonferroni *P*-value threshold was set at 3.09×10^{-5} (0.05/21 traits × 77 SNPs).

PheWAS with UK Biobank data.

For the 77 lead multiancestry SNPs and EA- and AA-specific SNPs, we performed a PheWAS in a GWAS of EHR-derived ICD billing codes from the white British participants of the UK Biobank using PheWeb[89]. In short, phenotypes were classified into 1,403 PheWAS codes excluding SNP–PheWAS code association pairs with case counts less than 50 (ref. [90]). All individuals were imputed using the Haplotype Reference Consortium panel[91], resulting in the availability of 28 million genetic variants for a total of 408,961 individuals. Effect estimates (e.g., regression coefficients) on binary outcomes were conducted using a model named SAIGE, adjusted for genetic relatedness, gender, year of birth and the first four PCs of white British genetic ancestry[92]. Results are shown in Supplementary Tables 30 and 31. SNP–trait associations are listed if they passed a nominal significance threshold of $P < 0.001$ and are considered Bonferroni significant when $P < 4.6 \times 10^{-7}$ (0.05/77 SNPs × 1,403 traits).

IEU OpenGWAS project SNP lookup.

An additional phenome-wide lookup was performed for 77 lead multiancestry SNPs and EA- and AA-specific SNPs in Bristol University's MRC IEU GWAS database[93]. This database consists of 126,114,500,026 genetic associations from 34,494 GWAS summary datasets, including UK Biobank (http://www.nealelab.is/uk-biobank), FinnGen (https://github.com/FINNGEN/pheweb), Biobank Japan (http://jenger.riken.jp/result), the NHGRI-EBI GWAS catalog (https://www.ebi.ac.uk/gwas), a large-scale blood metabolites GWAS[94], circulating metabolites GWAS[95], the MR-Base manually curated database[96] and a protein-level GWAS[97]. Results are shown in Supplementary Table 32.

Regional cardiometabolic cross-trait colocalization.

Bayesian colocalization tests between cALT-associated signals and the following trait- and disease-associated signals were performed using the COLOC R package[98]. To enable cross-trait associations, we compiled summary statistics of 36 cardiometabolic and blood cell-related quantitative traits and disease from GWASs conducted in individuals of EA ancestry and for MVP-based reports on individuals of AA or HISP ancestry. To summarize, for total, HDL and LDL cholesterol; triglycerides; alcohol use disorder and alcohol intake; systolic blood pressure and diastolic blood pressure; type 2 diabetes; BMI; and coronary artery disease, we used the summary statistics available from various MVP-based studies[47,51,99]. Of these, the summary statistics for coronary artery disease and BMI GWAS in MVP have not been published or deposited as of yet. Data on WHR were derived from GIANT Consortium[100], whereas summary statistics on CKD, gout, blood urea nitrogen, urate, urinary albumin-to-creatinine ratio, microalbuminuria, and eGFR were derived from CKD Genetics Consortium[101–103]. Finally, summary statistics of blood cell traits (for example platelet count, albumin, white blood cells, basophils, eosinophils, neutrophils, hemoglobin, hematocrit, immature reticulocyte fraction, lymphocytes, monocytes, reticulocytes, mean corpuscular hemoglobin, mean corpuscular volume, mean platelet volume, platelet distribution width, and red cell distribution width) were derived from a large-scale GWAS report performed in UK Biobank and INTERVAL studies[104]. A colocalization test was performed for all 77 cALT loci spanning 500-kb region around the lead SNP for all 36 compiled traits. For each association pair, COLOC was run with default parameters and priors to obtain posterior probabilities (PPs). Evidence of colocalization[105] was defined by PP3 + PP4 ≥ 0.99 and PP4/PP3 ≥ 5. Results are shown in Supplementary Table 33.

GRSs and histologically characterized NAFLD.

We constructed GRSs in four histologically characterized cohorts (e.g., Lundquist whites and HISPs, EPoS Consortium whites and BioVU whites) by calculating a linear combination of weights derived from the MVP dataset of lead 77 multiancestry cALT variants that passed conventional genome-wide significance (GRS-77, $P < 5.0 \times 10^{-8}$). The GRS-77 was standardized and the risk of histologically characterized NAFLD was assessed using a logistic regression model together with the potential confounding factors of age, gender and the first three to five PCs of ancestry. The regression coefficient (e.g., effect estimate) for GRS-77 represents the log odds change in NAFLD for each weighted unit of increase in effect alleles while holding the other independents in the model constant. To delineate the potential driving effects of known NAFLD loci, we divided the 77 loci into two sets and generated one PRS consisting of nine known NAFLD SNPs only (GRS-9), and one of newly identified 68 cALT SNPs (GRS-68). Both GRSs were added as independent predictors in a logistic regression model to explain histologically characterized NAFLD with the confounders of age, gender and PCs of ancestry. The individual effect sizes for each study were then meta-analyzed using the metagen package in R with random effects model comparing the standardized mean difference (SMD, mean differences divided by their respective standard deviations) (Supplementary Table 9). A forest plot was created to visualize the effect estimates between the studies (Supplementary Fig. 10). In similar fashion, SNPs were divided into three groups according to replication power, where SNPs were divided into a Bonferroni-replicated GRS consisting of 17 SNPs, a nominally significant with directional concordance GRS with 25 SNPs and nonreplicated GRSs with 35 SNPs (Supplementary Table 9 and Supplementary Fig. 11). Finally, a GRS subset was created based on the pleiotropy analysis and Venn diagram, where we generated a subset GRS that reflects liver + metabolic (17 SNPs), liver + metabolic + inflammation (38 SNPs), liver + inflammation (5 SNPs) and liver-only strata (17 SNPs) (Supplementary Table 9 and Supplementary Fig. 14).

TF analysis.

We identified nominated genes (Supplementary Table 28) that encode for TFs based on known motifs, inferred motifs from similar proteins and likely sequence-specific TFs according to literature or domain structure[106]. Target genes for these TFs were extracted using DoRothEA database[107] in OmniPath collection[108] using the associated Bioconductor R package OmnipathR[109], a gene set resource containing TF–TF target interactions curated from public literature resources (such as chromatin immunoprecipitation with sequencing peaks, TF binding site motifs and interactions inferred directly from gene expression.

Directional pleiotropy and gene cluster analysis.

We used the R package ‘pheatmap’ for a stratified agglomerative hierarchical clustering method named ‘complete linkage’, where each element is its own cluster at the beginning, and two clusters of the shortest distance in between them are sequentially combined into larger clusters until all elements are included in one single cluster, where distance is measured in Euclidean distance. We used the 77 lead SNPs and their corresponding single-trait effect estimates for 20 traits corresponding to three biological super groups (e.g., lipids, inflammation and metabolic) as input, with the sign of each cell determined by direction of effect and the strength by the $-\log^{10}(P$ value). The alleles were oriented as such that the cALT-increasing allele was set to the effect allele, which allows for direct comparison of the various association profiles. We selected the default ‘complete’ method and ‘Euclidean’ distance options to perform hierarchical clustering, stratified by the three super groups of metabolic, inflammation and lipid traits. The results of the clustering gene set are visualized with a dendrogram on the left side of the heatmap, which is broadly grouped into seven distinct gene clusters.

Ethics oversight.

