A genome-wide association study for nonalcoholic fatty liver disease identifies novel genetic loci and trait-relevant candidate genes in the Million Veteran Program.

Marijana Vujkovic[1]*, Shweta Ramdas[2]*, Kimberly M. Lorenz[1,2,3], Carolin V. Schneider[2], Joseph Park[2,4], Kyung M. Lee[5], Marina Serper[1], Rotonya M. Carr[4], David E. Kaplan[1], Mary E. Haas[6], Matthew T. MacLean[2], Walter R. Witschey[7], Xiang Zhu[8,9,10,11], Catherine Tcheandjieu[11,12], Rachel L. Kember[13,14], Henry R. Kranzler[13,14], Anurag Verma[1,2], Ayush Giri[15], Derek M. Klarin[16,17,18], Yan V. Sun[19,20], Jie Huang[21], Jennifer Huffman[16], Kate Townsend Creasy[2], Nicholas J. Hand[2], Ching-Ti Liu[22], Michelle T. Long[23], Jerome I. Rotter[24], Xiuqing Guo[24], Jie Yao[24], Matthew Budoff[25], Katherine A. Ryan[26], Braxton D. Mitchell[27], Dipender Gill[28], Andrew D. Wells[29], Elisabetta Manduchi[30], Yedidya Saiman[31], Nadim Mahmud[31], Donald R. Miller[32,33], Peter D. Reaven[34,35], Laurence S. Phillips[19,36], Sumitra Muralidhar[37], Scott L. DuVall[5,38], Jennifer S. Lee[11,12], Themistocles L. Assimes[11,12], Saiju Pyarajan[16,39,40], Kelly Cho[16,39], Todd L. Edwards[41,42], Scott M. Damrauer[1,43], Peter W. Wilson[19,44], John M. Gaziano[16,39], Christopher J. O'Donnell[16,39,40], Amit V. Khera[17,18,40], Struan F.A. Grant[45], Christopher D. Brown[2], Philip S. Tsao[11,12], Danish Saleheen[46,47,48], James B. Meigs[18,40,49] Julie A. Lynch[5,38], Daniel J. Rader*[2,4], Benjamin F. Voight*[1,2,3,51], Kyong-Mi Chang[1,3*]

*These authors contributed equally

Affiliations

1. Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA 2. Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA 3. Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA 4. Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA 5. VA Salt Lake City Health Care System, Salt Lake City, UT, USA 6. Broad Institute of MIT and Harvard, Cambridge, MA, USA 7. Department of Radiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. 8. Department of Statistics, The Pennsylvania State University, University Park, PA, USA 9. Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA 10. Department of Statistics, Stanford University, Stanford, CA, USA 11. VA Palo Alto Health Care System, Palo Alto, CA, USA 12. Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA 13. Mental Illness Research, Education and Clinical Center, Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA 14. Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia PA, USA 15. Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville TN, USA 16. VA Boston Healthcare System, Boston MA, USA 17. Center for Genomic Medicine, Massachusetts General Hospital, Boston MA, USA 18. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge MA, USA 19. Atlanta VA Medical Center, Decatur GA, USA. 20. Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta GA, USA 21. Department of Global Health, School of Public Health, Peking University, Beijing, China. 22. Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA 23. Department of Medicine, Section of Gastroenterology, Boston University School of Medicine, Boston, MA, USA 24. Department of Pediatrics, Genomic Outcomes, The Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA 25 Cardiology, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA 26 VISN 5 Capitol Health Care Network Mental Illness Research Education and Clinical Center, Baltimore, MD, USA 27 Program for Personalized and Genomic Medicine, Division of Endocrinology, Diabetes and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA 28. Department of Epidemiology and Biostatistics, School of Public Health, Imperial

College London, London, UK. 29. Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine and Children's Hospital of Philadelphia, Philadelphia PA, USA 30. Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia PA, USA. 31. Department of Medicine, Division of Gastroenterology, University of Pennsylvania Perelman School of Medicine, Philadelphia PA, USA. 32. Edith Nourse Rogers Memorial VA Hospital, Bedford, MA, USA. 33. Center for Population Health, University of Massachusetts, Lowell, MA, USA. 34. Phoenix VA Health Care System, Phoenix, AZ, USA. 35. College of Medicine, University of Arizona, Tuscon, AZ, USA. 36. Division of Endocrinology, Emory University School of Medicine, Atlanta, GA, USA. 37. Office of Research and Development, Veterans Health Administration, Washington, DC, USA. 38. Department of Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA. 39. Department of Medicine, Brigham Women's Hospital, Boston, MA, USA. 40. Department of Medicine, Harvard Medical School, Boston, MA, USA. 41. Nashville VA Medical Center, Nashville, TN, USA. 42. Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. 43. Department of Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. 44. Division of Cardiology, Emory University School of Medicine, Atlanta GA, USA. 45. Department of Pediatrics, University of Pennsylvania Perelman School of Medicine and Children's Hospital of Philadelphia, Philadelphia, PA, USA. 46. Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA. 47. Department of Cardiology, Columbia University Irving Medical Center, New York, NY, USA. 48. Center for Non-Communicable Diseases, Karachi, Sindh, Pakistan. 49. Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. 50. College of Nursing and Health Sciences, University of Massachusetts, Lowell, MA, USA. 51. Institute of Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

Abstract

Nonalcoholic fatty liver disease (NAFLD) is a prevalent, heritable trait that can progress to cancer and liver failure. Using our recently developed proxy definition for NAFLD based on chronic liver enzyme elevation without other causes of liver disease or alcohol misuse, we performed a multi-ancestry genome-wide association study in the Million Veteran Program with 90,408 NAFLD cases and 128,187 controls. Seventy-seven loci exceeded genome-wide significance of which 70 were novel, with an additional European-American specific and two African-American specific loci. Twelve of these loci were also significantly associated with quantitative hepatic fat on radiological imaging (n=44,289). Gene prioritization based on coding annotations, gene expression from GTEx, and functional genomic annotation identified candidate genes at 97% of loci. At eight loci, the allele associated with lower gene expression in liver was also associated with reduced risk of NAFLD, suggesting potential therapeutic relevance. Functional genomic annotation and gene-set enrichment demonstrated that associated loci were relevant to liver biology. We expand the catalog of genes influencing NAFLD, and provide a novel resource to understand its disease initiation, progression and therapy.

Introduction

Chronic liver disease is a major contributor to global morbidity and mortality, with complications of cirrhosis and hepatocellular carcinoma[1]. In particular, nonalcoholic fatty liver disease (NAFLD) is an increasingly common cause of chronic liver disease with an estimated world prevalence of 25% among adults and associated metabolic risk factors[1-5]. In the United States (US), NAFLD prevalence is projected to reach 33.5% among adult population by 2030, due in large part to the rising obesity and associated metabolic disorders[6]. NAFLD is defined by ≥5% fat accumulation in the liver (hepatic steatosis) in the absence of other known causes for liver disease, based on liver biopsy and/or non-invasive radiological imaging[3,4]. The clinical spectrum of NAFLD ranges from benign steatosis to nonalcoholic steatohepatitis (NASH) involving inflammation and hepatocellular injury with fibrosis progression. At least 20% of patients with NAFLD develop NASH with increased risk of consequent cirrhosis and liver cancer[5,6]. To date, there is no licensed drug approved to treat NAFLD and prevent its progression.

Individual susceptibility to NAFLD involves both genetic and environmental factors. Risk factors for NAFLD include obesity (in particular, abdominal adiposity), insulin resistance and features of metabolic syndrome[2,5-7], with current estimates of NAFLD heritability ranging between 20% to 50%[8]. Several genetic variants that promote the full spectrum of fatty liver disease have been identified in genome-wide association studies (GWAS) utilizing cohorts based on liver biopsy, imaging, and/or isolated liver enzyme values[9-22]. The most prominent of these include p.I148M in *PNPLA3* and p.E167K in *TM6SF2*, which increase NAFLD risk, and a loss-of-function variant in *HSD17B13* that confers protection against NASH[16]. However, the limited number of genetic associations in NAFLD contrasts with other cardiometabolic disorders where hundreds of loci have been mapped to date, traits that include obesity[23,24], type 2 diabetes[25] and

plasma lipids[26]. This also highlights the need for expanded discovery based on larger sample size and population diversity, with further integration with existing functional genomics data sets to identify candidate genes from leading, non-coding associations[27].

The Million Veteran Program (MVP) is among the world's largest and ancestrally diverse biobanks[28]. The availability of comprehensive, longitudinally collected Veterans Health Administration (VA) electronic health records for US Veteran participants in the MVP also makes it a promising resource for precision medicine. As NAFLD is markedly underdiagnosed clinically due to limited access to liver biopsy and variable use of imaging modalities[4], we recently developed and validated a proxy phenotype for NAFLD to facilitate case identification in MVP[21]. The proxy NAFLD phenotype is based on chronically elevated serum alanine aminotransferase (cALT) levels while excluding other conditions that are known to increase liver enzymes (e.g. viral hepatitis, alcohol dependence, autoimmune liver disease and known hereditary liver disease). We applied this cALT-based proxy NAFLD phenotype to the current build of 430,400 genotyped MVP participants with defined ancestry classification[29], and identified 90,408 NAFLD cases and 128,187 controls (inclusion/exclusion criteria for the remaining samples and study design described in **Supplementary Figure 1** and **Figure 1**). We performed a primary GWAS and identified 77 trans-ancestry loci that reached genome-wide significance. We used additional approaches to define NAFLD heritability and genetic correlations with various traits including quantitative hepatic fat measured by liver imaging with computed tomography (CT) and magnetic resonance imaging (MRI), in addition to identifying coding variants in putative causal genes.

Results

Diverse NAFLD case and control subjects enriched for metabolic disorders in MVP.

Our study consisted of 90,408 NAFLD cases and 128,187 controls across two stages and comprising four ancestral groups, namely European Americans (EA, 75.1%), African Americans (AA, 17.1%), Hispanic-Americans (HISP, 6.9%), and Asian-Americans (ASN, 0.9%, **Supplemental Table 1**) with the overall sample sizes and study design shown in **Figure 1 and Supplemental Figure 1**. Consistent with the US Veteran population, MVP cases and controls (n = 218,595) were predominantly male (92.3%) with an average age of 64 years at study enrollment (**Supplemental Table 1**). With the exclusion of other known causes of liver disease in our phenotype definition[21], our cohort was enriched for metabolic disorders, with higher prevalence in cases as compared to controls for type 2 diabetes (71% vs. 47%, $P<1x10^{-5}$), hypertension (73% vs. 60%, $P<1x10^{-5}$ and dyslipidemia (82% vs. 70%, $P<1x10^{-5}$).

