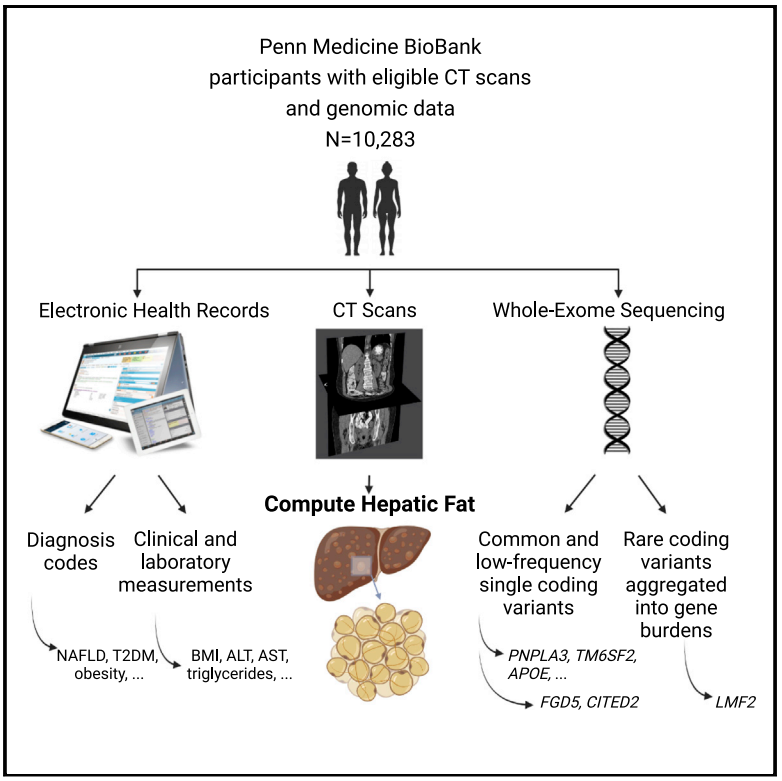
Cell Reports
Medicine

Article

Exome-wide association analysis of CT imaging-derived hepatic fat in a medical biobank

Graphical abstract



Authors

Joseph Park, Matthew T. MacLean, Anastasia M. Lucas, ..., Marylyn D. Ritchie, Walter R. Witschey, Daniel J. Rader

Correspondence

rader@pennmedicine.upenn.edu

In brief

Park et al. perform exome-wide association studies of clinical imaging-derived hepatic fat quantifications in a medical biobank, with cross-modality replication in UK Biobank. They confirm coding variants previously associated with hepatic fat and identify additional coding variants in *LMF2*, *CITED2*, and *FGD5* that replicated in UK Biobank.

Highlights

- CT-based hepatic fat is associated with cardiometabolic traits and diseases
- We confirm known single variants associated with hepatic fat (e.g., *PNPLA3*, *TM6SF2*)
- Additional single variants are associated with hepatic fat (e.g., *FGD5*, *CITED2*)
- Gene burdens of rare pLOF variants are associated with hepatic fat (e.g., *LMF2*)

Park et al., 2022, Cell Reports Medicine *3*, 100855
December 20, 2022
https://doi.org/10.1016/j.xcrm.2022.100855

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Exome-wide association analysis of CT imaging-derived hepatic fat in a medical biobank

Joseph Park,[1,2,3,10] Matthew T. MacLean,[1,4,10] Anastasia M. Lucas,[1,3] Drew A. Torigian,[4] Carolin V. Schneider,[1] Tess Cherlin,[3,5] Brenda Xiao,[1,3] Jason E. Miller,[1,3] Yuki Bradford,[1,3] Renae L. Judy,[6] Regeneron Genetics Center[7] Anurag Verma,[1,3] Scott M. Damrauer,[1,6,8] Marylyn D. Ritchie,[1,3] Walter R. Witschey,[4] and Daniel J. Rader[1,2,9,11,*]

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[5]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[6]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[7]Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, USA
[8]Department of Surgery, Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, USA
[9]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[10]These authors contributed equally
[11]Lead contact
*Correspondence: rader@pennmedicine.upenn.edu
https://doi.org/10.1016/j.xcrm.2022.100855

SUMMARY

Nonalcoholic fatty liver disease is common and highly heritable. Genetic studies of hepatic fat have not sufficiently addressed non-European and rare variants. In a medical biobank, we quantitate hepatic fat from clinical computed tomography (CT) scans via deep learning in 10,283 participants with whole-exome sequences available. We conduct exome-wide associations of single variants and rare predicted loss-of-function (pLOF) variants with CT-based hepatic fat and perform cross-modality replication in the UK Biobank (UKB) by linking whole-exome sequences to MRI-based hepatic fat. We confirm single variants previously associated with hepatic fat and identify several additional variants, including two (*FGD5* H600Y and *CITED2* S198_G199del) that replicated in UKB. A burden of rare pLOF variants in *LMF2* is associated with increased hepatic fat and replicates in UKB. Quantitative phenotypes generated from clinical imaging studies and intersected with genomic data in medical biobanks have the potential to identify molecular pathways associated with human traits and disease.