All participating studies were conducted in compliance with the Declaration of Helsinki and comply with all relevant ethical and local regulatory requirements. Specifically, the contributing genetic association studies were approved by the Department of Veteran’s Affairs central IRB (VA MVP), the Vanderbilt University Medical Center IRB (BioVU), the IRB of Perelman School of Medicine at the University of Pennsylvania (Penn Medicine Biobank), the North West Multi-centre Research Ethics Committee (UK Biobank) and the IRB at the Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center (Long QT Screening study). The IRB at the University of Maryland School of Medicine approved the Old World Order Amish study, and the IRBs of six field centers (Wake Forest University School of Medicine, University of Minnesota, Northwestern University, Columbia University, Johns Hopkins University and University of California, Los Angeles) have approved the study protocol of the Multi-Ethnic Study of Atherosclerosis. The Framingham Heart Study is approved by the IRB of the Boston University Medical Center. The EPoS Consortium obtained ethical approval from all participating field centers (North East - Tyne & Wear South Research Ethics Committee, Inselspital Direktion Lehre und Forschung Bestätigung, Universiteit Antwerpen ethisch comité, Les Comités de Protection des Personnes Ile-de-France VI, Ethics committee at University Hospital in Linköping, Landesärztekammer Rheinland-Pfalz Ethik Kommission, Comitato Etico Interaziendale, Comitato Etico Palermo and IRB of the Fondazione Ca'Granda IRCCS of Milan). The STELLAR-3 and STELLAR-4 clinical trials were approved by IRBs or independent ethics committees at all participating sites in the United States, Japan, Canada, France, South Korea, Australia, Hong Kong, Spain, Taiwan, United Kingdom, India, Germany, Singapore, Brazil, Israel, Belgium, Mexico, Italy, Argentina, Austria, Poland, Puerto Rico, Switzerland, Malaysia, Portugal, Netherlands and New Zealand. The NASH Boys study, FLINT trial and PIVENS trial were approved by local IRBs of each clinical center participating in the NASH-CRN (Case Western Reserve University; Duke University Medical Center; Indiana University School of Medicine; Saint Louis University; University of California, San Diego; University of California, San Francisco; Virginia Commonwealth University; and Virginia Mason Medical Center) and a central data safety and monitoring board appointed by the National Institute of Diabetes and Digestive and Kidney Diseases. The NASH Women study was reviewed and approved by the NASH-CRN Steering Committee and the IRB at the Cedars-Sinai Medical Center Los Angeles. The Cholesterol and Pharmacogenetics trial obtained approval from the Children's Hospital and Research Center IRB, University of California, San Francisco Committee on Human Research and University of California, Los Angeles Office of the Human Research Protection Program. The Geisinger Health System bariatric surgery biobank Regeneron collaboration study was approved by the Geisinger Health System IRB. In all pediatric studies (NASH boys and Long QT Screening study), all parents provided informed consent, and children older than 7 years provided assent for participation in the respective study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Marijana Vujkovic[1,2,69], Shweta Ramdas[3,69], Kim M. Lorenz[1,3,4], Xiuqing Guo[5], Rebecca Darlay[6], Heather J. Cordell[6], Jing He[7], Yevgeniy Gindin[8], Chuhan Chung[8], Robert P. Myers[8,9], Carolin V. Schneider[3], Joseph Park[2,3], Kyung Min Lee[10], Marina Serper[1,2], Rotonya M. Carr[11], David E. Kaplan[1,2], Mary E. Haas[12], Matthew T. MacLean[3], Walter R. Witschey[13], Xiang Zhu[14,15,16,17], Catherine Tcheandjieu[14,18], Rachel L. Kember[19,20], Henry R. Kranzler[19,20], Anurag Verma[1,3], Ayush Giri[21], Derek M. Klarin[14,22,23], Yan V. Sun[24,25], Jie Huang[26], Jennifer E. Huffman[27], Kate Townsend Creasy[3], Nicholas J. Hand[3], Ching-Ti Liu[28], Michelle T. Long[29], Jie Yao[5], Matthew Budoff[30], Jingyi Tan[5], Xiaohui Li[5], Henry J. Lin[5], Yii-Der Ida Chen[5], Kent D. Taylor[5], Ruey-Kang Chang[5], Ronald M. Krauss[31], Silvia Vilarinho[32], Joseph Brancale[32], Jonas B. Nielsen[33], Adam E. Locke[33], Marcus B. Jones[33], Niek Verweij[33], Aris Baras[33], K. Rajender Reddy[2], Brent A. Neuschwander-Tetri[34], Jeffrey B. Schwimmer[35], Arun J. Sanyal[36], Naga Chalasani[37], Kathleen A. Ryan[38], Braxton D. Mitchell[38], Dipender Gill[39], Andrew D. Wells[40,41], Elisabetta Manduchi[3], Yedidya Saiman[42], Nadim Mahmud[43], Donald R. Miller[44,45], Peter D. Reaven[46,47], Lawrence S. Phillips[24,48], Sumitra Muralidhar[49], Scott L. DuVall[10,50], Jennifer S. Lee[14,18], Themistocles L. Assimes[14,18,51], Saiju Pyarajan[27,52,53], Kelly Cho[27,52,53], Todd L. Edwards[54,55], Scott M. Damrauer[1,3,56], Peter W. Wilson[24,57], J. Michael

Gaziano[27,52], Christopher J. O'Donnell[27,52,53], Amit V. Khera[23,53,58], Struan F. A. Grant[3,59,60], Christopher D. Brown[3], Philip S. Tsao[14,18,51], Danish Saleheen[61,62,63], Luca A. Lotta[33], Lisa Bastarache[7], Quentin M. Anstee[64,65], Ann K. Daly[65], James B. Meigs[23,53,66], Jerome I. Rotter[5], Julie A. Lynch[10,50,67],

Regeneron Genetics Center*,

Geisinger-Regeneron DiscovEHR Collaboration*,

EPoS Consortium*,

VA Million Veteran Program*,

Daniel J. Rader[2,3,69], Benjamin F. Voight[1,3,4,68,69,✉], Kyong-Mi Chang[1,2,69,✉]

Affiliations

[1]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA.

[2]Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[3]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[4]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[5]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA.

[6]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK.

[7]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA.

[8]Gilead Sciences, Inc., Foster City, CA, USA.

[9]The Liver Company, Palo Alto, CA, USA.

[10]VA Salt Lake City Health Care System, Salt Lake City, UT, USA.

[11]Division of Gastroenterology, University of Washington, Seattle, WA, USA.

[12]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[13]Department of Radiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[14]VA Palo Alto Health Care System, Palo Alto, CA, USA.

[15]Department of Statistics, The Pennsylvania State University, University Park, PA, USA.

[16]Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA.

[17]Department of Statistics, Stanford University, Stanford, CA, USA.

[18]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA.

[19]Mental Illness Research Education and Clinical Center, Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA.

[20]Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[21]Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN, USA.

[22]Division of Vascular Surgery, Stanford University School of Medicine, Palo Alto, CA, USA.

[23]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[24]Atlanta VA Medical Center, Decatur, GA, USA.

[25]Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA, USA.

[26]School of Public Health and Emergency Management, Southern University of Science and Technology, Shenzhen, Guangdong, China.

[27]VA Boston Healthcare System, Boston, MA, USA.

[28]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.

[29]Department of Medicine, Section of Gastroenterology, Boston University School of Medicine, Boston, MA, USA.

[30]Department of Cardiology, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA.

[31]Departments of Pediatrics and Medicine, University of California, San Francisco, San Francisco, CA, USA.

[32]Section of Digestive Diseases, Department of Internal Medicine, and Department of Pathology, Yale School of Medicine, New Haven, CT, USA.