Identification of novel trans-ancestry and ancestry-specific NAFLD-associated loci in the diverse MVP population

To identify loci associated with NAFLD, we performed ancestry-specific genome-wide scans by meta-analyzing summary statistics derived from each ancestry followed by trans-ancestry meta-analysis combining data across all ancestries and stages (**Methods** and **Figure 1**). In the trans-ancestry scan across stages, 77 independent sentinel SNPs exceeded genome-wide significance ($P < 5x10^{-8}$), of which 70 were novel whereas 7 (namely *PNPLA3*, *TM6SF2*, *ERLIN1*, *TNKS [PPP1R3B]*, *MARC1*, *HSD17B13*, and *LYPLAL1*) had previously reported genome-wide significant associations with NAFLD (within 500kb and/or CEU r^2 LD > 0.05; **Figure 2 and Supplemental Table**

2)[9-22]. In addition, 55 loci in EAs, 8 loci in AAs, and 3 loci in HISPs, exceeded genome-wide significance (**Supplemental Tables 3-5** and **Supplemental Figures 2-4)**. One SNP (rs4940689) reached genome-wide significance in an ancestry-specific analysis of EAs only (**Supplemental Table 3**), whereas two SNPs (rs144127357; rs2666559) reached genome-wide significance among AAs only (**Supplemental Table 4**). No loci in ASNs achieved genome-wide significance, likely due to limited sample size in this group (**Supplemental Figure 5**).

Among the eight AA-specific lead SNPs, three were intronic: rs115038698 in the *ABCB4* locus, rs144127357 in *TJP2,* and rs2666559 in *NRXN2*. Two of these variants were nearly monomorphic in EA but polymorphic in AA (rs115038698 MAF AA: 1.2%, MAF EA: 0%; rs144127357 MAF AA: 3.14%, MAF EA: <0.001%). In contrast, the tagged variant rs2666559 was common in both AA (MAF = 19.1% in AA, gnomAD AF = 17.2% in Africans) and EA (AF = 69.1% in EA, gnomAD AF = 68.4% in non-Finnish Europeans).

Internal and external replication of NAFLD-associated loci.

We next compared the extent of association across both Stage 1 (primary analysis) and Stage 2 replication stage internally in MVP and externally in the Penn Medicine Biobank (PMBB, **Methods**). Of the 77 associated SNPs from the trans-ancestry meta-analysis, 56 reached genome-wide significance in Stage 1 subset, of which 32 passed Bonferroni significance (0.05/56) in Stage 2 replication in MVP (**Supplemental Table 2**). All 77 SNPs showed directional concordance in effect estimates between the two stages. External replication for our trans-ancestry lead SNPs in PMBB (n=72 of our loci were genotyped) with 2,570 cases and 3,802 controls demonstrated that 8 out of 72 variants were directionally consistent and nominally associated (signed binomial-test

175 P=4.4x10^{-4}). Furthermore, 4 out of 8 loci discovered in the AA-specific scan (signed bionomial-
176 test P=2.5x10^{-5}) and 1 of 3 loci discovered in the HISP-specific scan (signed bionomial test P=0.07)
177 were also directionally consistent and nominally associated in PMBB (**Supplemental Tables 6-9**).
178 In summary, we found 73 novel loci associated with NAFLD that were identified by trans and
179 single-ancestry association studies and supported by replication in multiple stages and studies.
180
181 *Concordance of cALT-based NAFLD loci with CT/MRI-based quantitative hepatic fat*
182 To place our discoveries into physiological context, we next investigated the extent to which the
183 77 trans-ancestry SNPs from our NAFLD GWAS associated with quantitative measures of hepatic
184 fat, derived from CT/MRI imaging studies (**Methods**). We performed a trans-ancestry meta-
185 analysis among 44,289 participants in the UK Biobank, PMBB, Framingham Heart Study,
186 University of Maryland Old Order Amish Study, and Multi-Ethnic Study of Atherosclerosis
187 (**Supplemental Table 10**). We found that 24 SNPs were nominally associated with quantitative
188 hepatic fat (P < 0.05), of which 12 (15.6% of 77 loci) exceeded Bonferroni multi-test correction (P
189 < 6.5x10^{-4}, including *PNPLA3, TM6SF2, APOC1;APOE, GPAM, MARC1, KIAA0196 [TRIB1], MTTP,*
190 *APOH, PIK3R2;IFI30;MPV17L2, TNKS [PPP1R3B], COBLL1;SCN2A* and *PPARG*). Notably, *PNPLA3*,
191 *TM6SF2*, and *TNKS [PPP1R3B]* were previously identified using imaging data[11,12,14], and the
192 direction of effect for all significant SNPs was concordant between chronic ALT elevation and
193 hepatic fat, with the known exception of the variant at the *PPP1R3B* locus[12].
194
195 *Identification of additional independent NAFLD-associated variants by conditional analysis*

To discover additional variants independent of our lead NAFLD signals, we next performed approximate conditional analysis using the leading sentinel variants at our 77 trans-ancestry associated loci. We detected a total of 41 conditionally independent SNPs flanking three known (*PNPLA3*, *HSD17B13*, and *ERLIN1*) and 17 novel NAFLD loci in EA (**Supplemental Table 11**). Nine conditionally independent SNPs were observed at the *PNPLA3* locus in MVP, indicative of the complexity of this locus. For one of the novel loci, located on chromosome 12 between 121-122Mb, the trans-ancestry lead variant (rs1626329) was located in *P2RX7*, whereas the lead peak for EA mapped to *HNF1A* (rs1169292, **Figure 3**). Both are strongly linked to distinct coding variants (*P2RX7*: rs1718119, Ala348Thr and *HNF1A*: rs1169288, Ile27Leu) and are compelling candidate genes for metabolic liver disease. In AA, we observed eight conditionally independent variants at four genomic loci, one at *PNPLA3* and three novel loci (**Supplemental Table 12**), including four in *GPT*, two in *AKNA*, one in *ABCB4*. In HISP, two conditionally independent variants in the *PNPLA3* locus were identified (**Supplemental Table 13**). Collectively, 51 additional variants were identified at 24 loci across ancestries based on conditional analysis.

Fine mapping to define potential causal variants in 95% credible sets

To leverage increased sample size and population diversity to improve fine-mapping resolution, we computed 95% credible sets using Wakefield's approximate Bayes' factors[30] derived from the trans-ancestry meta-regression, EA, AA, and HISP scans (**Supplemental Table 14-17, Methods**). In a comparison of the trans-ancestry and EA-only scans, the trans-ancestry meta-regression reduced the median 95% credible set size from 9 (IQR 3 - 17) to 7.5 variants (IQR 2 - 13). A total of 11 distinct NAFLD associations were resolved to a single SNP in the trans-ancestry meta-

regression, with 4 additional loci suggestive a single SNP in the EA (n=2) and AA (n=2) ancestry-specific meta-analyses that were not already resolved to a single SNP via the trans-ancestry analysis.

Heritability of NAFLD and genetic correlations with other phenotypes.

To tabulate trait heritability and genetic correlation with others, we utilized LD score regression[31-33] (**Methods**). Consistent with our discovery of novel genetic associations, we estimated the SNP-based liability-scaled heritability at 16% (95% CI: 12-19, $P < 1.0x10^{-6}$) in EA. Genome-wide genetic correlations of NAFLD were calculated with a total of 774 complex traits and diseases by comparing allelic effects using LD score regression with the EA-specific NAFLD summary statistics. A total of 116 significant associations were observed (Bonferroni correction for 774 traits $P < 6.5x10^{-5}$, **Supplemental Table 18**). Consistent with the previous epidemiological associations with metabolic syndrome traits, we observed strong correlations with cardiometabolic risk factors including measures of obesity and adiposity, type 2 diabetes, hypertension, dyslipidemia, coronary artery disease, family history of metabolic risk factors and general health conditions in addition to educational attainment.

Liver-specific enrichment of NAFLD heritability

To ascertain the tissues contributing to the disease-association underlying NAFLD heritability, we performed tissue-specific analysis using stratified LD score regression. The strongest associations were observed in genomic annotations surveyed in liver, hepatocytes, adipose, and immune cell types among others (e.g., liver histone H3K36me3 and H3K4me1, adipose nuclei H3K27ac, spleen

TCRγδ, eosinophils in visceral fat; **Supplemental Table 19**). Medical subject heading (MeSH)-based analysis showed enrichment mainly in hepatocytes and liver (False Discovery Rate (FDR) < 5%, **Supplemental Table 20**). Gene set analysis showed strongest associations for liver and lipid-related traits (P-value < $1x10^{-6}$**, Supplemental Table 21**). Enrichment analyses using publicly-available epigenomic data (implemented in GREGOR enrichment analysis, **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 (e.g. H3K27Ac, H3K9ac, H3K4me1, H3K4me3; **Supplemental Table 22 and 23**). These analysis support the hypothesis that our GWAS captures multiple physiological mechanisms that contribute heritability to NAFLD. Finally, DEPICT-based predicted gene function nominated gene candidates for 28 genes, including the known genes *PNPLA3* and *ERLIN1* (FDR <5%, **Supplemental Table 24**), as well as well-known cardiometabolic disease genes (e.g., *PPARG***)**.

Coding variants in putative causal genes driving NAFLD associations.

There were six novel trans-ancestry loci for which the lead SNP itself is a coding missense variant (**Supplementary Table 25**): Thr1412Asn in *CPS1* (rs1047891, β=0.037, $P=2.8x10^{-8}$), Glu430Gln in *GPT* (rs141505249, β=-2.023, $P=9.0x10^{-62}$), Val112Phe in *TRIM5* (rs11601507, β=0.099, $P=1.5x10^{-14}$), Ala163Thr in *DNAJC22* (rs146774114, β=-0.157, $P=2.5x10^{-8}$), Glu366Lys in *SERPINA1* (rs28929474, β=0.492, $P=9.01x10^{-73}$) and Cys325Gly in *APOH* (rs1801689, β=0.17, $P=1.5x10^{-18}$). To identify additional coding variants that may drive the association between the lead SNPs and NAFLD risk, we investigated predicted loss of function (pLoF) and missense variants strongly linked to the identified NAFLD lead variants ($r^2 > 0.7$, **Supplemental Table 25-28**). Four previously

described missense variants were replicated in the current study, including Thr165Ala in *MARC1*, Ile292Val in *ERLIN1*, Glu167Lys in *TM6SF2* and Ile148Met in *PNPLA3*. Among novel loci, missense variants linked ($r^2 > 0.7$) with lead variants included the genes *CCDC18, MERTK, APOL3, PPARG, MTTP, MLXIPL, ABCB4, AKNA, GPAM, SH2B3, P2RX7, NYNRIN, ANPEP,IFI30* and *MPV17L2*. Among the trans-ancestry coding missense variants, ten (*CCDC18, MLXIPL, ABCB4, AKNA, DNAJC22, SERPINA1, ANPEP, APOH, IFI30, MPV17L2,* and *PNPLA3*) were predicted based on two methods (SIFT, PolyPhen-2) to have potentially deleterious and/or damaging effects in protein function[34,35]. An AA-specific locus on chromosome (rs115038698, chr7:87024718) was strongly linked to a nearby missense variant Ala934Thr in *ABCB4* (rs61730509, AFR r^2=0.92) with predicted deleterious effect, where the T-allele confers an increased risk of NAFLD (β=0.617, P=1.8x10^{-20}). In summary, 24 of our 77 trans-ancestry loci prioritized a candidate gene based on a missense variant in tight linkage with the lead SNP (**Supplemental Table 25**).