INTRODUCTION

Hepatic steatosis, or excess accumulation of intrahepatic fat, is a major risk factor for metabolic dysfunction, liver inflammation, and end-stage liver disease accompanied by high morbidity and mortality.[1] In particular, nonalcoholic fatty liver disease (NAFLD) is the most common cause of chronic liver disease in Western countries, and there is growing evidence that the clinical burden of NAFLD extends beyond liver-related morbidity and mortality such as increasing risk for type 2 diabetes mellitus, cardiovascular disease, and chronic kidney disease.[2] While NAFLD has high heritability relative to other metabolic disorders such as obesity and diabetes and our understanding of the genetic underpinnings of NAFLD has advanced,[3,4] known genetic risk variants (e.g., in *PNPLA3* and *TM6SF2*) still explain only a fraction of heritability, suggesting the existence of additional genetic variation such as rare coding variants and non-European-ancestry-predominant variants that may confer risk for or protection from NAFLD that have yet to be uncovered.[5]

Systematic quantification of hepatic fat in population-based epidemiological studies has led to discovery of common genetic variants associated with NAFLD.[6–12] Medical centers collect enormous quantities of advanced abdominal imaging data in the course of clinical care, but imaging-derived quantitative traits such as hepatic fat are not systematically generated for research or clinical use. To address the challenges of conventional analysis of large numbers of images obtained in clinical care, machine learning can be brought to bear to provide quantitative image analysis using automation.[13,14] With the growth of medical biobanks linked to large-scale genomic data generation, there is potential for leveraging imaging-derived phenotypes (IDPs) from clinical imaging studies and integrating with genomic data for discovery.

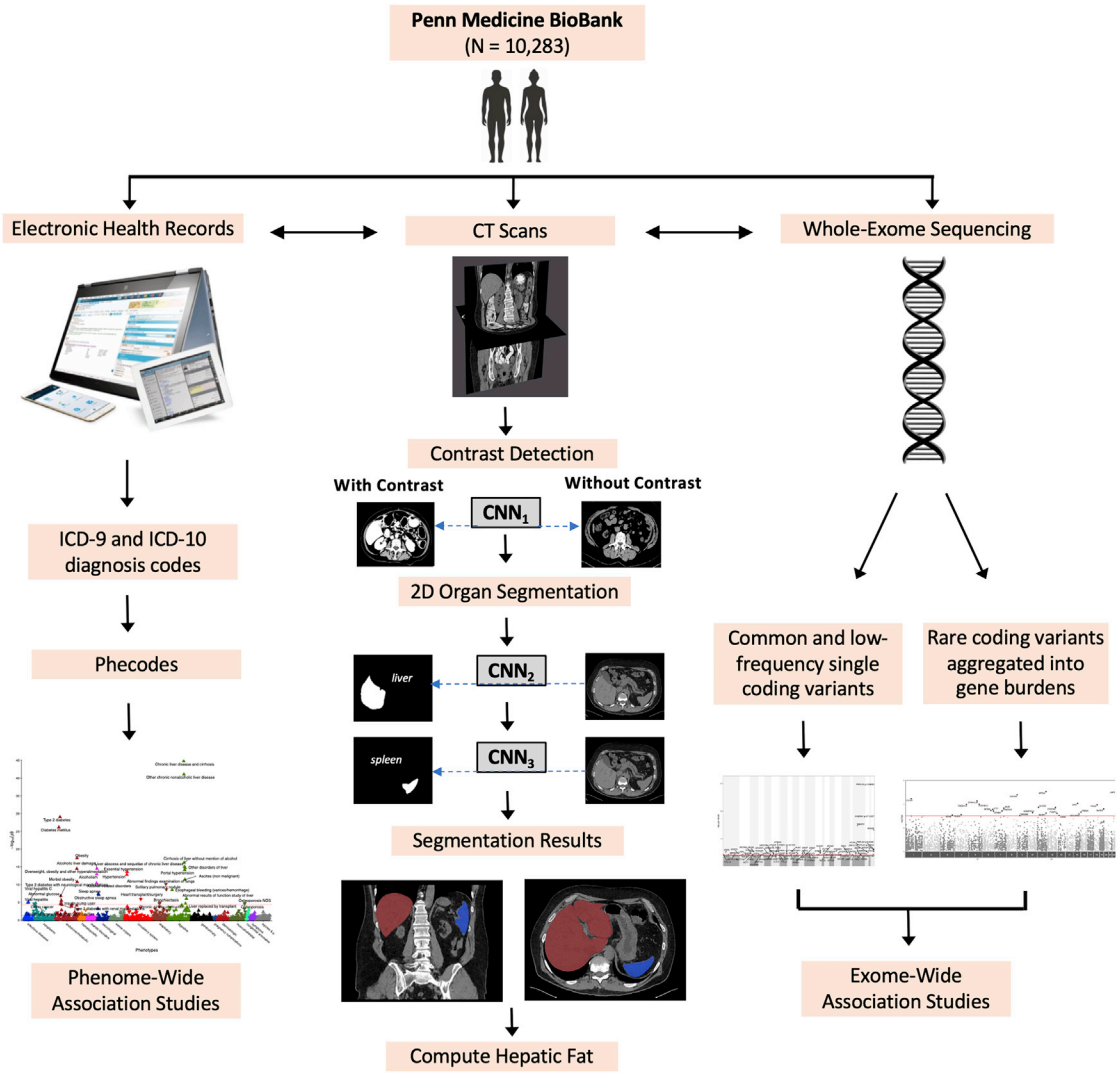


Figure 1. Flowchart of analysis pipeline for phenome-wide and exome-wide association analyses of hepatic fat in the Penn Medicine BioBank