[33]Regeneron Genetics Center, Tarrytown, NY, USA.

[34]Department of Internal Medicine, Saint Louis University, St. Louis, MO, USA.

[35]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA.

[36]Department of Internal Medicine, Virginia Commonwealth University School of Medicine, Richmond, VA, USA.

[37]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA.

[38]Program for Personalized and Genomic Medicine, Division of Endocrinology, Diabetes and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA.

[39]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK.

[40]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[41]Department of Pathology, Children's Hospital of Philadelphia, Philadelphia, PA, USA.

[42]Department of Medicine, Section of Hepatology, Lewis Katz School of Medicine at Temple University, Temple University Hospital, Philadelphia, PA, USA.

[43]Department of Medicine, Division of Gastroenterology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[44]Center for Healthcare Organization and Implementation Research, Bedford VA Healthcare System, Bedford, MA, USA.

[45]Center for Population Health, Department of Biomedical and Nutritional Sciences, University of Massachusetts, Lowell, MA, USA.

[46]Phoenix VA Health Care System, Phoenix, AZ, USA.

[47]College of Medicine, University of Arizona, Phoenix, AZ, USA.

[48]Division of Endocrinology, Emory University School of Medicine, Atlanta, GA, USA.

[49]Office of Research and Development, Veterans Health Administration, Washington, DC, USA.

[50]Department of Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA.

[51]Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA, USA.

[52]Department of Medicine, Brigham Women's Hospital, Boston, MA, USA.

[53]Department of Medicine, Harvard Medical School, Boston, MA, USA.

[54]Nashville VA Medical Center, Nashville, TN, USA.

[55]Division of Epidemiology, Department of Medicine, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA.

[56]Department of Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[57]Division of Cardiology, Emory University School of Medicine, Atlanta, GA, USA.

[58]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA.

[59]Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[60]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA.

[61]Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA.

[62]Department of Cardiology, Columbia University Irving Medical Center, New York, NY, USA.

[63]Center for Non-Communicable Diseases, Karachi, Sindh, Pakistan.

[64]Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK.

[65]Newcastle NIHR Biomedical Research Centre, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK.

[66]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA.

[67]College of Nursing and Health Sciences, University of Massachusetts, Lowell, MA, USA.

[68]Institute of Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[69]These authors contributed equally: Marijana Vujkovic, Shweta Ramdas, Daniel J. Rader, Benjamin F. Voight, Kyong-Mi Chang.

Acknowledgements

This research is based on data from the MVP, Office of Research and Development, Veterans Health Administration and was supported by award MVP000. This publication does not represent the views of the Department of Veterans Affairs, the US Food and Drug Administration or the US Government. This research was also supported by funding from the Department of Veterans Affairs awards I01- BX003362 (P.S.T. and K.-M.C.) and I01 BX003341 (H.R.K. co-principal investigator) and the VA Informatics and Computing Infrastructure VA HSR RES 130457 (S.L.D. and J.A.L.). B.F.V. acknowledges support for this work from the National Institutes of Health (NIH)/ National Institute of Diabetes and Digestive and Kidney Diseases (grants DK101478 and DK126194) and a Linda Pechenik Montague Investigator award. K.-M.C., S.M.D., J.M.G., C.J.O., L.S.P. and P.S.T. are supported by the VA Cooperative Studies Program. S.M.D. is supported by the VA (IK2 CX001780). Funding support is also acknowledged for M.S. (K23 DK115897), R.M.C. (R01 AA026302), D.E.K. (T32 HL007734), A.D.W. (R01 DK122586, R01 AI123539), J.B.M. (R01 HL151855, UM1 DK078616), W.R.W. (R01 HL137984 P41 EB029460), S.F.A.G. (R01 HD056465), L.B. (R01 LM010685), A.V.K. (K08 HG010155, U01 HG011719) and L.S.P. (VA awards CSP #2008, I01 CX001899, I01 CX001737 and I01 BX005831; NIH awards R01 DK127083, R03 AI133172, R21 AI156161, UL1 TR002378 and P30 DK111024; and Cystic Fibrosis Foundation award PHILLI12A0). The Rader lab was supported by NIH grants HL134853 (N.J.H. and D.J.R.) and DK114291-01A1 (K.T.C., N.J.H. and D.J.R.). We thank all study participants for their contribution. Support for imaging studies was provided by the University of Pennsylvania's Institute for Translational Medicine and Therapeutics (NIH NCATS UL1 TR001878), the Penn Center for Precision Medicine Accelerator Fund.

Regeneron Genetics Center

Luca A. Lotta[33], Jonas B. Nielsen[33], Adam E. Locke[33], Marcus B. Jones[33], Niek Verweij[33] and Aris Baras[33]

Geisinger-Regeneron DiscovEHR Collaboration

Luca A. Lotta[33], Jonas B. Nielsen[33], Adam E. Locke[33], Marcus B. Jones[33], Niek Verweij[33] and Aris Baras[33]

EPoS Consortium

Quentin M. Anstee[64,,65], Rebecca Darlay[6], Heather J. Cordell[6] and Ann K. Daly[65] VA Million Veteran Program

Marijana Vujkovic[1,,2,,69], Kim M. Lorenz[1,,3,,4], Kyung Min Lee[10], Marina Serper[1,,2], David E. Kaplan[1,,2], Xiang Zhu[14,,15,,16,,17], Catherine Tcheandjieu[14,,18], Rachel L. Kember[19,,20], Henry R. Kranzler[19,,20], Anurag Verma[1,,3], Ayush Giri[21], Derek M. Klarin[14,,22,,23], Yan V. Sun[24,,25], Jennifer E. Huffman[27], Donald R. Miller[44,,45], Peter D. Reaven[46,,47], Lawrence S. Phillips[24,,48], Sumitra Muralidhar[49], Scott L. DuVall[10,,50], Jennifer S. Lee[14,,18], Themistocles L. Assimes[14,,18,,51], Saiju Pyarajan[27,,52,,53], Kelly Cho[27,,52,,53], Todd L. Edwards[54,,55], Scott M. Damrauer[1,,3,,56], Peter W. Wilson[24,,57], J. Michael Gaziano[27,,52], Christopher J. O’Donnell[27,,52,,53], Philip S. Tsao[14,,18,,51], James B. Meigs[23,,53,,66], Julie A. Lynch[10,,50,,67], Benjamin F. Voight[1,,3,,4,,68,,69] and Kyong-Mi Chang[1,,2,,69]

Data availability

The full summary-level association data from the multiancestry, EA, AA, HISP and ASN analyses from this report are available through dbGAP under accession number phs001672.v7.p1 (Veterans Administration MVP Summary Results from Omics Studies). Source data are provided with this paper.