Additional approaches to nominate putative causal genes

We performed colocalization analyses with gene expression and splicing across 48 tissues measured by the GTEx project, and overlapped our lead SNPs with histone quantitative trait locus (QTL) data from livers to identify NAFLD-associated variants that are also associated with change in gene expression (eQTLs), splice isoforms (sQTLs), or histone modifications (hQTLs, **Methods**, **Supplemental Table 29**). Across all tissues, a total of 123 genes were prioritized with 20 genes in liver tissue (**Methods**). In liver tissue alone, a total of eight variant-gene pairs were identified where the allele associated with protection against NAFLD was also associated with reduced gene expression (i.e., the direction of effect was concordant between the GTEx eQTL and GWAS

sentinel variant): *AC091114.1, PANX1, FADS2*, *SHROOM3, U2SURP*, *NYNRIN*, *CD276* and *EFHD1*. Furthermore, sQTL analysis in GTEx v8 identified two genes in liver, *HSD17B13* and *ANPEP*, and 12 genes (*MARC1, HSD17B13*, *ABO*, *FADS1-FADS2*, *TMEM258*, *MLXIP*, *ANPEP, KAT7, STRADA*, *DDX42*, *TRC4AP*, and *APOL3*) that were affected in at least two tissues (**Supplemental Table 30**). Finally, two of our lead SNPs were in high LD ($r^2 > 0.8$) with variants that regulated H3K27ac levels in liver tissue (hQTLs), namely *EFHD1* (hQTL SNPs rs2140773, rs7604422 in *EFHD1*) and *FADS2* loci (hQTL rs174566 in *FADS2*)[36].

We next mapped our NAFLD loci to regions of open chromatin using ATAC-seq in three biologically-relevant liver-derived tissues (human liver, liver cancer cell line [HepG2], and hepatocyte-like cells [HLC] derived from pluripotent stem cells)[37]. 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) (**Supplemental Table 31**). These datasets are useful entry points for deciphering regulatory mechanisms involved in the pathophysiology of NAFLD. Most notably, the genes *DHODH*, *H2AFZ*, *PAQR9*, *FTO*, *MIR644A*, *BCL7B* and *KRT82* showed interactions with NAFLD loci that were also in open chromatin in both HLC and HepG2 cells.

Based on DEPICT gene prediction, coding variant linkage analysis, and QTL colocalization (**Supplemental Tables 24-31**), 215 potentially relevant genes for NAFLD were identified for the 77 loci. A protein-protein interaction (PPI) analysis revealed that among the 192 available proteins, 86 nodes were observed, with a PPI enrichment ($P < 9.0 \times 10^{-8}$) indicating that the network has substantially more interactions than expected by chance (**Supplemental Table 32 and Supplemental Figure 6**).

306 For each gene identified from all of the above described analyses, we counted the number
307 of times that the gene was identified for each of the analyses (DEPICT gene prediction, coding
308 variant linkage, QTL colocalization, promotor Capture-C and/or ATAC-Seq peak overlap, and PPI
309 network analysis) and divided this by the number of analyses (e.g., 8). We labeled this measure
310 as the gene nomination score, which reflects the cumulative evidence supporting the respective
311 gene as causal for the observed association. Based on our gene nomination scheme, we found
312 evidence for a single gene nomination at 52 genomic loci, two genes at 14 loci, and three genes
313 at six loci. Six loci had more than three genes nominated (one of which was HLA), and only two
314 loci lacked any data to support a nomination (**Supplemental Table 33**). We further prioritized
315 those loci which were prioritized by at least 3 sources of evidence (or 4 sources of evidence for
316 coding variants). This resulted in a total of 27 loci supported by multiple lines of evidence (**Table**
317 **1**), which included 6 loci with co-localizing eQTLs in liver or adipose tissues and connection to the
318 predicted gene via Promoter CaptureC data (i.e., *EPHA2, IL1RN, SHROOM3, HKDC1, PANX1,*
319 *DHODH;HP).*
320 Interestingly, 14 of the nominated genes are transcription factors (TF) (**Supplemental**
321 **Table 34**). Of particular interest, two of these TFs have several downstream target genes
322 identified using the DoRothEA data in OmniPath (**Methods**). Notably, the CEBPA TF targets the
323 downstream genes *PPARG*, *TRIB1*, *GPAM*, *FTO*, *IRS1*, *CRIM1*, *HP, TBC1D8*, and *CPS1*, but also
324 *NCEH1,* a gene in the vicinity of one of our associations that lacked a nominated candidate gene.
325 Similarly, *HNF1A*, the lead gene in EA scan (and corresponding to the trans-ancestry *P2RX7* locus)
326 targets *SLC2A2*, *MTTP*, and *APOH*.
327

328 *Polygenic Risk Score analyses.*

329 We calculated a candidate SNP polygenic risk score (PRS) based on Stage 1 350K dataset (primary
330 set) to perform a Phewas in an independent sample in MVP (Stage 2 replication set). We observed
331 that an increased NAFLD PRS was associated with abnormal results of function study of liver
332 (Bonferroni $P < 3.1 \times 10^{-5}$), and showed suggestive significance with bacterial pneumonia, otalgia,
333 gout and other crystal arthropathies and non-infectious gastroenteritis ($P < 0.001$, **Supplemental**
334 **Table 35**). Furthermore, a NAFLD PRS based on the Stage 1 set was positively associated with
335 NAFLD prediction in the Stage 2 replication set ($P=3.8 \times 10^{-5}$, **Supplemental Table 36**).

336

337 *Investigation of pleiotropy of lead NAFLD SNPs.*

338 We next sought to identify additional traits that were also associated with our 77 trans-ancestry
339 lead SNPs. First, we performed a LabWAS of distinct clinical laboratory test results[38] in MVP
340 (**Methods**), yielding 304 significant SNP-trait associations (**Supplemental Table 37, Supplemental**
341 **Figure 7**). Second, we performed a PheWAS Analysis in UK Biobank data using SAIGE (**Methods**),
342 which identified various SNP-trait associations that mapped to loci previously associated with
343 liver traits, cardiometabolic traits, as well as additional enriched association for gallstones, gout,
344 arthritis, and hernias (**Supplemental Tables 38 and 39**). In particular, we examined all
345 associations for PheCode 571.5, "Other chronic nonalcoholic liver disease" which comprised
346 1,664 cases and 400,055 controls. Of the n=73 variants found, we noted that 14/73 were both
347 nominally associated and directionally consistent with our scan (signed binomial test $P=3.4 \times 10^{-9}$),
348 providing additional validation for our scan (**Supplementary Table 40**). Third, we performed a
349 SNP lookup using the curated data in the IEU OpenGWAS project (**Supplemental Tables 41 and**

42), which identified 2,891 genome-wide significant SNP-trait associations for trans-ancestry SNPs, and additional 283 SNP-trait associations for the ancestry-specific lead SNPs. Finally, we performed cross-trait colocalization analyses using COLOC of EA, AA, and HISP lead loci with 36 other GWAS statistics of cardiometabolic and blood cell related traits (**Methods**). This resulted in significant regional colocalization for 64 SNP-trait pairs in EA, 32 SNP-trait pairs in AA, and 12 SNP trait pairs in HISP (**Supplemental Table 43-45**).

Based on the four analyses described above, we categorized relevant phenotypes observed as liver traits, metabolic traits, or inflammatory traits based on all significant SNP-trait associations and their nominated candidate genes (**Supplemental Tables 37-44, Figure 4**). Across the trans-ancestry lead variants (n=77), ancestry-specific (n=3), and secondary proximal associations (*HNF1A,* n=1), 22 SNPs showed association with only liver traits (such as ALT, ALP, AST, and GGT) (**Figure 4).** By contrast, 23 loci showed associations with both liver and cardiometabolic traits (such as HDL, LDL, and total cholesterol, triglycerides, BMI, glucose, and HbA1c) whereas 3 loci (*IL1RN*, *TMEM147*;*ATP4A* and *RORA*) showed association with both liver traits and inflammatory traits (e.g., C-reactive protein, white blood cell count, lymphocyte count). Finally, 25 loci showed association with all three traits: liver, cardiometabolic, and inflammation. Notably, among 12 loci that showed significant association with hepatic fat (color-coded in red in **Figure 4**), 11 were associated with both liver and metabolic traits, including five that were also associated with inflammatory traits. Collectively, our findings identify novel NAFLD-associated genetic loci with pleotropic effects that may impact hepatic, metabolic and inflammatory traits.

Discussion

In this study, the largest and most diverse GWAS of NAFLD to date, we report a total of 77 trans-ancestry (of which 70 are novel) and 3 additional ancestry-specific loci that show significant genome-wide association with NAFLD. While our NAFLD definition is a proxy for chronic hepatocellular injury in the absence of other known causes of liver disease, we further used CT/MRI imaging data to compare to what extent these SNPs also associated with hepatic fat accumulation. Overall, 24 (~30%) of these loci were nominally associated with hepatic fat based on CT or MRI, and the majority of these overlapping SNPs were associated with metabolic and/or inflammatory traits. Thus, SNPs that are associated with liver enzymes, metabolic risk factors, and inflammatory biomarkers may be the most likely to be associated with liver steatosis and should be prioritized for further follow-up. Furthermore, detailed genetic correlation analyses showed significant enrichment of these SNPs for cardiometabolic traits, metabolic pathways, and genomic annotations relevant for NAFLD. We found that most of our index NAFLD-associated SNPs were associated with metabolic and/or inflammatory traits - the most common being lipid-related, followed by glycemic traits, hypertension, and cardiovascular disease, as well as cholelithiasis (gallstones), cholecystitis, osteoarthritis, hypothyroidism, and thrombophlebitis. Collectively, our findings offer a comprehensive and refined view of the genetic contribution to NAFLD with potential clinical, pathogenic, and therapeutic relevance. Integration of these with extant phenotypic association data sets allowed us to further characterize the functional mechanisms through which our identified loci may mediate NAFLD risk.

Previous studies for liver enzyme levels, particularly serum ALT activity, have been performed [10,11,16]. While there is overlap in the discoveries made by studies of natural variation

394 in circulating levels of this biomarker, our cohort and approach to phenotyping make our results
395 and interpretation unique. First, the diversity of our cohort provides both additional power and
396 potential for discovery, as the bulk of studies to date have been performed in predominantly
397 European-ancestry cohorts. Second, our approach ascertains individuals with *chronic* elevation
398 of this enzyme, consistent with genuine chronic liver disease. At the same time, we excluded
399 individuals with known causes of liver disease outside of NAFLD via ICD code definition, which
400 served to further enrich for metabolic disorders in our cohort. We further excluded control
401 individuals who maybe have intermittent ALT elevation, focusing on a healthier, 'super-control'
402 subset of the population. The result is that our approach should have higher specificity to
403 ascertained risk alleles that predispose to metabolic-induced fatty liver disease. In contrast, a
404 standard-ALT scan would be powered to discover the full spectrum of causes of liver disease (and
405 perhaps many more loci), many of which will not be specific to NAFLD and may be due to other
406 causes. As we have shown in validation studies using quantitative measures of hepatic fat as well
407 as ICD-code definitions of NAFLD, our results are highly directionally concordant, demonstrating
408 the relevance of our proxy phenotype to liver disease and physiology. Genetic correlation analysis
409 demonstrated strong correlation with cardiometabolic traits and disease, again consistent with
410 the relevance of our trait relative to simple enzyme measures.