Flowchart diagram showing associations between EHR phenotypes, whole-exome sequence data, and automated quantification of hepatic fat from CT scans in the PMBB. We focused our analyses on the individuals with both whole-exome sequences and quantitated hepatic fat (N = 10,283). For automated quantitation of hepatic fat from clinical CT scans, CNN_1 is the contrast detection network that removes contrasted studies, CNN_2 the liver segmentation network, and CNN_3 the spleen segmentation network. For EHR diagnosis codes, we mapped ICD-9 and ICD-10 codes to Phecodes to conduct phenome-wide association studies. From whole-exome sequences, we analyzed on an exome-wide scale common and low-frequency (MAF > 0.1%) single coding variants as well as gene burdens aggregating rare (MAF $\leq$ 0.1%) coding variants for associations with quantitated hepatic fat.

The Penn Medicine BioBank (PMBB) is a large academic medical biobank in which participants are agnostically recruited from the outpatient setting and consented for access to their electronic health record (EHR) data and permission to generate genomic and biomarker data. Compared with population-based biobanks such as the UK Biobank (UKB), the PMBB is enriched for a wide range of disease[15] and includes a substantial number of participants who have received abdominal and chest computed tomography (CT) scans in the course of routine clinical care. We built a fully automated image curation and organ-labeling technique using deep learning applied to CT scans to quantify hepatic fat and linked this CT-derived hepatic fat quantitation with whole-exome sequence data to identify coding variants associated with hepatic fat that may confer risk for or protect from development of NAFLD.

RESULTS

Automated quantification of hepatic fat from clinical CT scans and validation of the approach

We developed a fully automated approach for extraction of hepatic fat quantifications from abdominal and chest CT scans (Figures 1 and S1A). The contrast detection network was

Table 1. PMBB discovery cohort characteristics, related to Figures 1 and 2

Basic demographics			
Total population, N	10,283		
Female, N (%)	4,551 (44.3)		
Median age, years	69		
Genetically informed ancestry	N (%)		
African (AFR)	2,814 (27.4)		
Mixed American (AMR)	110 (1.1)		
East Asian (EAS)	103 (1.0)		
European (EUR)	7,096 (69.0)		
South Asian (SAS)	84 (0.8)		
Phecodes	N (%)	OR	p
Chronic liver disease and cirrhosis	1,000 (9.7)	1.057	1.70×10^{-45}
Other chronic nonalcoholic liver disease	872 (8.5)	1.057	8.89×10^{-42}
Alcoholic liver damage	129 (1.3)	1.072	3.50×10^{-15}
Viral hepatitis	611 (5.9)	1.022	1.17×10^{-5}
Liver replaced by transplant	260 (2.5)	1.032	2.12×10^{-5}
Portal hypertension	135 (1.3)	1.070	7.72×10^{-15}
Type 2 diabetes	3,018 (29.3)	1.034	9.42×10^{-30}
Obesity	2,998 (29.2)	1.026	3.31×10^{-18}
Essential hypertension	6,089 (59.2)	1.026	2.40×10^{-14}

Basic demographic characteristics and representative Phecodes identified by PheWAS of median hepatic fat in PMBB. Each characteristic is labeled with count data and percentage prevalence where appropriate. Phecodes are additionally labeled with OR corresponding to one unit change in hepatic fat (ΔHounsfield units [HU] = spleen HU – liver HU) and p value as identified from PheWAS. Individuals were determined to be a case for a Phecode if they had the corresponding ICD diagnosis on two or more dates, while controls consisted of individuals who never had the ICD code. Individuals with an ICD diagnosis on only one date as well as those under control exclusion criteria based on Phecode mapping protocols were not considered.

evaluated on 400 randomly selected scans (200 with IV contrast and 200 without) and classified 399 correctly. We then assessed the performance of the liver and spleen segmentation networks by comparing automated versus manual full-volume contours (Figure S1B) and achieved mean percentage overlaps (Dice ± SD) of 0.95 ± 0.02 and 0.92 ± 0.07 respectively. Finally, we compared the mean attenuation computed by the automated method versus expert radiologic review and found excellent agreement with interclass correlation coefficients of 0.976 and 0.956 for the liver and spleen, respectively.