References

1. Asrani SK, Devarbhavi H, Eaton J & Kamath PS Burden of liver diseases in the world. J. Hepatol 70, 151–171 (2019). [PubMed: 30266282]
2. Younossi Z, Anstee QM & Marietti M Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat. Rev. Gastroenterol. Hepatol 15, 11–20 (2018). [PubMed: 28930295]
3. Estes C, Razavi H, Loomba R, Younossi Z & Sanyal AJ Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. Hepatology 67, 123–133 (2018). [PubMed: 28802062]
4. Carr RM, Oranu A & Khungar V Nonalcoholic fatty liver disease: pathophysiology and management. Gastroenterol. Clin. North Am 45, 639–652 (2016). [PubMed: 27837778]
5. Chalasani N et al. The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. Hepatology 67, 328–357 (2018). [PubMed: 28714183]
6. Sookoian S & Pirola CJ Genetic predisposition in nonalcoholic fatty liver disease. Clin. Mol. Hepatol 23, 1–12 (2017). [PubMed: 28268262]

7. Romeo S et al. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet 40, 1461–1465 (2008). [PubMed: 18820647]

8. Chambers JC et al. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nat. Genet 43, 1131–1138 (2011). [PubMed: 22001757]

9. Speliotes EK et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet 7, e1001324 (2011). [PubMed: 21423719]

10. Emdin CA et al. A missense variant in mitochondrial amidoxime reducing component 1 gene and protection against liver disease. PLoS Genet 16, e1008629 (2020). [PubMed: 32282858]

11. Anstee QM et al. Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort. J. Hepatol 73, 505–515 (2020). [PubMed: 32298765]

12. Gaziano JM et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol 70, 214–223 (2016). [PubMed: 26441289]

13. Husain N et al. Nonalcoholic fatty liver disease (NAFLD) in the Veterans Administration population: development and validation of an algorithm for NAFLD using automated data. Aliment Pharm. Ther 40, 949–954 (2014).

14. Serper M et al. Validating a non-invasive non-alcoholic fatty liver phenotype in the Million Veteran Program. PLoS One 15, e0237430 (2020). [PubMed: 32841307]

15. de Vries PS et al. Multiancestry genome-wide association study of lipid levels incorporating gene-alcohol interactions. Am. J. Epidemiol 188, 1033–1054 (2019). [PubMed: 30698716]

16. Kozlitina J et al. Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet 46, 352–356 (2014). [PubMed: 24531328]

17. Abul-Husn NS et al. A protein-truncating HSD17B13 variant and protection from chronic liver disease. N. Engl. J. Med 378, 1096–1106 (2018). [PubMed: 29562163]

18. Young KA et al. Genome-wide association study identifies loci for liver enzyme concentrations in Mexican Americans: the GUARDIAN Consortium. Obes. (Silver Spring) 27, 1331–1337 (2019).

19. 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 Med 17, 135 (2019). [PubMed: 31311600]

20. Chalasani N et al. Genome-wide association study identifies variants associated with histologic features of nonalcoholic Fatty liver disease. Gastroenterology 139, 1567–1576 (2010). 1576 e1–6. [PubMed: 20708005]

21. Chen VL et al. Genome-wide association study of serum liver enzymes implicates diverse metabolic and liver pathology. Nat. Commun 12, 816 (2021). [PubMed: 33547301]

22. Pazoki R et al. Genetic analysis in European ancestry individuals identifies 517 loci associated with liver enzymes. Nat. Commun 12, 2579 (2021). [PubMed: 33972514]

23. Stephens CR et al. The impact of education and age on metabolic disorders. Front Public Health 8, 180 (2020). [PubMed: 32671006]

24. Wakefield J Bayes factors for genome-wide association studies: comparison with *P*-values. Genet. Epidemiol 33, 79–86 (2009). [PubMed: 18642345]

25. Baxter M et al. Phenotypic and functional analyses show stem cell-derived hepatocyte-like cells better mimic fetal rather than adult hepatocytes. J. Hepatol 62, 581–589 (2015). [PubMed: 25457200]

26. Brancale J & Vilarinho S A single cell gene expression atlas of 28 human livers. J. Hepatol 75, 219–220 (2021). [PubMed: 34016468]

27. Goldstein JA et al. LabWAS: novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. PLoS Genet 16, e1009077 (2020). [PubMed: 33175840]

28. Sliz E et al. NAFLD risk alleles in PNPLA3, TM6SF2, GCKR and LYPLAL1 show divergent metabolic effects. Hum. Mol. Genet 27, 2214–2223 (2018). [PubMed: 29648650]

29. Stender S et al. Relationship between genetic variation at PPP1R3B and levels of liver glycogen and triglyceride. Hepatology 67, 2182–2195 (2018).

30. Mehta MB et al. Hepatic protein phosphatase 1 regulatory subunit 3B (Ppp1r3b) promotes hepatic glycogen synthesis and thereby regulates fasting energy homeostasis. J. Biol. Chem 292, 10444–10454 (2017). [PubMed: 28473467]

31. Brouwers M, Jacobs C, Bast A, Stehouwer CDA & Schaper NC Modulation of glucokinase regulatory protein: a double-edged sword? Trends Mol. Med 21, 583–594 (2015). [PubMed: 26432016]

32. Fagerberg L et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol. Cell Proteom 13, 397–406 (2014).

33. Duff MO et al. Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature 521, 376–379 (2015). [PubMed: 25970244]

34. Jamialahmadi O et al. Exome-wide association study on alanine aminotransferase identifies sequence variants in the GPAM and APOE associated with fatty liver disease. Gastroenterology 160, 1634–1646 e7 (2021). [PubMed: 33347879]

35. Hammond LE et al. Mitochondrial glycerol-3-phosphate acyltransferase-deficient mice have reduced weight and liver triacylglycerol content and altered glycerolipid fatty acid composition. Mol. Cell. Biol 22, 8204–8214 (2002). [PubMed: 12417724]

36. Cuchel M et al. Inhibition of microsomal triglyceride transfer protein in familial hypercholesterolemia. N. Engl. J. Med 356, 148–156 (2007). [PubMed: 17215532]

37. Soubeyrand S, Martinuk A & McPherson R TRIB1 is a positive regulator of hepatocyte nuclear factor 4-alpha. Sci. Rep 7, 5574 (2017). [PubMed: 28717196]

38. Laudadio I et al. A feedback loop between the liver-enriched transcription factor network and miR-122 controls hepatocyte differentiation. Gastroenterology 142, 119–129 (2012). [PubMed: 21920465]

39. Kim JY, Han YH, Nam MW, Kim HJ & Lee MO RORalpha suppresses interleukin-6-mediated hepatic acute phase response. Sci. Rep 9, 11798 (2019). [PubMed: 31409825]

40. Laatsch A et al. Low density lipoprotein receptor-related protein 1 dependent endosomal trapping and recycling of apolipoprotein E. PLoS One 7, e29385 (2012). [PubMed: 22238606]

41. Sanyal AJ et al. Pioglitazone, vitamin E, or placebo for nonalcoholic steatohepatitis. N. Engl. J. Med 362, 1675–1685 (2010). [PubMed: 20427778]

42. Musso G, Cassader M, Paschetta E & Gambino R Thiazolidinediones and advanced liver fibrosis in nonalcoholic steatohepatitis: a meta-analysis. JAMA Intern Med 177, 633–640 (2017). [PubMed: 28241279]

43. Ratziu V et al. Long-term efficacy of rosiglitazone in nonalcoholic steatohepatitis: results of the fatty liver improvement by rosiglitazone therapy (FLIRT 2) extension trial. Hepatology 51, 445–453 (2010). [PubMed: 19877169]

44. Cusi K et al. Long-term pioglitazone treatment for patients with nonalcoholic steatohepatitis and prediabetes or type 2 diabetes mellitus: a randomized trial. Ann. Intern Med 165, 305–315 (2016). [PubMed: 27322798]

45. Tilg H, Adolph TE & Moschen AR Multiple parallel hits hypothesis in nonalcoholic fatty liver disease: revisited after a decade. Hepatology 73, 833–842 (2021). [PubMed: 32780879]

46. Hamada M, Tsunakawa Y, Jeon H, Yadav MK & Takahashi S Role of MafB in macrophages. Exp. Anim 69, 1–10 (2020). [PubMed: 31582643]