411 There are several aspects of our study that are worth highlighting. We demonstrate the
412 strength of trans-ancestry GWAS for the discovery and interrogation of NAFLD susceptibility loci,
413 discoveries made possible by the diversity and sample size of the Million Veteran Program cohort,
414 of which 25% of participants are of non-European ancestry. Utilizing this data allows us to narrow
415 down putatively causal variants through trans-ancestry fine-mapping and construction of

credible sets likely to harbor the likely culprit variant(s). Construction of credible sets using trans-ancestry data has been shown to facilitate fine-mapping by producing smaller credible sets compared to sets based on single ancestries[39], an effect we also observed at our loci. Moreover, we identified eight NAFLD-associated loci in AAs. In particular, the lead SNP at the *ABCB4* locus (rs115038698) was in high LD with the missense variant rs61730509 (Ala934Thr, AFR r^2=0.92) and segregated a very potent effect (OR=1.87, CI=1.64-2.14, P=1.8x10^{-20}). This variant is of low frequency in AA (MAF=1.2%) but virtually absent in EA and ASN. *ABCB4*, also known as multidrug resistance protein 3 (*MDR3*), is a compelling candidate gene, as it has been previously implicated in cholestasis, gallbladder disease, and adult biliary fibrosis/cirrhosis[40-42]. Finally, for a number of variant gene-pairs, the observed effect on NAFLD risk and the impact of gene expression in the liver was consistent with our understanding of the expected effect given what is known about gene function, suggesting possible relevance as therapeutic targets. Among those, genetic deletion of Pannexin 1 (encoded by *PANX1*) was reported to have a protective effect in mouse model of acute and chronic liver disease[43,44], and is consistent with the data we report here.

Twelve of our loci were associated with quantitative measures of hepatic fat after multiple-test correction. These included loci previously associated with NAFLD or all-cause cirrhosis (e.g., *PNPLA3, TM6SF2, TNKS [PPP1R3B]*, *KIAA0196* [*TRIB1],* and *MARC1*), but also included novel loci reported here (e.g., *GPAM, APOE;APOC1, MTTP, APOH, IFI30;MPV17L2, SCN2A;COBLL1,* and *PPARG*). In all cases except *TNKS [PPP1R3B]*, the directional effect on hepatic fat was consistent with cALT levels. A discordance between measures of hepatic fat based on radiological and histological evaluation has been noted[12] and may be explained by the role of the

437 *PPP1R3B*-encoded protein in promoting the accumulation of hepatic glycogen[45] which may
438 influence the contrast in hepatic images[46,47].

439 Through functional genomic and bioinformatics prioritization analyses beyond those
440 based on coding variants or eQTLs, we were able to nominate loci that have at least one
441 candidate gene nominations at 75 out of our 77 (97%) identified loci. We found that these genes
442 were often highly expressed in liver and have prior biological connections to liver physiology and
443 disease, making this list compelling for further interrogation. As an example, *GPAM*, tagged by
444 the missense variant rs2792751 (Ile43Val, EA r^2 = 0.99), encodes the mitochondrial glycerol-3-
445 phosphate-acyltransferase 1, a protein used in the mitochondria to convert saturated fatty acids
446 into glycerolipids. *GPAM* is highly expressed in liver[48,49] and associated with metabolic disease[50],
447 consistent with our pleiotropy analyses. Mouse knockouts of *GPAM* had reduced weight, lower
448 hepatic triacylglycerol content, and decreased plasma triacylglycerol[51]. Another example is *MTTP*
449 which is tagged by the missense variant rs3816873 (Ile128Thr, EA r^2=1.0) and encodes the
450 microsomal triglyceride transfer protein, which loads lipids onto assembling VLDL particles and
451 facilitate their secretion by hepatocytes. Liver-specific *MTTP* knockout mice have reduced VLDL
452 secretion and increased hepatic steatosis[52]. Lomitapide, a small molecule inhibitor of MTTP, is
453 approved as a treatment for lowering LDL cholesterol in homozygous familial
454 hypercholesterolemia, but increases liver lipid by inhibiting VLDL secretion[53]. *TRIM5* (Val112Phe)
455 is a member of the tripartite motif (TRIM) family with E3 ubiquitin ligase activity with a key role
456 in innate immune signaling and antiviral host defense[54], and *TRIM5* SNPs have been associated
457 with increased risk of liver fibrosis in HIV/HCV co-infected patients[55]. *APOH* (Cys325Gly) encodes
458 the apolipoprotein H which is exclusively expressed in liver tissue[48] and which is associated with

459 ALT, AST, triglycerides, LDL cholesterol and platelets in the MVP labWAS. Two coding variants
460 (strongly linked) in MerTK (Arg466Lys and Ile518Val, r^2=0.98) were associated with NAFLD; MerTK
461 signaling in hepatic macrophages was recently shown to mediate hepatic stellate cell activation
462 and promote hepatic fibrosis progression[56], and variants in *MERTK* were associated with liver
463 fibrosis progression in HCV-infected patients[57], raising the possibility for MerTK as a novel
464 therapeutic target against fibrosis[58]. We emphasize that functional studies of our nominated
465 causal genes are needed to demosntrate casual relevance, their impact on hepatosteatosis, and
466 ultimately to determine their underlying mechanisms.

467 Given the complex etiology and progression of NAFLD, we anticipated that our study
468 would identify novel loci with putatively causal genes that span multiple molecular pathways.
469 Indeed, our novel loci include genes that play roles in obesity (e.g., *FTO, PPARG*), insulin
470 resistance (e.g., *COBLL1*, *MIR5702 [IRS1]*), and diabetes (e.g., *HNF1A*). Relevant for hepatic
471 inflammation in the two-hit hypothesis of NAFLD[59], our novel loci also implicate immune-
472 mediated or inflammatory contributions to NAFLD progression, including *HLA, RORA*[60,61],
473 *IFI30*[62,63], *CD276*[64], *ILRN*[62,65], *ITCH*[66,67] and *P2RX7*[68-70]. Among these, *RORA* encodes the retinoic
474 acid receptor related orphan receptor A which may be involved in NASH pathogenesis through
475 macrophage polarization and miRNA122, which comprises 70% of the total miRNA in liver [60,61]. It
476 is also known that loss of TRIB1 substantially decreases miR-122 levels via its impact on HNF4 and
477 HNF1A[71]. *IFI30* encodes gamma-interferon-inducible lysosomal thiol reductase (GILT) which is
478 involved in antigen processing and presentation and the production of reactive oxygen species
479 during cellular stress and autophagy. Finally, *P2RX7* encodes the purinergic receptor P2X7 which
480 is involved in inflammasome activation and IL-1β processing in liver inflammation and fibrosis[68-

[70]. Encouragingly, these and additional pathways have emerged despite the proxy nature of our phenotype, and almost certainly underestimate the true number of loci contributing to NAFLD.

In conclusion, we define 77 trans-ancestry loci (70 novel) with 3 additional ancestry-specific loci associated with NAFLD by using chronic ALT elevation in a large, ancestrally diverse cohort enriched for metabolic disorders without other known causes of liver disease. The abundance of NAFLD loci identified by our analyses constitutes a much-needed large-scale, multi-ancestry genetic resource that can be used to build prediction models, identify causal mechanisms, and understand biological pathways contributing to NAFLD initiation and disease progression.

Methods

We performed a large-scale trans-ancestry NAFLD GWAS in the Million Veteran Program. We subsequently conducted analyses to facilitate the prioritization of these individual findings, including transcriptome-wide predicted gene expression for NAFLD, secondary signal analysis, coding variant mapping, phenome-wide association analyses in various public data sources, and various forms of cardiometabolic cross-trait colocalization analyses to fine-map the genomic loci to putatively causal genes.

Discovery cohort.

The Million Veteran Program (MVP) is a large cohort of fully consented veterans of the United States military forces recruited from 63 participating Department of Veterans Affairs (VA) medical facilities[28]. Recruitment for this ongoing sample started in 2011, and all veterans are eligible to

participate. This study analyzed clinical data through July 2017 for participants who were enrolled since January 2011. All MVP study participants provided blood samples for DNA extraction and genotyping, completed surveys about their health, lifestyle, and military experiences. Consent to participate and permission to re-contact was provided after veterans received information materials by mail and met with research staff to address their questions. Study participation also includes access to the participant's electronic health records for research purposes. Each veteran's electronic health care record is integrated into the MVP biorepository, including inpatient International Classification of Diseases (ICD-9-CM and ICD-10-CM) diagnosis codes, Current Procedural Terminology (CPT) procedure codes, clinical laboratory measurements, and reports of diagnostic imaging modalities. Researchers are provided with de-identified data, and have neither the ability nor authorization to link these details with a participant's identity. Blood samples are collected by phlebotomists and banked at the VA Central Biorepository in Boston, where DNA is extracted and shipped to two external centers for genotyping. 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.

Genotyping*:* DNA extracted from buffy coat was genotyped using a custom Affymetrix Axiom biobank array. The MVP 1.0 genotyping array contains a total of 723,305 SNPs, enriched for 1) low frequency variants in AA and HISP populations, and 2) variants associated with diseases common to the VA population [28].

Genotype quality-control: The MVP genomics working group applied standard quality control and genotype calling algorithms to the data in three batches using the Affymetrix Power Tools Suite (v1.18). Excluded were duplicate samples, samples with more heterozygosity than expected, and

samples with an over 2.5% missing genotype calls. We excluded related individuals (halfway between second- and third-degree relatives or closer) with KING software[72]. Before imputation, variants that were poorly called or that deviated from their expected allele frequency based on reference data from the 1000 Genomes Project were excluded[73]. After prephasing using EAGLE v2, genotypes were imputed via Minimac4 software[74] from the 1000 Genomes Project phase 3, version 5 reference panel. The top 30 principal components (PCs) were computed using FlashPCA in all MVP participants and an additional 2,504 individuals from 1000 Genomes. These PCs were used to unify of self-reported race/ancestry and genetically inferred ancestry to compose ancestral groups[29].