Hepatic fat extracted from clinical CT scans is highly significantly associated with a range of cardiometabolic diseases and traits

Among exome-sequenced individuals in PMBB (N = 10,283; Table 1), we conducted a phenome-wide association study (PheWAS) of the quantitative trait of median hepatic fat to interrogate the clinical EHR diagnosis phenotypes associated with hepatic fat (Figure 2). Hepatic fat quantity was associated with increased risk for chronic liver disease and cirrhosis ($p = 1.70 \times 10^{-45}$) and other chronic nonalcoholic liver disease (Phecode representing NAFLD; $p = 8.89 \times 10^{-42}$) at phenome-wide significance. Hepatic fat also showed phenome-wide significant associations with increased risk for cardiometabolic comorbidities such as type 2 diabetes ($p = 9.42 \times 10^{-30}$), obesity ($p = 3.31 \times 10^{-18}$), and hypertension ($p = 1.64 \times 10^{-13}$). Additionally, viral hepatitis ($p = 1.17 \times 10^{-5}$) and alcoholic liver damage ($p = 3.50 \times 10^{-15}$) were associated with increased hepatic fat at phenome-wide significance. Hepatic fat was also highly significantly associated with the quantitative trait of body mass index (BMI) (Table S1), consistent with the known relationship between obesity and hepatic steatosis. We also analyzed the association of hepatic fat with several clinical EHR laboratory quantitative traits (Table S1). We found that hepatic fat values were significantly positively associated with serum alanine aminotransferase (ALT), aspartate transaminase (AST), alkaline phosphatase, hemoglobin A1C, random glucose, and random triglycerides, and significantly inversely associated with high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, and total cholesterol. Thus, our CT-derived hepatic fat quantitation has the expected associations with cardiometabolic diseases and traits, helping to further validate this approach.

Exome-wide analyses of single coding variants identify variants associated with hepatic fat

After excluding individuals with alcohol-related and viral hepatitis diagnoses in order to study the genetic etiologies of NAFLD, we conducted a univariate exome-wide analysis of hepatic fat for all nonsynonymous coding variants of sufficient frequency (minor allele frequency [MAF] >0.1% in gnomAD) (Figures 3A, S2A, and S4). Among 120,315 total variants with at least 10 carriers, we identified 91 variants in 86 genes with exome-wide significant ($p < 4.2 \times 10^{-7}$) or suggestive ($p < 9.9 \times 10^{-5}$) associations with hepatic fat (Tables 2 and S2A). These included variants