47. Klarin D et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet 50, 1514–1523 (2018). [PubMed: 30275531]

48. Stoffel W et al. Obesity resistance and deregulation of lipogenesis in Delta6-fatty acid desaturase (FADS2) deficiency. EMBO Rep 15, 110–120 (2014). [PubMed: 24378641]

49. Mirea AM, Tack CJ, Chavakis T, Joosten LAB & Toonen EJM IL-1 family cytokine pathways underlying NAFLD: towards new treatment strategies. Trends Mol. Med 24, 458–471 (2018). [PubMed: 29665983]

50. Miao Z et al. Identification of 90 NAFLD GWAS loci and establishment of NAFLD PRS and causal role of NAFLD in coronary artery disease. HGG Adv 3, 100056 (2022). [PubMed: 35047847]

51. zler HR et al. Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. Nat. Commun 10, 1499 (2019). [PubMed: 30940813]

52. Justice AC et al. AUDIT-C and ICD codes as phenotypes for harmful alcohol use: association with ADH1B polymorphisms in two US populations. Addiction 113, 2214–2224 (2018). [PubMed: 29972609]

53. Chang CC et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015). [PubMed: 25722852]

54. Willer CJ, Li Y & Abecasis GR METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010). [PubMed: 20616382]

55. Hutchinson A, Watson H & Wallace C Improving the coverage of credible sets in Bayesian genetic fine-mapping. PLoS Comput. Biol 16, e1007829 (2020). [PubMed: 32282791]

56. MacLean MT et al. Quantification of abdominal fat from computed tomography using deep learning and its association with electronic health records in an academic biobank. J. Am. Med. Inform. Assoc 28, 1178–1187 (2021). [PubMed: 33576413]

57. Haas ME et al. Machine learning enables new insights into genetic contributions to liver fat accumulation. Cell Genom 1, 100066 (2021). [PubMed: 34957434]

58. Wattacheril J et al. Genome-wide associations related to hepatic histology in nonalcoholic fatty liver disease in Hispanic boys. J. Pediatr 190, 100–107 (2017). [PubMed: 28918882]

59. Patton HM et al. Clinical correlates of histopathology in pediatric nonalcoholic steatohepatitis. Gastroenterology 135, 1961–1971 (2008). [PubMed: 19013463]

60. Lin HJ et al. Home use of a compact, 12lead ECG recording system for newborns. J. Electrocardiol 53, 89–94 (2019). [PubMed: 30716528]

61. Weinshilboum RM & Wang L Pharmacogenomics: precision medicine and drug response. Mayo Clin. Proc 92, 1711–1722 (2017). [PubMed: 29101939]

62. Simon JA et al. Phenotypic predictors of response to simvastatin therapy among African-Americans and Caucasians: the Cholesterol and Pharmacogenetics (CAP) study. Am. J. Cardiol 97, 843–850 (2006). [PubMed: 16516587]

63. Hardy T et al. The European NAFLD Registry: a real-world longitudinal cohort study of nonalcoholic fatty liver disease. Contemp. Clin. Trials 98, 106175 (2020). [PubMed: 33045403]

64. Angulo P et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology 149, 389–397 (2015). [PubMed: 25935633]

65. Dewey FE et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science 354, aaf6814 (2016). [PubMed: 28008009]

66. Harrison SA et al. Selonsertib for patients with bridging fibrosis or compensated cirrhosis due to NASH: Results from randomized phase III STELLAR trials. J. Hepatol 73, 26–39 (2020). [PubMed: 32147362]

67. Roden DM et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin. Pharmacol. Ther 84, 362–369 (2008). [PubMed: 18500243]

68. Bulik-Sullivan B et al. An atlas of genetic correlations across human diseases and traits. Nat. Genet 47, 1236–1241 (2015). [PubMed: 26414676]

69. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]

70. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]

71. Andersson R et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014). [PubMed: 24670763]

72. Finucane HK et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet 47, 1228–1235 (2015). [PubMed: 26414678]

73. Finucane HK et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet 50, 621–629 (2018). [PubMed: 29632380]

74. CARDIoGRAMplusC4D Consortium et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat. Genet 45, 25–33 (2013). [PubMed: 23202125]

75. Fehrmann RS et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat. Genet 47, 115–125 (2015). [PubMed: 25581432]

76. Cahoy JD et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J. Neurosci 28, 264–278 (2008). [PubMed: 18171944]

77. Heng TS & Painter MW, Immunological Genome Project Consortium. The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol 9, 1091–1094 (2008). [PubMed: 18800157]

78. Pers TH et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun 6, 5890 (2015). [PubMed: 25597830]

79. Schmidt EM et al. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics 31, 2601–2606 (2015). [PubMed: 25886982]

80. McLaren W et al. The ensembl variant effect predictor. Genome Biol 17, 122 (2016). [PubMed: 27268795]

81. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). [PubMed: 26432245]

82. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat. Genet 45, 580–585 (2013). [PubMed: 23715323]

83. Chesi A et al. Genome-scale Capture-C promoter interactions implicate effector genes at GWAS loci for bone mineral density. Nat. Commun 10, 1260 (2019). [PubMed: 30890710]

84. Pashos EE et al. Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci. Cell Stem Cell 20, 558–570 (2017). [PubMed: 28388432]

85. Caliskan M et al. Genetic and epigenetic fine mapping of complex trait associated loci in the human liver. Am. J. Hum. Genet 105, 89–107 (2019). [PubMed: 31204013]

86. Wingett S et al. HiCUP: pipeline for mapping and processing Hi-C data. F1000Res 4, 1310 (2015). [PubMed: 26835000]

87. Cairns J et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol 17, 127 (2016). [PubMed: 27306882]

88. Szklarczyk D et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47, D607–D613 (2019). [PubMed: 30476243]

89. Gagliano Taliun SA et al. Exploring and visualizing large-scale genetic associations by using PheWeb. Nat. Genet 52, 550–552 (2020). [PubMed: 32504056]

90. Denny JC et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26, 1205–1210 (2010). [PubMed: 20335276]

91. Loh PR et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet 48, 1443–1448 (2016). [PubMed: 27694958]

92. Zhou W et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet 50, 1335–1341 (2018). [PubMed: 30104761]

93. Elsworth B et al. The MRC IEU OpenGWAS data infrastructure. Preprint at *bioRxiv*, 10.1101/2020.08.10.244293 (2020).

94. Shin S et al. CREB mediates the insulinotropic and anti-apoptotic effects of GLP-1 signaling in adult mouse beta-cells. Mol. Metab 3, 803–812 (2014). [PubMed: 25379405]

95. Kettunen J et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat. Commun 7, 11122 (2016). [PubMed: 27005778]

96. Hemani G et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife 7, e34408 (2018). [PubMed: 29846171]

97. Sun BB et al. Genomic atlas of the human plasma proteome. Nature 558, 73–79 (2018). [PubMed: 29875488]

98. Giambartolomei C et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 10, e1004383 (2014). [PubMed: 24830394]

99. Giri A et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. Nat. Genet 51, 51–62 (2019). [PubMed: 30578418]

100. Pulit SL et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum. Mol. Genet 28, 166–174 (2019). [PubMed: 30239722]

101. Teumer A et al. Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. Nat. Commun 10, 4130 (2019). [PubMed: 31511532]

102. Tin A et al. Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. Nat. Genet 51, 1459–1474 (2019). [PubMed: 31578528]