Phenotype classification: MVP NAFLD phenotype definitions were developed by combining a previously published VA CDW ALT-based approach with non-invasive clinical parameters available to practicing clinicians at the point of care. The primary NAFLD phenotype (labeled "ALT-threshold") was defined by: (i) elevated ALT >40 U/L for men and >30 U/L for women during at least two time points at least 6 months apart within a two-year window period at any point prior to enrollment and (ii) exclusion of other causes of liver disease (e.g. presence of chronic viral hepatitis B or C [defined as positive hepatitis C RNA > 0 international units/mL or positive hepatitis B surface antigen], chronic liver diseases or systemic conditions [e.g. hemochromatosis, primary biliary cholangitis, primary sclerosing cholangitis, autoimmune hepatitis, alpha-1-antitrypsin deficiency, sarcoidosis, metastatic liver cancer, secondary biliary cirrhosis, Wilson's disease], and/or alcohol use disorder [e.g. alcohol use disorder, alcoholic liver disease, alcoholic hepatitis and/or ascites, alcoholic fibrosis and sclerosis of liver, alcoholic cirrhosis of liver and/or ascites, alcoholic hepatic failure and/or coma, and unspecified alcoholic liver disease). The

548 control group was defined by having a: normal ALT (≤30 U/L for men, ≤20 U/L for women) and
549 no apparent causes of liver disease or alcohol use disorder or related conditions[21]. Habitual
550 alcohol consumption was assessed with the age-adjusted Alcohol Use Disorders Identification
551 Test (AUDIT-C) score, a validated questionnaire annually administered by VA primary care
552 practitioners and used previously in MVP[75,76].
553
554 *Single-variant autosomal analyses.*
555 We tested imputed SNPs that passed quality control (i.e. HWE > $1x10^{-10}$, INFO > 0.3, call rate >
556 0.975) for association with NAFLD through logistic regression assuming an additive model of
557 variants with MAF > 0.1% in European American (EA), and MAF > 1% in African Americans (AA),
558 Hispanics (HISP), and Asians (ASN) using PLINK2a software[77]. Covariates included age, gender,
559 age-adjusted AUDIT-C score, and 10 principal components of genetic ancestry. We aggregated
560 association summary statistics from the ancestry-specific analyses and performed a trans-
561 ancestry meta-analysis. The association summary statistics for each analysis were meta-analyzed
562 in a fixed-effects model using METAL with inverse-variance weighting of log odds ratios[78].
563 Variants were clumped using a range of 500kb and/or CEU r^2 LD > 0.05, and were considered
564 genome-wide significant if they passed the conventional p-value threshold of $5.0x10^{-8}$.
565
566 *Secondary signal analysis.*
567 GCTA[79] was used to conduct conditional analyses to detect ancestry-specific distinct association
568 signals at each of the lead SNPs utilizing the GWAS summary statistics in EA, AA, and HISP; these
569 ancestry-stratified MVP cohorts were used to model LD patterns between variants. The reference

570 panel of genotypes consisted of the variants with allele frequencies > 0.1% in EA, >1% in AA, and
571 >1% in HISP that passed quality control criteria in the MVP-specific NAFLD GWAS (INFO > 0.3,
572 HWE P > $1.0x10^{-10}$, call rate > 0.975). For each lead SNP, conditionally independent variants that
573 reached locus-wide significance (P < $1.0x10^{-5}$) were considered secondary signals of distinct
574 association. If the minimum distance between any distinct signals from two separate loci was less
575 than 500kb, we performed an additional conditional analysis that included both regions and
576 reassessed the independence of each signal.
577
578 *Credible Sets.*
579 We calculated Wakefield's approximate Bayes' factors [30] based on the marginal summary
580 statistics of the trans-ancestry meta-analysis and ancestry specific summary statistics using the
581 CRAN R package corrcoverage[80]. For each locus, the posterior probabilities of each variant being
582 causal were calculated and a 95% credible set was generated which contains the minimum set of
583 variants that jointly have at least 95% posterior probability (PP) of including the causal variant.
584
585 *Concordance of NAFLD with qHF.*
586 For 77 lead trans-ancestry SNPs a concordance analysis was performed to evaluate the extent to
587 which genetic predictors of hepatocellular injury (cALT) correspond with quantitative hepatic fat
588 derived from computed tomography (CT) / magnetic resonance imaging (MRI)-measured hepatic
589 fat in the Penn Medicine Biobank (PMBB), UK Biobank, Multi-Ethnic Study of Atherosclerosis
590 (MESA), Framingham Heart Study (FHS), and University of Maryland Older Order Amish study.
591 Attenuation was measured in Hounsfield units. The difference between the spleen and liver

attenuation was measured for PMBB; a ratio between liver attenuation/spleen attenuation was used for MESA and Amish; and liver attenuation/phantom attenuation ratio in FHS as previously described by Speliotes *et al*[12]. Abdominal MRI data from UK Biobank data were used to quantify liver fat using a two-stage machine learning approach with deep convolutional neural networks[81]. CT-measured hepatic fat was estimated using a multi-stage series of neural networks for presence of scan contrast and liver segmentation using convolutional neural networks. The PMBB included CT data on 2,979 EA and 1,250 AA participants[82], the FHS included a total of 3,011 EA participants, the Amish study 754 EA participants, and MESA contributed 1,525 EA, 1,048 AA, 923 HISP, and 360 ASN participants for concordance analysis. The UK Biobank included MRI image data from 36,703 EA participants. All cohorts underwent individual-level linear regression analysis on hepatic fat, adjusted for the covariates of age, gender, first 10 principal components of genetic ancestry, and alcohol intake if available. If the lead SNP was not available in any of the studies, a proxy SNP in high LD with the lead variant was used ($r^2 > 0.7$) or if no such variant was identified, the SNP was set to missing for that respective study. The study-specific ancestry-stratified summary statistics 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[78]. In a first round of meta-analysis, ancestry-specific summary statistics were generated, which then served as input for a subsequent round of meta-analysis that represents the trans-ancestry effects of our lead SNPs on quantitative hepatic fat.

Heritability estimates and genetic correlations analysis.

LD-score regression was used to estimate the heritability coefficient, and subsequently population and sample prevalence estimates were applied to estimate heritability on the liability scale[83]. A genome-wide genetic correlation analysis was performed to investigate possible co-regulation or a shared genetic basis between T2D and other complex traits and diseases. Pairwise genetic correlation coefficients were estimated between the meta-analyzed NAFLD 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 NAFLD heritability.

We analyzed cell type-specific annotations to identify enrichments of NAFLD heritability. First, a baseline gene model was generated consisting of 53 functional categories, including UCSC gene models, ENCODE functional annotations[84], Roadmap epigenomic annotations[85], and FANTOM5 enhancers[86]. Gene expression and chromatin data were also analyzed to identify disease-relevant tissues, cell types, and tissue-specific epigenetic annotations. We used LDSC[31-33] to test for enriched heritability in regions surrounding genes with the highest tissue-specific expression. Sources of data that were analyzed included 53 human tissue or cell type RNA-seq data from GTEx[27]; human, mouse, or rat tissue or cell type array data from the Franke lab[87]; 3 sets of mouse brain cell type array data from Cahoy *et al*[88]; 292 mouse immune cell type array data from ImmGen[89]; and 396 human epigenetic annotations from the Roadmap Epigenomics Consortium [85].

635

636 *Pathway Annotation enrichment.*

637 Enrichment analyses in DEPICT[90] were conducted using genome-wide significant ($P < 5x10^{-8}$)
638 NAFLD GWAS lead SNPs. DEPICT is based on predefined phenotypic gene sets from multiple
639 databases and Affymetrix HGU133a2.0 expression microarray data from >37k subjects to build
640 highly-expressed gene sets for Medical Subject Heading (MeSH) tissue and cell type annotations.
641 Output includes a P-value for enrichment and a yes/no indicator of whether the FDR q-value is
642 significant ($P < 0.05$). Tissue and gene-set enrichment features are considered. We tested for
643 epigenomic enrichment of genetic variants using GREGOR software[91]. We selected EA-specific
644 NAFLD lead variants with a p-value less than $5x10^{-8}$. We tested for enrichment of the resulting
645 GWAS lead variants or their LD proxies (r^2 threshold of 0.8 within 1 Mb of the GWAS lead, 1000
646 Genomes Phase I) in genomic features including ENCODE, Epigenome Roadmap, and manually
647 curated data (**Supplemental Table 24**). Enrichment was considered significant if the enrichment
648 p-value was less than the Bonferroni-corrected threshold of $P=1.8x10^{-5}$ (0.05/2,725 tested
649 features).

650

651 *Coding variant mapping.*

652 All imputed variants in MVP were evaluated with Ensemble variant effect predictor[92], and all
653 predicted LoF and missense variants were extracted. The LD was calculated with established
654 variants for trans-ancestry, EA, AA, and HISP lead SNPs based on 1000 Genomes reference
655 panel[73]. For SNPs with low allele frequencies, the MVP dataset was used for LD calculation for
656 the respective underlying population. For the trans-ancestry coding variant, 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-value < $1x10^{-5}$) were considered the putative causal drivers of the observed association at the respective locus.

Colocalization with gene expression

GWAS summary statistics were lifted over from GRCh37 to GRCh38 using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Colocalization analysis was run separately for each of the 49 tissues in GTEx v8[27]. For each tissue, we obtained an LD block for the genome with a sentinel SNP at P < $5x10^{-8}$, and then restricted analysis to the LD blocks. For each LD block with a sentinel SNP, all genes within 1Mb 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 1Mb 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 pass 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 trans-ancestry meta-analysis, we looked for overlaps between any variant in the credible set, and regions of open chromatin previously identified using ATAC-Seq experiments in two cell types—3 biological replicates of HepG2[93] and 3 biological replicates of hepatocyte-like cells (HLC)[94] produced by differentiating three biological

679 replicates of iPSCs, which in turn were generated from peripheral blood mononuclear cells using
680 a previously published protocol[36].
681
682 *Overlap with Promoter Capture-C data.*
683 We used two promoter Capture-C datasets from two cell/tissue types to capture physical
684 interactions between gene promoters and their regulatory elements and genes; three biological
685 replicates of HepG2 liver carcinoma cells, and hepatocyte-like cells (HLC)[93]. The detailed protocol
686 to prepare HepG2 or HLC cells for the promoter Capture-C experiment is previously described[36].
687 Briefly, for each dataset, 10 million cells were used for promoter Capture-C library generation.
688 Custom capture baits were designed using an Agilent SureSelect library design targeting both
689 ends of DpnII restriction fragments encompassing promoters (including alternative promoters)
690 of all human coding genes, noncoding RNA, antisense RNA, snRNA, miRNA, snoRNA, and lincRNA
691 transcripts, totaling 36,691 RNA baited fragments. Each library was then sequenced on an
692 Illumina NovoSeq (HLC), or Illumina HiSeq 4000 (HLC), generating 1.6 billion read pairs per sample
693 (50 base pair read length.) HiCUP[95] was used to process the raw FastQ files into loop calls; we
694 then used CHiCAGO[96] to define significant looping interactions; a default score of 5 was defined
695 as significant. We identified those NAFLD loci at which at least one variant in the credible set
696 interacted with an annotated bait in the Capture-C data.
697
698 *Protein-Protein Interaction Network Analysis*
699 We employed the search tool for retrieval of interacting genes (STRING) v11[97] (https://string-
700 db.org) 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 were applied to construct the PPI networks, in which the nodes correspond to the proteins and the edges represent the interactions (**Figure 4, Supplemental Table 32**).

Gene Nomination.

Based on DEPICT gene prediction, coding variant linkage analysis, QTL analysis, and annotation enrichment, and PPI networks **(Supplemental Tables 24-33**), a total of 215 potentially relevant genes for NAFLD were mapped to trans-ancestry 77 loci. For each locus with multiple mapped genes, we counted how many times each gene was identified through each of the analysis, and divided this by the total number of experiments (i.e., 8) to calculate an evidence burden that ranges from 0 to 100%. For each genomic locus, the gene that was most frequently identified as potentially relevant was selected as the putative causal gene. In the case of a tie break, and if the respective genes have identical nomination profiles, the gene with more eQTLs 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 3 pieces of evidence or greater are listed for coding variant (**Table 1A**) and non-coding variants (**Table 1B**), respectively.

MVP LabWAS.