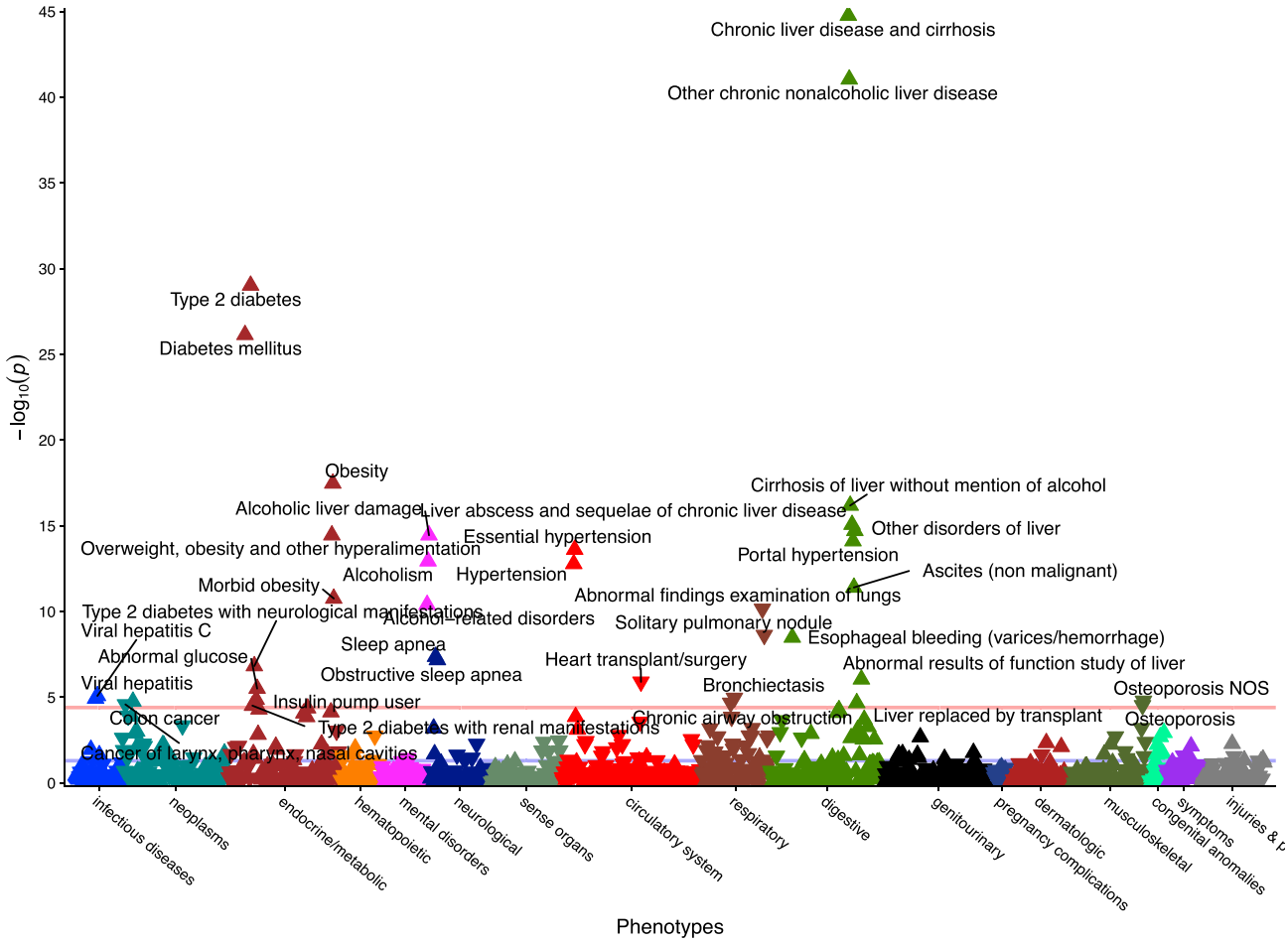


Figure 2. PheWAS of median hepatic fat in PMBB

PheWAS plot of the associations of Phecodes with median hepatic fat among exome-sequenced individuals in PMBB (N = 10,283). Associations were conducted as trans-ancestral cosmopolitan analyses adjusted for age, genetically determined sex, and PC1-10. Phecodes are plotted along the x axis to represent the phenome, and the association of median hepatic fat with each Phecode is plotted along the y axis representing $-\log_{10}$(p value). The red line represents the Bonferroni-corrected significance threshold to adjust for multiple testing ($p = 3.58 \times 10^{-5}$), and the blue line represents a nominal significance threshold ($p = 0.05$).

previously reported to be associated with hepatic fat and/or NAFLD: *PNPLA3* variants I148M, K434E, and S453I, *TM6SF2* E167K, and APOE4 (C130R). Additional positive control associations were found below the significance threshold (Table S2B), including *GCKR* L446P, *MTARC1* T165A, and *TM6SF2* L156P. Twenty-seven of the 91 single variants were African-ancestry-specific variants (African/European MAF ratio >10 in gnomAD; Table S2A).

For replication, we tested the association of these 91 variants with liver proton density fat fraction (PDFF) measurements derived from abdominal MRI scans in 9,049 participants in the UKB who also had whole-exome sequencing available (Table S2C). Not only were the methods of hepatic fat quantitation different but there were notable differences in the distribution of hepatic fat between PMBB and UKB (Figure S3). Furthermore, there was insufficient power in UKB for replication of the 27 African-ancestry-specific variants, as 20 of 27 variants had N < 5 carriers in the cosmopolitan analyses. Despite this, in addition to replicating a number of the previously known associations, we also replicated associations of H600Y in *FGD5* and a 6-bp deletion (S198_G199del) in *CITED2* with hepatic fat.

To characterize the clinical implications of these replicated variants, we conducted PheWAS of each variant. The H600Y variant in *FGD5* was nominally associated with Phecodes related to insulin resistance in PMBB including impaired fasting glucose (odds ratio [OR] 1.700, p = 0.0131) and hyperglyceridemia (OR = 1.800, p = 0.0464) as well as diagnoses for both alcohol-related and nonalcoholic liver diseases such as alcoholic liver damage (OR = 2.217, p = 0.0216), liver replaced by transplant (OR = 1.920, p = 0.0230), and other chronic nonalcoholic liver disease (OR = 1.388, p = 0.0386). Additionally, the 6-bp deletion (p.S198_G199del) in *CITED2* was associated with increased risk for liver replaced by transplant (OR = 54.379, p = 1.043E-04), impaired fasting glucose (OR = 19.581, $p = 3.484 \times 10^{-3}$),