103. Wuttke M et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat. Genet 51, 957–972 (2019). [PubMed: 31152163]

104. Astle WJ et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell 167, 1415–1429 (2016). [PubMed: 27863252]

105. Guo H et al. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. Hum. Mol. Genet 24, 3305–3313 (2015). [PubMed: 25743184]

106. Lambert SA et al. The human transcription factors. Cell 175, 598–599 (2018). [PubMed: 30290144]

107. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D & Saez-Rodriguez J Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res 29, 1363–1375 (2019). [PubMed: 31340985]

108. Turei D, Korcsmaros T & Saez-Rodriguez J OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat. Methods 13, 966–967 (2016). [PubMed: 27898060]

109. Ceccarelli F, Turei D, Gabor A & Saez-Rodriguez J Bringing data from curated pathway resources to Cytoscape with OmniPath. Bioinformatics 36, 2632–2633 (2020). [PubMed: 31886476]

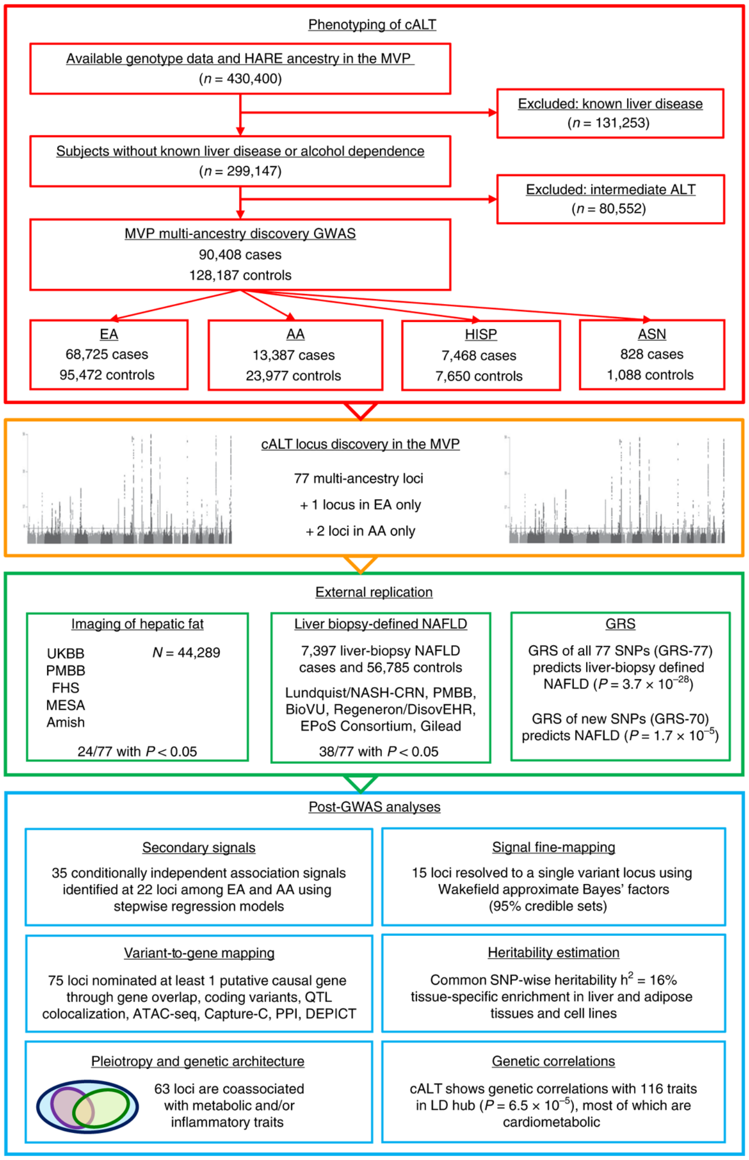
Phenotyping of cALT
Available genotype data and HARE ancestry in the MVP
(n = 430,400)
Excluded: known liver disease
(n = 131,253)
Subjects without known liver disease or alcohol dependence
(n = 299,147)
Excluded: intermediate ALT
(n = 80,552)
MVP multi-ancestry discovery GWAS
90,408 cases
128,187 controls
EA
68,725 cases
95,472 controls
AA
13,387 cases
23,977 controls
HISP
7,468 cases
7,650 controls
ASN
828 cases
1,088 controls
cALT locus discovery in the MVP
77 multi-ancestry loci
+ 1 locus in EA only
+ 2 loci in AA only
External replication
Imaging of hepatic fat
UKBB
PMBB
FHS
MESA
Amish
N = 44,289
24/77 with P < 0.05
Liver biopsy-defined NAFLD
7,397 liver-biopsy NAFLD cases and 56,785 controls
Lundquist/NASH-CRN, PMBB, BioVU, Regeneron/DisovEHR, EPoS Consortium, Gilead
38/77 with P < 0.05
GRS
GRS of all 77 SNPs (GRS-77) predicts liver-biopsy defined NAFLD ($P = 3.7 \times 10^{-28}$)
GRS of new SNPs (GRS-70) predicts NAFLD ($P = 1.7 \times 10^{-5}$)
Post-GWAS analyses
Secondary signals
35 conditionally independent association signals identified at 22 loci among EA and AA using stepwise regression models
Signal fine-mapping
15 loci resolved to a single variant locus using Wakefield approximate Bayes' factors (95% credible sets)
Variant-to-gene mapping
75 loci nominated at least 1 putative causal gene through gene overlap, coding variants, QTL colocalization, ATAC-seq, Capture-C, PPI, DEPICT
Heritability estimation
Common SNP-wise heritability h^2 = 16%
tissue-specific enrichment in liver and adipose tissues and cell lines
Pleiotropy and genetic architecture
63 loci are coassociated with metabolic and/or inflammatory traits
Genetic correlations
cALT shows genetic correlations with 116 traits in LD hub ($P = 6.5 \times 10^{-5}$), most of which are cardiometabolic

Fig. 1 |. Overview of analysis pipeline.

The flow diagram shows in the red box our study design with initial inclusion of 430,400 MVP participants with genotyping and ancestry classification by HARE, exclusion of individuals with known liver disease or alcohol dependence and inclusion of participants based on cALT (case) or normal ALT (control). This resulted in 90,408 proxy NAFLD cases and 128,187 controls with EA, AA, HISP and ASN ancestries that were examined in primary multiancestry and ancestry-specific genome-wide association scans. The orange box of the flow diagram highlights our results of multiancestry and ancestry-specific meta-analyses identifying 77 multiancestry loci + 1 EA-specific locus + 2 AA-specific loci that met genome-wide significance. The green box summarizes the results from external replication cohorts, whereas the blue box indicates all the post-GWAS annotation analyses that we performed, which include secondary signal analysis, fine-mapping (95% credible sets), (tissue-specific) heritability estimation, genetic correlations analysis, variant-to-gene-mapping and pleiotropy analysis. ATAC-seq, assay for transposase-accessible chromatin with high-throughput sequencing; FHS, Framingham Heart Study; GRS, genetic risk score; MESA, Multi-Ethnic Study of Atherosclerosis; NASH-CRN, Non-Alcoholic Steatohepatitis Clinical Research Network; PMBB, Penn Medicine Biobank; PPI, protein–protein interaction; QTL, quantitative trait locus; UKBB, UK Biobank; HARE, harmonized ancestry and race/ethnicity; BioVU, the DNA databank at Vanderbilt; Regeneron/DiscovEHR, collaboration between the Regeneron Genetics Center and Geisinger Health System; DEPICT, data-driven expression prioritized integration for complex traits; ldhub, a centralized database and web interface to perform LD score regression; Amish, University of Maryland Old Order Amish Study.