A total of 21 continuous traits in the discovery MVP dataset, e.g. AST, ALP, fasting TG, HDL, LDL, TC, random glucose, HbA1c, albumin, bilirubin, platelet count, BMI, blood urea nitrogen (BUN),

723 creatinine, eGFR, SBP, DBP, ESR, INR, and C-reactive protein were tested in 186,681 EA's with
724 association of 77 SNPs using linear regression of log-linear values. Covariates included age,
725 gender and first 10 principal components of EA ancestry.
726
727 *PheWAS with UK Biobank data.*
728 For the 77 lead trans-ancestry SNPs and EA and AA specific SNPs, we performed a PheWAS in a
729 genome-wide association study of EHR-derived ICD billing codes from the White British
730 participants of the UK Biobank using PheWeb[98]. In short, phenotypes were classified into 1,403
731 PheWAS codes excluding SNP-PheWAS code association pairs with case counts less than fifty[99].
732 All individuals were imputed using the Haplotype Reference Consortium panel [100], resulting in
733 the availability of 28 million genetic variants for a total of 408,961 subjects. Analyses on binary
734 outcomes were conducted using a model named SAIGE, adjusted for genetic relatedness, gender,
735 year of birth and the first 4 principal components of white British genetic ancestry[101]. SAIGE
736 stands for Scalable and Accurate Implementation of GEneralized mixed model and represents a
737 generalized mixed-model association test that accounts for case-control imbalance and sample
738 relatedness[101].
739
740 *IEU OpenGWAS project SNP lookup.*
741 An additional phenome-wide lookup was performed for 77 lead trans-ancestry SNPs and EA and
742 AA specific SNPs in Bristol University's MRC Integrative Epidemiology Unit (IEU) GWAS
743 database[102]. This database consists of 126,114,500,026 genetic associations from 34,494 GWAS
744 summary datasets, including UK Biobank (http://www.nealelab.is/uk-biobank), FinnGen

(https://github.com/FINNGEN/pheweb), Biobank Japan (http://jenger.riken.jp/result), NHGRI-EBI GWAS catalog (https://www.ebi.ac.uk/gwas), blood metabolites GWAS[103], circulating metabolites GWAS[104], the MR-Base manually curated database[105], and protein level GWAS[106].

Regional cardiometabolic cross-trait colocalization.

Bayesian colocalization tests between NAFLD-associated signals and the following trait- and disease-associated signals were performed using the COLOC R package[107]. To enable cross-trait associations, we compiled summary statistics of 36 cardiometabolic and blood cell-related quantitative traits and disease from GWAS studies conducted in EA ancestry individuals, and for MVP-based reports also on AA and HISP. To summarize, for total, HDL, and LDL cholesterol, triglycerides, alcohol use disorder, alcohol intake, systolic blood pressure, diastolic blood pressure, type 2 diabetes, BMI, CAD, we used the summary statistics available from various MVP-based studies[26,75,108]. Of these, the summary statistics for CAD and BMI GWAS have not been published or deposited as of yet. Data on WHR were derived from GIANT Consortium[109], whereas summary statistics on CKD, gout, blood urea nitrogen, urate, urinary albumin-to-creatinine ratio, microalbuminuria, and eGFR were derived from CKD Genetics Consortium[110-112]. Finally, summary statistics of blood cell traits (e.g. 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[113]. A colocalization test was performed for all 77 NAFLD loci spanning 500kb region around the lead SNP for all 36

767 compiled traits. COLOC requires for each SNP data on allele frequency, sample size, beta-
768 coefficients and variance or p values. For each association pair COLOC was run with default
769 parameters and priors. COLOC computed posterior probabilities for the following five
770 hypotheses: PP0, no association with trait 1 (NAFLD GWAS signal) or trait 2 (e.g., co-associated
771 metabolic signal); PP1, association with trait 1 only (i.e., no association with trait 2); PP2,
772 association with trait 2 only (i.e., no association with trait 1); PP3, association with trait 1 and
773 trait 2 by two independent signals; and PP4, association with trait 1 and trait 2 by shared variants.
774 Evidence of colocalization[114] was defined by PP3 + PP4 ≥ 0.99 and PP4/PP3 ≥ 5.
775
776 *NAFLD Polygenic risk score and NAFLD risk.*
777 We constructed polygenic risk score (PRS) for NAFLD in the Stage 2 replication data set containing
778 of 73,580 MVP participants of EA ancestry by calculating a linear combination of weights derived
779 from the discovery MVP dataset of lead 77 trans-ancestry variants. The PRS was divided into
780 quintiles and the risk of NAFLD was assessed using a logistic regression model using the lowest
781 decile as a reference (e.g. the 20% of participants with lowest of NAFLD PRS), together with the
782 potential confounding factors of age, gender, age-adjusted AUDIT-C, and the first 10 principal
783 components of EA ancestry.
784
785 *NAFLD PRS Phewas*
786 For the NAFLD PRS that was generated using the Stage 1 350K dataset, we performed a PheWAS
787 study in the Stage 2 108K replication dataset to fully leverage full catalog of available ICD-9/ICD-
788 10 diagnosis codes. Of genotyped veterans, participants were included in the PheWAS analysis if

their respective electronic health record reflected two or more separate encounters in the VA Healthcare System in MVP up to July 2017. Using this method, a total of 73,580 veterans of EA ancestry were available for PheWAS analysis. ICD-9/ICD-10 diagnosis codes were collapsed to clinical disease groups and corresponding controls using predefined groupings [99]. Phenotypes were required to have a case count over 25 in order to be included in the PheWAS analysis, and a multiple testing thresholds for statistical significance was set to $P < 2.8x10^{-5}$ (Bonferroni method). The NAFLD PRS was used as a continuous exposure variable in a logistic regression adjusting for age, sex, age-adjusted AUDIT-C, and 10 principal components in an additive effects model using the PheWAS R package in R v3.2.065. The results from these analyses are reported as odds ratios, in which the estimate is the average change in odds of the PheWAS trait per NAFLD-increasing polygenic risk score.

Transcription Factor Analysis.

We identified nominated genes (**Supplemental Table 34**) that encode for TFs based on known motifs, inferred motifs from similar proteins, and likely sequence specific TFs according to literature or domain structure[115]. Target genes for these TFs were extracted using DoRothEA database[116] in OmniPath collection[117] using the associated Bioconductor R package OmnipathR[118], a gene set resource containing TF-TF target interactions curated from public literature resources, such as ChIP-seq peaks, TF binding site motifs and interactions inferred directly from gene expression.

Acknowledgements

This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration and was supported by award no. 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 I01BX003341 (H.R.K. Co-Principal Investigator) and the VA Informatics and Computing Infrastructure (VINCI) VA HSR RES 130457 (S.L.D). B.F.V. acknowledges support for this work from the NIH/NIDDK (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 Veterans Administration [IK2 CX001780]. Funding support is also acknowledged for MS (K23 DK115897), R.M.C (R01 AA026302), D.K. (National Heart, Lung, and Blood Institute of the National Institutes of Health [T32 HL007734]), L.S.P. (VA awards I01 CX001025, and I01 CX001737, NIH awards R21 DK099716, U01 DK091958, U01 DK098246, P30 DK111024, and R03 AI133172, and a Cystic Fibrosis Foundation award PHILLI12A0). The Rader lab was supported by NIH grants HL134853 (NJH and DJR) 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 ITMAT (NIH NCATS UL1TR001878), the Penn Center for Precision Medicine Accelerator Fund and R01 HL137501. Data for external replication and hepatic fat concordance were provided by investigators using United Kingdom BioBank, Multi-Ethnic Study of Atherosclerosis (MESA), Old Order Amish Study (Amish), Framingham Heart Study (FHS) and Penn Medicine Biobank (PMBB).

MESA/MESA SHARe Acknowledgements: MESA and the MESA SHARe projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. This research was supported by R01 HL071739 and MESA was supported by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

Ethics statement

The Central Veterans Affairs Institutional Review Board (IRB) and site-specific Research and Development Committees approved the Million Veteran Program study. All other cohorts participating in this meta-analysis have ethical approval from their local institutions. All relevant ethical regulations were followed.

Data availability

The full summary level association data from the trans-ancestry, European, African American, Hispanic, and Asian meta-analysis from this report will be available through dbGAP (Accession codes will be available before publication).

856 **Disclosures**

857 H.R.K. is a member of a Dicerna scientific advisory board; a member of the American Society of
858 Clinical Psychopharmacology's Alcohol Clinical Trials Initiative, which during the past three
859 years was supported by Alkermes, Amygdala Neurosciences, Arbor Pharmaceuticals, Dicerna,
860 Ethypharm, Indivior, Lundbeck, Mitsubishi, and Otsuka; and is named as an inventor on PCT
861 patent application #15/878,640 entitled: "Genotype-guided dosing of opioid agonists," filed
862 January 24, 2018.

863 **Legends**

864

865 **Table 1a. Gene nominations at loci with strongest evidence for coding variants.**

SNP	Position	Gene	AA-Change	SIFT/PP2*	e/sQTL**	Other †	Pleiotropy‡
rs6541349	1:93787867	CCDC18	p.Leu1134Val	+/-	+	.	M
rs2642438	1:220970028	MARC1	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

866 Genes nominated with various sources of evidence are listed as follows.
867 *Prio to the slash symbol: '+' indicates 'deleterious' in SIFT and '-' otherwise. After slash symbol: '+'
868 denotes probably damaging in Polyphen-2 and '-' otherwise.
869 ** The '+' indicates colocalization between NAFLD GWAS variant and GTEx QTL varint (COLOC
870 PP4/(PP3+PP4) > 0.9). (L) denotes QTL effect in Liver, (A) denotes QTL in Adipose.
871 †Each '+' represent evidence from DEPICT, PPI data, or if the lead SNP is within the transcript; coding
872 variants also include '+' from hQTLs/Capture-C evidence.
873 ‡Pleiotropy is limited to association with Metabolic (M) or Inflammatory (I) Traits
874