CellPress
OPEN ACCESS

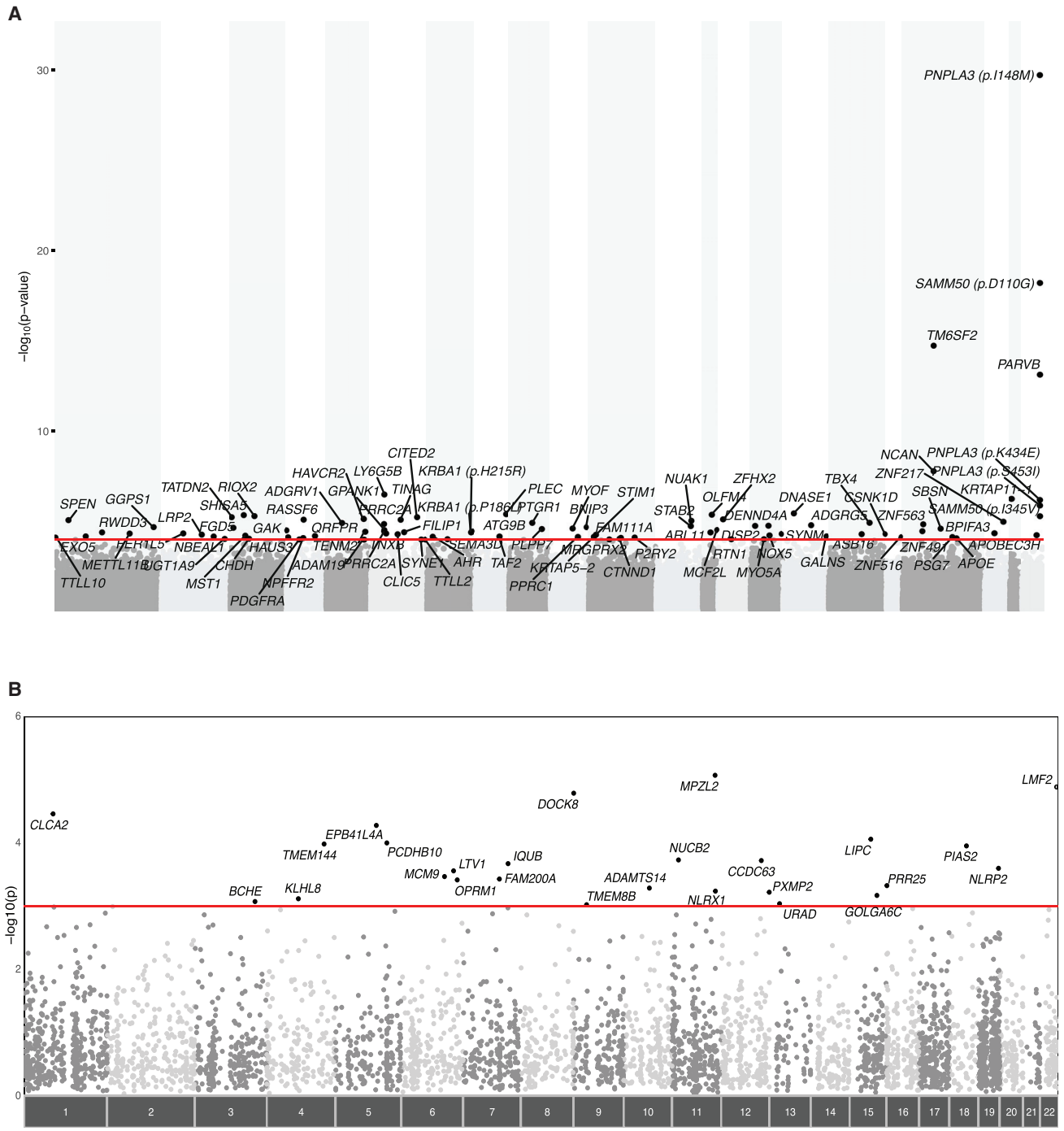


Figure 3. Manhattan plots of exome-wide discovery analyses in PMBB

(A) Manhattan plot showing the results of the exome-wide single-variant discovery analysis in PMBB (N = 9,594) for coding variants of sufficient frequency (MAF > 0.1%, N ≥ 10). The x axis represents the exome and is organized by chromosomal location. The location of each single variant along the x axis corresponds to the genomic location for each variant according to Genome Reference Consortium Human build 38 (GRCh38). The most significant association of each single variant with hepatic fat is plotted vertically above each variant, and the height of each point represents the $-\log_{10}$(p value) of the association. Each variant is annotated with its corresponding gene name and amino acid change for genes with multiple exome-wide significant variants. The red line represents the suggestive significance threshold at $p = 9.9 \times 10^{-5}$ to account for multiple hypothesis testing.

(legend continued on next page)

and diabetes mellitus (OR = 2.531, p = 8.187 × 10^{-3}) in UKB. A comprehensive review of the clinical indications for the CT scans included in the PMBB discovery study for carriers of the *FGD5* and *CITED2* variants did not show signs of bias for other abdominal conditions, which may have confounded our original findings.

To test whether rare predicted deleterious coding variants in the 86 genes containing the 91 single variants were also associated with differences in hepatic fat, we aggregated non-overlapping rare (MAF $\leq$ 0.1% in gnomAD) predicted loss-of-function (pLOF) variants into gene burdens for targeted associations with hepatic fat in PMBB (Table S3A). Gene burdens of rare pLOF variants in eight genes were significantly associated with hepatic fat, including genes with previously described common variation associated with differences in hepatic fat such as *PNPLA3* and *PARVB*, as well as additional findings such as *PTGR1*. We also aggregated the combination of rare pLOF and rare predicted deleterious missense (pDM) variants (rare exonic variant ensemble learner [REVEL] $\geq$ 0.5) per gene for targeted gene burden association with hepatic fat in PMBB (Table S3A). We found 11 additional genes associated with differences in hepatic fat in PMBB by adding rare pDM variants to pLOFs, including *SAMM50*, which contains previously described common variation associated with differences in hepatic fat. Of note, the combined *CITED2* gene burden was nominally associated with increased median hepatic fat, although underpowered (N = 3, beta = 9.723, p = 0.0276). We also tested the burden of rare pLOFs in the 86 genes for association with hepatic fat in UKB (Table S3B) and found three concordant gene burdens, namely *ADAM19*, *EXO5*, and *SEMA3D*. Additionally, the combined burden of rare pLOFs and rare pDM variants (Table S3B) revealed three additional significant genes, namely *ADGRG5*, *GPS1*, and *NOX5*.