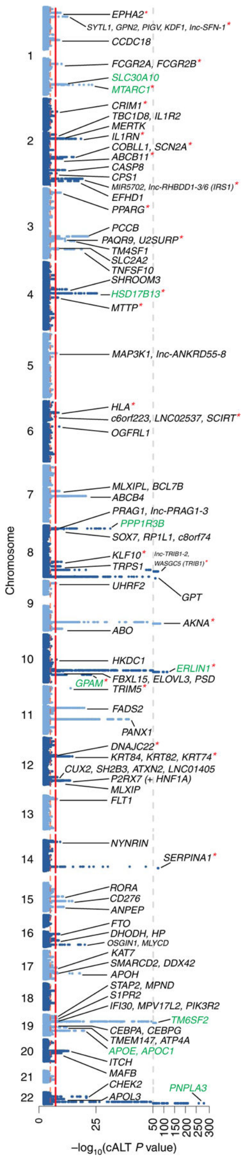

Fig. 2 |. Manhattan plot of GWAS of 90,408 cALT cases and 128,187 controls in multiancestry meta-analysis.

Nominated genes are indicated for 77 loci reaching genome-wide significance ($P < 5 \times 10^{-8}$). Previously reported NAFLD loci with genome-wide significant association are indicated in green font. Red asterisks indicate the SNPs that have been replicated with liver biopsy and/or radiologic imaging.

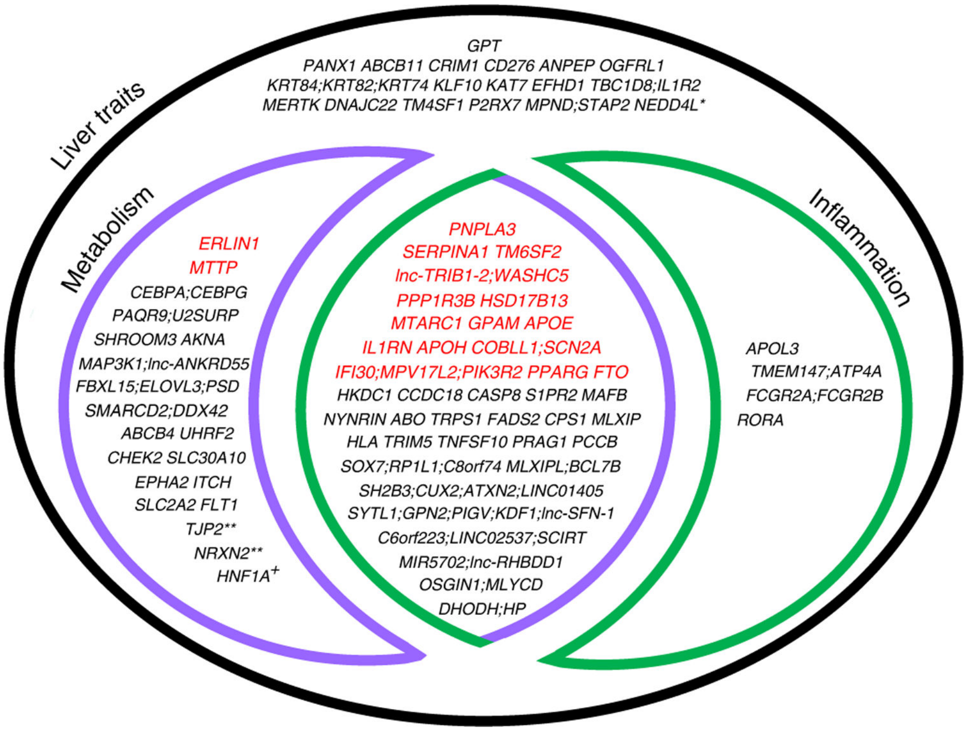

Fig. 3 |. Venn diagram depicting overlapping liver, metabolic and inflammatory traits among cALT-associated loci.

Overlapping liver (black), metabolic (purple) and/or inflammatory (green) traits are shown in association with 77 multiancestry and additional ancestry-specific lead SNPs. The trait categorizations reflect significant SNP–trait associations identified by (1) LabWAS of clinical laboratory results in MVP, (2) PheWAS with UK Biobank data using SAIGE, (3) SNP lookup using the curated data in the IEU OpenGWAS projects and (4) cross-trait colocalization analyses using colocalization analysis in EA, AA and HISP lead loci with 36 other GWAS statistics of cardiometabolic and blood cell-related traits. Genes denoted in bold and color-coded in red refer to the loci also associated with qHF on imaging analyses or histologically characterized NAFLD from liver biopsies. *Locus identified in European-only GWAS. **Locus identified in AA-restricted analysis. +Secondary signal from EA analysis (e.g., *HNF1A*/*P2RX7*).