875

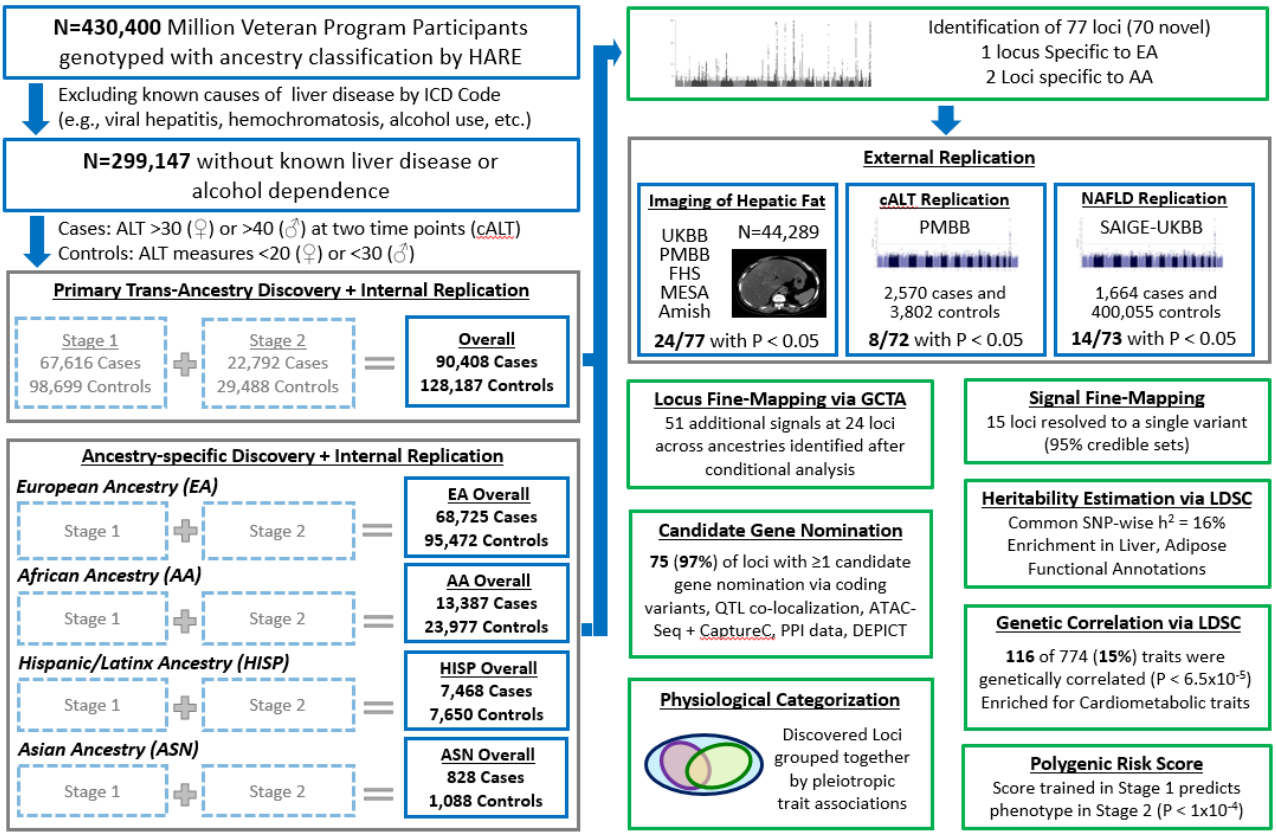
876 **Table 1b. Gene nominations at loci with strongest evidence for non-coding variants.**

SNP	Position	Gene	hQTL	CaptureC	e/sQTL**	Other †	Pleiotropy‡
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

877 Genes nominated with various sources of evidence are listed as follows.
878 *Prio to the slash symbol: '+' indicates 'deleterious' in SIFT and '-' otherwise. After slash symbol: '+'
879 denotes probably damaging in Polyphen-2 and '-' otherwise.
880 ** The '+' indicates colocalization between NAFLD GWAS variant and GTEx QTL varint (COLOC
881 PP4/(PP3+PP4) > 0.9). (L) denotes QTL effect in Liver, (A) denotes QTL in Adipose.
882 †Each '+' represent evidence from DEPICT, PPI data, or if the lead SNP is within the transcript; coding
883 variants also include '+' from hQTLs/Capture-C evidence.
884 ‡Pleiotropy is limited to association with Metabolic (M) or Inflammatory (I) Traits
885

886

887 **Figure 1. Overview of analysis pipeline.**

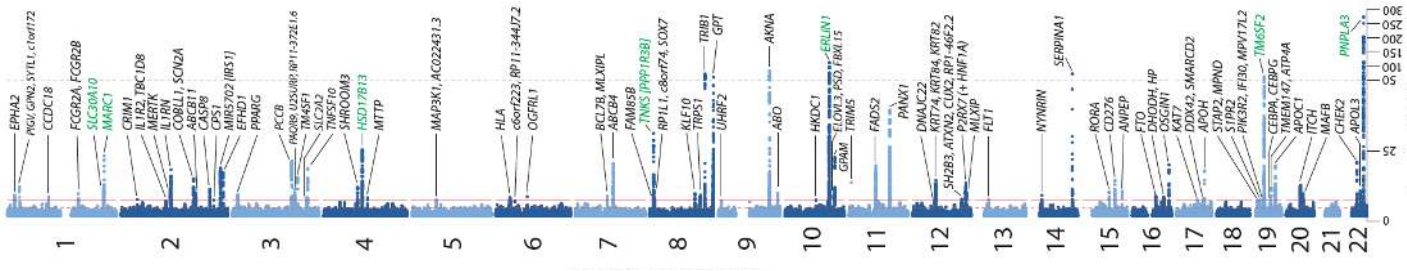


888

889 Left side of the flow diagram shows our study design with initial inclusion of 430,000 Million Veteran Program
890 participants with genotyping and ancestry classification by HARE, exclusion of individuals with known liver disease
891 or alcohol dependence and inclusion of subjects based on chronic ALT elevation (case) or normal ALT (control). This
892 resulted in 90,408 NAFLD cases and 128,187 controls with EA, AA, HISP and ASN ancestries that were examined in
893 primary trans-ancestry and ancestry-specific genome-wide association scans in discovery (stage 1) and internal
894 replication stages (stage 2) with further meta-analysis. Right side of the flow diagram highlights our results of trans-
895 ancestry and ancestry-specific meta-analyses identifying 77 trans-ancestry loci + 1 EA-specific + 2 AA-specific loci
896 that met genome-wide significance, with additional results of external replications, locus fine-mapping via GCTA,
897 signal fine-mapping (95% credible sets), heritability estimation and genetic correlations by LDSC, physiological
898 categorization of discovered loci based on pleotropic trait associations (mainly liver, metabolic and inflammation),
899 candidate gene nomination and polygenic risk score.

900

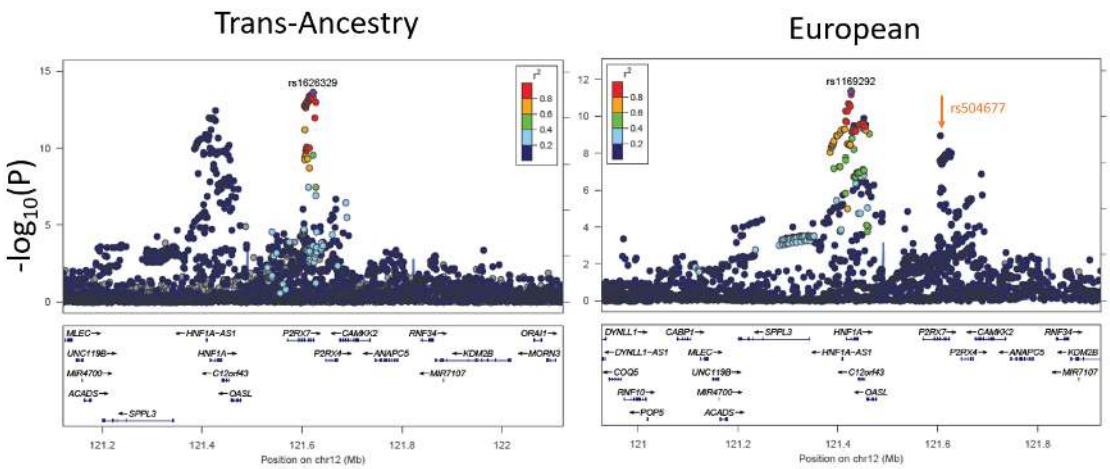
Figure 2. Manhattan plot of NAFLD GWAS of 90,408 NAFLD and 128, 187 controls in trans-ancestry meta-analysis.



Nominated genes are indicated for 77 loci reaching genome-wide significance ($P<5x10^{-8}$). Previously reported NAFLD-loci with genome-wide significant association are indicated in green font.

909 **Figure 3. Chromosome 12 locus points to different genes in trans-ancestry (left) and European-**

910 **only (right) analyses.**



911

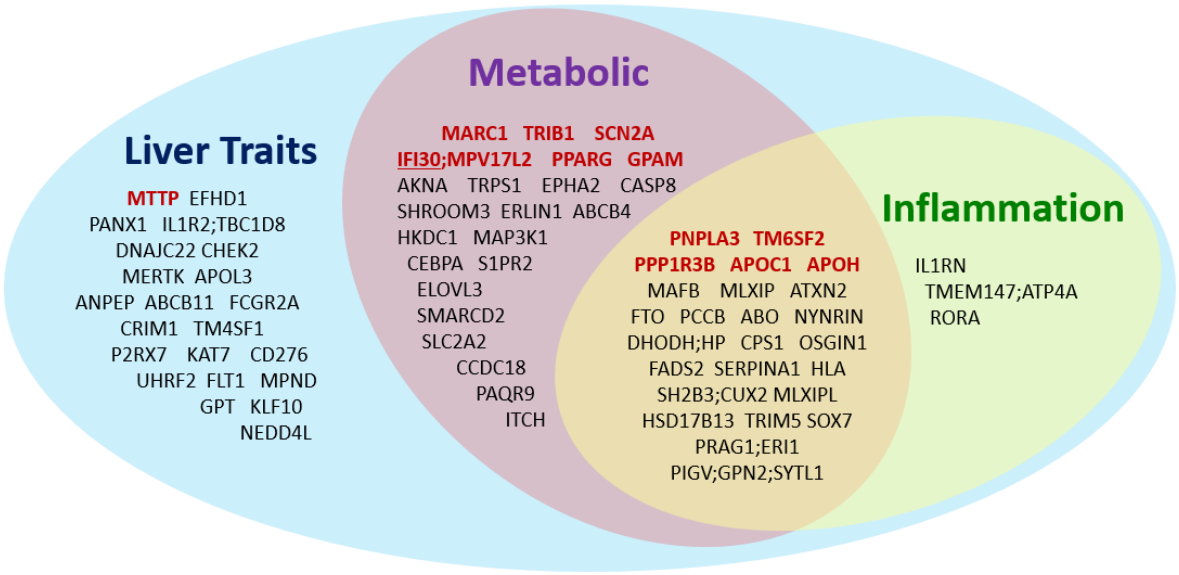
912 The lead variants in each analysis are highlighted. The orange arrow refers to the proxy SNP of

913 rs1626329 in the European-only analysis.

914

915 **Figure 4. Venn diagram depicting overlapping liver, metabolic and inflammatory traits among**

916 **NAFLD-associated loci.**



917

918 Overlapping liver (light blue), metabolic (pink) and/or inflammatory (green) traits are shown in
919 association with 77 trans-ancestry and additional ancestry-specific lead SNPs. The trait
920 categorizations reflect significant SNP-trait associations identified by: 1) LabWAS of clinical
921 laboratory results in MVP; 2) PheWAS with UKBB data using SAIGE; 3) SNP lookup using the
922 curated data in the IEU OpenGWAS projects; and 4) cross-trait colocalization analyses using
923 COLOC of EA, AA and HISP lead loci with 36 other GWAS statistics of cardiometabolic and blood
924 cell related traits. **Red/bold** font refers to the loci also associated with quantitative hepatic fat
925 on imaging analyses.