On exome-wide analysis, a gene burden of rare pLOFs in *LMF2* is associated with increased hepatic fat in PMBB and UKB

We also performed an exome-wide rare pLOF gene burden analysis for association with hepatic fat in PMBB. We aggregated rare (MAF $\leq$ 0.1% in gnomAD) pLOFs per gene: among 4,187 genes with at least 10 carriers for rare pLOFs who have CT-derived hepatic fat quantifications available, there were 26 genes that had exome-wide significant ($p < 1.2 \times 10^{-5}$) or suggestive ($p < 9.9 \times 10^{-4}$) associations with hepatic fat (Figures 3B, S2B, and S4; Table S4). For these 26 genes, we attempted replication in UKB using a similar rare pLOF gene burden approach. We found that the *LMF2* pLOF gene burden had a significant association with hepatic fat (beta = 0.429, p = 5.79 × 10^{-3}, N = 5) and, importantly, in the same direction (increased) as we observed in PMBB.

We conducted an analysis of the association of the gene burden of rare pLOFs in *LMF2* in PMBB (N = 105 het carriers) with Phecodes in the digestive group (520–579.8). In addition to a significant association with the NAFLD Phecode, the burden of pLOF variants in *LMF2* had significant associations with an array of biliary-related Phecodes such as cholangitis, calculus of bile duct, cholelithiasis with acute cholecystitis, and primary biliary cirrhosis (Table S5A). Furthermore, the *LMF2* pLOF gene burden was associated with increased serum alkaline phosphatase, total cholesterol levels, and BMI (Table S5B), as well as a trend to an increase in triglycerides (beta = 0.0474, p = 0.0957). There were no significant associations with HDL, LDL, ALT, or AST levels. Additionally, a comprehensive review of the clinical indications for the CT scans included in the PMBB discovery study for carriers of *LMF2* pLOF variants did not show signs of bias for other abdominal conditions, which may have confounded our original findings.

Hepatic *LMF2* expression is significantly increased in histologically proven NAFLD cases versus controls

In a primary analysis of publicly available data, we found that *LMF2* expression was significantly increased in human livers of histologically proven NAFLD cases compared with control (Table S6). We conducted differential gene expression analyses for genes nominated by exome-wide association analyses in PMBB with replication in UKB, namely *LMF2*, *FGD5*, and *CITED2*. We found that hepatic *LMF2* expression was significantly increased in both histologically characterized early and moderate NAFLD cases compared with controls, while there was no significant difference in hepatic expression of *FGD5* and *CITED2* in NAFLD versus controls.

DISCUSSION

Hepatic steatosis is common but markedly underdiagnosed, hindering EHR-based research into the spectrum of NAFLD. CT scans performed during clinical care allow for opportunities to quantitate IDPs such as hepatic fat that can be used for clinical and genomic research. We developed and applied a machine learning approach to automating the quantification of hepatic fat from non-contrast chest and abdomen/pelvis CT images. By integrating the hepatic fat quantitative trait with coding variants obtained on whole-exome sequencing in PMBB, we conducted exome-wide association of rare, low-frequency, and common coding variants with hepatic fat in a medical biobank. Our single variant analyses confirmed several previously described coding variants associated with hepatic fat, supporting the validity of our approach. Several additional variants were associated with CT-derived hepatic fat, two of which were replicated in a separate MRI-based quantitation of liver fat in the UK Biobank.

Our findings include coding variants in the genes *FGD5* and *CITED2* significantly associated with hepatic fat. The variant H600Y in *FGD5*, a predicted deleterious missense variant (Combined Annotation Dependent Depletion score, i.e.,

(B) Manhattan plot showing the results of the exome-wide gene burden discovery analysis in PMBB (N = 9,594) aggregating rare (MAF $\leq$ 0.1%) predicted loss-of-function (pLOF) variants per gene (N $\geq$ 10). The x axis represents the exome and is organized by chromosomal location. The location of each gene along the x axis corresponds to the genomic location for each variant according to GRCh38. The most significant association of each gene burden with hepatic fat is plotted vertically above each gene, and the height of each point represents the $-\log_{10}$(p value) of the association. Each gene is annotated with its gene name. The red line represents the suggestive significance threshold at p = 9.9 × 10^{-4} to account for multiple hypothesis testing.

Table 2. Significant results from single-variant discovery analyses in the PMBB and replication in UKB, related to Figure 3