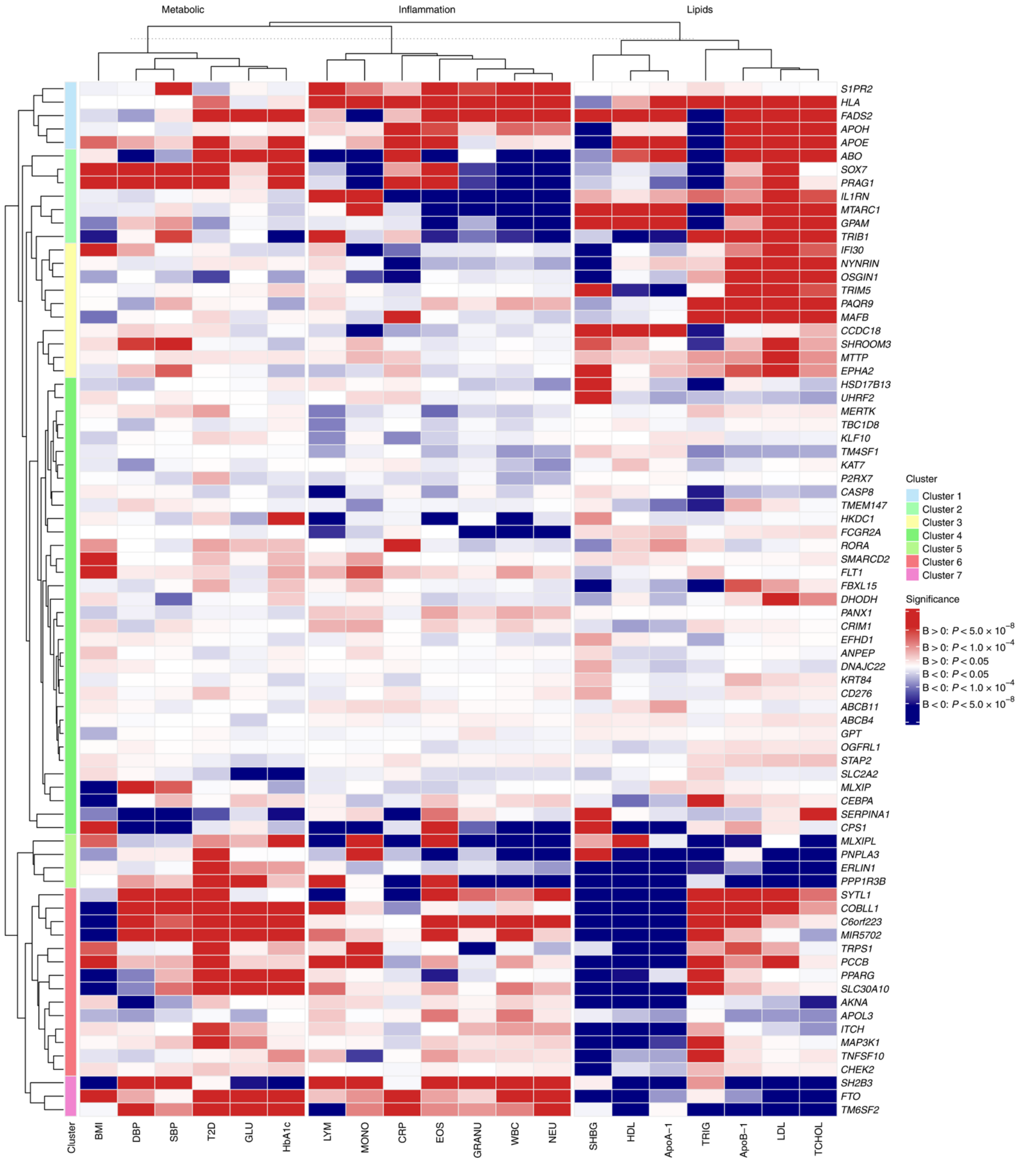

Fig. 4 |. Seven gene clusters with distinct biomarker association profiles.

The 77 loci cluster along seven groups using stratified linkage hierarchical clustering. Each cluster has a distinct biomarker association profile, which is visualized with a heatmap. The 26 traits are clustered within their biological strata (e.g., lipids, inflammation, metabolic and liver). The color coding corresponds to the direction of association of the cALT-increasing allele (red, positive association; blue, negative association) and the strength of the association based on the *P* value. DBP, diastolic blood pressure; SBP, systolic blood pressure; T2D, type 2 diabetes; GLU, glucose; HbA1c, hemoglobin A1c; LYM, lymphocyte count; MONO, monocyte count; CRP, C-reactive protein; EOS, eosinophil count; GRANU, granulocyte count; WBC, white blood cell count; NEU, neutrophil count; SHBG, sex hormone binding globulin; ApoA-1, apolipoprotein A1; TRIG, triglycerides; ApoB-1, apolipoprotein B1; TCHOL, total cholesterol.

Table 1 |

Gene nominations at loci with strongest evidence for coding variants

SNP	Position	Gene	Amino Acid change	SIFT/PolyPhen-2[a]	Expression/ splicingQTL[b]	Other[c]	Pleiotropy[d]
rs6541349	1:93787867	*CCDC18*	p.Leu1134Val	+/−	+	.	M
rs2642438	1:220970028	*MTARC1*	p.Thr165Ala	−/−	+ (A)	+	M
rs11683409	2:112770134	*MERTK*	p.Arg466Lys	−/−	.	++	.
rs17036160	3:12329783	*PPARG*	p.Pro12Ala	−/−	+	++	M
rs17598226	4:100496891	*MTTP*	p.Ile128Thr	−/−	+	+	.
rs115038698	7:87024718	*ABCB4*	p.Ala934Thr	+/+	+	+	M,I
rs799165	7:73052057	*MLXIPL*	p.Gln241His	−/+	+	+	M,I
			p.Ala358Val	−/−	+	+	M,I
rs7041363	9:117146043	*AKNA*	p.Pro624Leu	+/−	+	+	M
rs10883451	10:101924418	*ERLIN1*	p.Ile291Val	−/−	.	++	M
rs4918722	10:113947040	*GPAM*	p.Ile43Val	−/−	+	++	M
rs11601507	11:5701074	*TRIM5*	p.Val112Phe	−/−	.	++	M,I
rs1626329	12:121622023	*P2RX7*	p.Ala348Thr	−/−	+	+	.
rs11621792	14:24871926	*NYNRIN*	p.Ala978Thr	−/−	+ (L,A)	+	M,I
rs28929474	14:94844947	*SERPINA1*	p.Glu366Lys	−/+		+++	M,I
rs7168849	15:90346227	*ANPEP*	p.Ala311Val	−/−	+ (L)	+	.
rs1801689	17:64210580	*APOH*	p.Cys325Gly	+/+	.	++	M,I
rs132665	22:36564170	*APOL3*	p.Ser39Arg	−/−	+ (A)	+	.
rs738408	22:44324730	*PNPLA3*	p.Ile148Met	+/+	.	+++	M,I

Genes nominated with various sources of evidence are listed as follows.

[a] Before the slash symbol, '+' indicates 'deleterious' in SIFT ('−' otherwise). After the slash symbol, '+' denotes probably damaging in PolyPhen-2 ('−' otherwise).

[b] The '+' indicates colocalization between NAFLD GWAS variant and GTEx QTL variant (posterior probability (PP) PP4/(PP3 + PP4) > 0.9). (L) denotes QTL effect in liver, (A) denotes QTL in adipose.

[c] Each '+' represents evidence from DEPICT or PPI data, or if the lead SNP is within the transcript; coding variants also include '+' from hQTLs/Capture-C evidence.

[d] Pleiotropy is limited to association with metabolic (M) or inflammatory (I) traits.

Table 2 |

Gene nominations at loci with strongest evidence for noncoding variants

SNP	Position	Gene	hQTL	Capture-C	Expression/splicingQTL[a]	Other[b]	Pleiotropy[c]
rs36086195	1:16510894	*EPHA2*	.	+	+ (L,A)	+	M
rs6734238	2:113841030	*IL1RN*	.	+	+ (A)	++	I
rs10201587	2:202202791	*CASP8*	.	+	+	+	M
rs11683367	2:233510011	*EFHD1*	+	.	+ (L)	+	.
rs61791108	3:170732742	*SLC2A2*	.	+	.	+++	M
rs7653249	3:136005792	*PCCB*	.	.	+	++	M,I
rs12500824	4:77416627	*SHROOM3*	.	+	+ (L)	+	M
rs10433937	4:88230100	*HSD17B13*	.	.	+ (L,A)	+	M,I
rs799165	7:73052057	*BCL7B*	.	+	+	+	M,I
rs687621	9:136137065	*ABO*	.	.	+	+	M,I
rs35199395	10:70983936	*HKDC1*	.	+	+ (L,A)	+	M
rs174535	11:61551356	*FADS2*	+	.	+ (A)	++	M,I
rs56175344	11:93864393	*PANX1*	.	.	+ (L,A)	++	.
rs34123446	12:122511238	*MLXIP*	.	+	+	+	M,I
rs12149380	16:72043546	*DHODH*	.	+	+	+	M,I
		HP	.	+	+ (A)	.	M,I
rs2727324	17:61922102	*DDX42*	.	+	+	+	M
		SMARCD2	.	.	+	+	M
rs5117	19:45418790	*APOC1*	.	.	+	++	M,I

Genes nominated with various sources of evidence are listed as follows.

[*] Prior to the slash symbol: '+' indicates 'deleterious' in SIFT and '−' otherwise. After slash symbol: '+' denotes probably damaging in PolyPhen-2 and '−' otherwise.

[a] The '+' indicates colocalization between NAFLD GWAS variant and GTEx QTL varint (PP4/(PP3 + PP4) > 0.9). (L) denotes QTL effect in liver, (A) denotes QTL in adipose.

[b] Each '+' represent evidence from DEPICT, PPI data, or if the lead SNP is within the transcript; coding variants also include '+' from hQTLs/ Capture-C evidence.

[c] Pleiotropy is limited to association with metabolic (M) or inflammatory (I) traits.