926

927

References

1. Asrani, S.K., Devarbhavi, H., Eaton, J. & Kamath, P.S. Burden of liver diseases in the world. *J Hepatol* **70**, 151-171 (2019).
2. Younossi, Z., Anstee, Q.M. & Marietti, M. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. *Nat Rev Gastroenterol Hepatol* **15**(2018).
3. Carr, R.M., Oranu, A. & Khungar, V. Nonalcoholic Fatty Liver Disease: Pathophysiology and Management. *Gastroenterol Clin North Am* **45**, 639-652 (2016).
4. 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).
5. Friedman, S.L., Neuschwander-Tetri, B.A., Rinella, M. & Sanyal, A.J. Mechanisms of NAFLD development and therapeutic strategies. *Nat Med* **24**, 908-922 (2018).
6. Estes, C., Razavi, H., Loomba, R., Younossi, Z. & Sanyal, A.J. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. *Hepatology* **67**, 123-133 (2018).
7. Jarvis, H. *et al.* Metabolic risk factors and incident advanced liver disease in non-alcoholic fatty liver disease (NAFLD): A systematic review and meta-analysis of population-based observational studies. *PLoS Med* **17**, e1003100 (2020).
8. Sookoian, S. & Pirola, C.J. Genetic predisposition in nonalcoholic fatty liver disease. *Clin Mol Hepatol* **23**, 1-12 (2017).
9. Romeo, S. *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **40**, 1461-5 (2008).
10. Yuan, X. *et al.* Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am J Hum Genet* **83**, 520-8 (2008).
11. Chambers, J.C. *et al.* Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet* **43**, 1131-8 (2011).
12. Speliotes, E.K. *et al.* Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet* **7**, e1001324 (2011).
13. Feitosa, M.F. *et al.* The ERLIN1-CHUK-CWF19L1 gene cluster influences liver fat deposition and hepatic inflammation in the NHLBI Family Heart Study. *Atherosclerosis* **228**, 175-80 (2013).
14. Kozlitina, J. *et al.* Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **46**, 352-6 (2014).
15. Liu, Y.L. *et al.* TM6SF2 rs58542926 influences hepatic fibrosis progression in patients with non-alcoholic fatty liver disease. *Nat Commun* **5**, 4309 (2014).
16. Abul-Husn, N.S. *et al.* A Protein-Truncating HSD17B13 Variant and Protection from Chronic Liver Disease. *N Engl J Med* **378**, 1096-1106 (2018).
17. Young, K.A. *et al.* Genome-Wide Association Study Identifies Loci for Liver Enzyme Concentrations in Mexican Americans: The GUARDIAN Consortium. *Obesity (Silver Spring)* **27**, 1331-1337 (2019).
18. 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).

19. Emdin, C.A. *et al.* A missense variant in Mitochondrial Amidoxime Reducing Component 1 gene and protection against liver disease. *PLoS Genet* **16**, e1008629 (2020).
20. Anstee, Q.M. *et al.* Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort(). *J Hepatol* **73**, 505-515 (2020).
21. Serper, M. *et al.* Validating a Non-Invasive Non-Alcoholic Fatty Liver Phenotype in the Million Veteran Program. *PLoS One* **(in press)**(2020).
22. Chalasani, N. *et al.* Genome-wide association study identifies variants associated with histologic features of nonalcoholic Fatty liver disease. *Gastroenterology* **139**, 1567-76, 1576 e1-6 (2010).
23. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
24. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649 (2018).
25. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet* **52**, 680-691 (2020).
26. 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).
27. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
28. Gaziano, J.M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214-23 (2016).
29. Fang, H. *et al.* Harmonizing genetic ancestry and self-identified race/ethnicity in genome-wide association studies. *Am J Hum Gen* **105**, 763-772 (2019).
30. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol* **33**, 79-86 (2009).
31. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
32. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
33. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).
34. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20 (2013).
35. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-4 (2003).
36. 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).
37. 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**, 581-9 (2015).

38. Goldstein, J.A. *et al.* LabWAS: novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. *medRxiv*, 2020.04.08.19011478 (2020).
39. Replication, D.I.G. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234-44 (2014).
40. Sticova, E. & Jirsa, M. ABCB4 disease: Many faces of one gene deficiency. *Ann Hepatol* **19**, 126-133 (2020).
41. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435-44 (2015).
42. Stattermayer, A.F., Halilbasic, E., Wrba, F., Ferenci, P. & Trauner, M. Variants in ABCB4 (MDR3) across the spectrum of cholestatic liver diseases in adults. *J Hepatol* **73**, 651-663 (2020).
43. Willebrords, J. *et al.* Protective effect of genetic deletion of pannexin1 in experimental mouse models of acute and chronic liver disease. *Biochim Biophys Acta Mol Basis Dis* **1864**, 819-830 (2018).
44. Cooreman, A. *et al.* Connexin and Pannexin (Hemi)Channels: Emerging Targets in the Treatment of Liver Disease. *Hepatology* **69**, 1317-1323 (2019).
45. Mehta, M.B. *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).
46. Stender, S. *et al.* Relationship between genetic variation at PPP1R3B and levels of liver glycogen and triglyceride. *Hepatology* **67**, 2182-2195 (2018).
47. Dwyer, A. *et al.* Influence of glycogen on liver density: computed tomography from a metabolic perspective. *J Comput Assist Tomogr* **7**, 70-3 (1983).
48. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* **13**, 397-406 (2014).
49. Duff, M.O. *et al.* Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature* **521**, 376-9 (2015).
50. de Vries, P.S. *et al.* Multiancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions. *Am J Epidemiol* **188**, 1033-1054 (2019).
51. Hammond, L.E. *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**, 8204-14 (2002).
52. Raabe, M. *et al.* Analysis of the role of microsomal triglyceride transfer protein in the liver of tissue-specific knockout mice. *J Clin Invest* **103**, 1287-98 (1999).
53. Cuchel, M. *et al.* Inhibition of microsomal triglyceride transfer protein in familial hypercholesterolemia. *N Engl J Med* **356**, 148-56 (2007).
54. van Gent, M., Sparrer, K.M.J. & Gack, M.U. TRIM Proteins and Their Roles in Antiviral Host Defenses. *Annu Rev Virol* **5**, 385-405 (2018).
55. Medrano, L.M. *et al.* Relationship of TRIM5 and TRIM22 polymorphisms with liver disease and HCV clearance after antiviral therapy in HIV/HCV coinfected patients. *J Transl Med* **14**, 257 (2016).

56. Cai, B. *et al.* Macrophage MerTK Promotes Liver Fibrosis in Nonalcoholic Steatohepatitis. *Cell Metab* **31**, 406-421 e7 (2020).
57. Patin, E. *et al.* Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. *Gastroenterology* **143**, 1244-1252 e12 (2012).
58. Wen, Y. & Ju, C. MerTK - A Novel Potential Target to Treat NASH Fibrosis. *Hepatology* (2020).
59. Day, C.P. From fat to inflammation. *Gastroenterology* **130**, 207-10 (2006).
60. Han, Y.H. *et al.* A maresin 1/RORalpha/12-lipoxygenase autoregulatory circuit prevents inflammation and progression of nonalcoholic steatohepatitis. *J Clin Invest* **129**, 1684-1698 (2019).
61. Chai, C. *et al.* Agonist of RORA Attenuates Nonalcoholic Fatty Liver Progression in Mice via Up-regulation of MicroRNA 122. *Gastroenterology* **159**, 999-1014 e9 (2020).
62. West, L.C. & Cresswell, P. Expanding roles for GILT in immunity. *Curr Opin Immunol* **25**, 103-8 (2013).
63. Chiang, H.S. & Maric, M. Lysosomal thiol reductase negatively regulates autophagy by altering glutathione synthesis and oxidation. *Free Radic Biol Med* **51**, 688-99 (2011).
64. Chapoval, A.I. *et al.* B7-H3: a costimulatory molecule for T cell activation and IFN-gamma production. *Nat Immunol* **2**, 269-74 (2001).
65. Mirea, A.M., Tack, C.J., Chavakis, T., Joosten, L.A.B. & Toonen, E.J.M. IL-1 Family Cytokine Pathways Underlying NAFLD: Towards New Treatment Strategies. *Trends Mol Med* **24**, 458-471 (2018).
66. Mueller, D.L. E3 ubiquitin ligases as T cell anergy factors. *Nat Immunol* **5**, 883-90 (2004).
67. Kleine-Eggebrecht, N. *et al.* Mutation in ITCH Gene Can Cause Syndromic Multisystem Autoimmune Disease With Acute Liver Failure. *Pediatrics* **143**(2019).
68. Baeza-Raja, B. *et al.* Pharmacological inhibition of P2RX7 ameliorates liver injury by reducing inflammation and fibrosis. *PLoS One* **15**, e0234038 (2020).
69. Di Virgilio, F., Dal Ben, D., Sarti, A.C., Giuliani, A.L. & Falzoni, S. The P2X7 Receptor in Infection and Inflammation. *Immunity* **47**, 15-31 (2017).
70. Giuliani, A.L., Sarti, A.C., Falzoni, S. & Di Virgilio, F. The P2X7 Receptor-Interleukin-1 Liaison. *Front Pharmacol* **8**, 123 (2017).
71. Soubeyrand, S., Martinuk, A., Naing, T., Lau, P. & McPherson, R. Role of Tribbles Pseudokinase 1 (TRIB1) in human hepatocyte metabolism. *Biochim Biophys Acta* **1862**, 223-32 (2016).
72. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-73 (2010).
73. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
74. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
75. Kranzler, H.R. *et al.* Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat Commun* **10**, 1499 (2019).

76. Justice, A.C. *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).
77. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
78. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
79. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
80. Hutchinson, A., Watson, H. & Wallace, C. Correcting the coverage of credible sets in Bayesian genetic fine-mapping. *bioRxiv*, 781062 (2019).
81. Haas, M.E. *et al.* Machine learning enables new insights into clinical significance of and genetic contributions to liver fat accumulation. *medRxiv*, 2020.09.03.20187195 (2020).
82. MacLean, M.T. *et al.* Linking abdominal imaging traits to electronic health record phenotypes. *medRxiv*, 2020.09.08.20190330 (2020).
83. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
84. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
85. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
86. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).
87. Fehrmann, R.S. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat Genet* **47**, 115-25 (2015).
88. Cahoy, J.D. *et al.* A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci* **28**, 264-78 (2008).
89. Heng, T.S., Painter, M.W. & Immunological Genome Project, C. The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol* **9**, 1091-4 (2008).
90. Pers, T.H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6**, 5890 (2015).
91. Schmidt, E.M. *et al.* GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601-6 (2015).
92. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
93. 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).
94. Pashos, E.E. *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 e10 (2017).
95. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310 (2015).

96. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol* **17**, 127 (2016).
97. 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).
98. Gagliano Taliun, S.A. *et al.* Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat Genet* **52**, 550-552 (2020).
99. Denny, J.C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205-10 (2010).
100. Loh, P.R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448 (2016).
101. 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).
102. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv*, 2020.08.10.244293 (2020).
103. Shin, S.Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-550 (2014).
104. 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).
105. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**(2018).
106. Sun, B.B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).
107. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
108. Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet* **51**, 51-62 (2019).
109. Pulit, S.L. *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).
110. Teumer, A. *et al.* Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. *Nat Commun* **10**, 4130 (2019).
111. Tin, A. *et al.* Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat Genet* **51**, 1459-1474 (2019).
112. 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).
113. Astle, W.J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429 e19 (2016).
114. 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**, 3305-13 (2015).
115. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **175**, 598-599 (2018).
116. Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* **29**, 1363-1375 (2019).

117. Turei, D., Korcsmaros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* **13**, 966-967 (2016).
118. Ceccarelli, F., Turei, D., Gabor, A. & Saez-Rodriguez, J. Bringing data from curated pathway resources to Cytoscape with OmniPath. *Bioinformatics* **36**, 2632-2633 (2020).