Discovery in PMBB					
Gene	AA change	rs ID	N	Beta	p
PNPLA3	p.I148M	rs738409	3,639	1.564	1.91×10^{-30}
SAMM50	p.D110G	rs3761472	2,884	1.325	6.02×10^{-19}
*TM6SF2**	p.E167K	rs58542926	847	2.529	1.74×10^{-15}
PARVB	p.W37R	rs1007863	4,240	1.202	7.12×10^{-14}
*NCAN**	p.P92S	rs2228603	885	1.737	1.71×10^{-8}
LY6G5B	p.R176C	rs9267532	861	1.535	3.16×10^{-7}
KRTAP11-1	p.S78F	rs79258920	13	12.711	5.28×10^{-7}
*PNPLA3***	p.K434E	rs2294918	5,581	0.795	6.09×10^{-7}
PNPLA3	p.S453I	rs6006460	477	−1.678	1.23×10^{-6}
DNASE1	p.R207C	rs148373909	12	10.554	3.42×10^{-6}
OLFM4	p.R214X	rs34067666	11	12.032	3.98×10^{-6}
PLEC	p.P4259S	rs202040785	25	6.859	4.25×10^{-6}
SHISA5	p.P236L	rs141742404	14	9.525	4.35×10^{-6}
*SAMM50***	p.I345V	rs8418	5,495	0.723	4.81×10^{-6}
RIOX2	p.P140L	rs41265444	192	2.699	4.92×10^{-6}
TATDN2	p.M416T	rs61730105	502	1.510	5.66×10^{-6}
CITED2	p.S198_G199del	rs531316452	11	12.473	5.66×10^{-6}
HAVCR2	p.T154S	rs150536405	16	10.130	7.18×10^{-6}
ZFHX2	p.S129P	rs142184428	11	11.147	7.47×10^{-6}
TINAG	p.K407N	rs140019555	10	11.128	7.84×10^{-6}
RASSF6	p.Q39K	rs145319675	11	9.876	7.85×10^{-6}
SPEN	p.A970V	rs848208	1,458	−0.933	8.39×10^{-6}
NUAK1	p.R595L	rs141618950	27	6.644	8.78×10^{-6}
ZNF217	p.R903Q	rs61748378	39	5.243	9.75×10^{-6}
PTGR1	splicing	rs146469061	13	10.050	1.18×10^{-5}
ADGRV1	p.R249K	rs41303344	10	10.919	1.19×10^{-5}
TBX4	p.P288A	rs193204039	10	10.141	1.20×10^{-5}
PRRC2A†	p.S1219Y	rs41273264	150	2.915	1.35×10^{-5}
ZNF563	p.R353X	rs112896133	186	2.913	1.41×10^{-5}
ADGRG5	p.C425Y	rs114796383	13	8.763	1.60×10^{-5}
DENND4A	p.S248N	rs201378259	20	−7.579	1.69×10^{-5}
STAB2	p.D2021N	rs116894406	168	2.619	1.72×10^{-5}
DISP2	p.R114Q	rs35070171	647	1.540	1.72×10^{-5}
BNIP3	p.R143K	rs143231747	48	5.419	2.04×10^{-5}
GGPS1	p.A146D	rs147202180	15	8.039	2.11×10^{-5}
FGD5	p.H600Y	rs144177006	16	7.801	2.17×10^{-5}
MYOF	p.T498M	rs190149415	21	8.179	2.40×10^{-5}
SBSN	p.N529S	rs75962883	20	6.952	2.45×10^{-5}
PLPP7	p.D54E	rs141024100	10	10.452	2.66×10^{-5}
MCF2L	p.G1000R	rs12429945	24	7.799	2.78×10^{-5}
GPANK1††	p.R41L	rs3130618	2,076	1.164	2.92×10^{-5}
GAK	p.K1265R	rs2306242	516	1.474	3.05×10^{-5}
PRRC2A††	p.P2006S	rs10885	2,074	1.156	3.34×10^{-5}
ATG9B	p.N493S	rs7804893	2,680	0.745	3.46×10^{-5}
ZNF491	p.E109Q	rs149778854	10	−11.327	3.50×10^{-5}
TENM2	p.T662M	rs201157994	29	7.010	3.63×10^{-5}
FILIP1	p.T1126M	rs35227190	42	5.353	3.81×10^{-5}

(*Continued on next page*)

Table 2. ***Continued***

Gene	AA change	rs ID	N	Beta	p
Discovery in PMBB					
ARL11	p.V64I	rs143660006	13	8.993	3.92×10^{-5}
KRBA1	p.P186L	rs114410482	30	6.528	3.95×10^{-5}
KRBA1	p.H215R	rs188335425	30	6.528	3.95×10^{-5}
RWDD3	p.I15V	rs142820652	12	8.690	3.97×10^{-5}
BPIFA3	p.D166N	rs142257117	25	6.787	4.22×10^{-5}
TNXB†	p.R1064H	rs61995676	166	2.628	4.49×10^{-5}
FER1L5	p.T687A	rs7599598	2,417	−1.347	4.50×10^{-5}
CSNK1D	p.M404V	rs112902236	33	6.431	4.72×10^{-5}
SYNM	p.A266V	rs140039713	106	3.089	4.76×10^{-5}
ASB16	p.R428W	rs75035743	14	10.252	4.87×10^{-5}
CLIC5	p.T114A	rs723580	202	2.224	4.93×10^{-5}
LRP2	p.L726F	rs144451000	22	7.136	5.25×10^{-5}
STIM1	splicing	rs118128831	176	2.392	5.34×10^{-5}
APOBEC3H	p.E7Kfs*28	rs760113060	23	7.062	5.38×10^{-5}
NOX5	p.R759G	rs7168025	179	2.547	5.79×10^{-5}
MST1	splicing	rs201139286	22	6.281	5.92×10^{-5}
QRFPR	p.I232S	rs139457842	12	9.558	6.24×10^{-5}
GALNS	p.R376Q	rs150734270	28	6.828	6.38×10^{-5}
EXO5	p.L151P	rs35672330	810	1.106	6.61×10^{-5}
AHR	p.V570I	rs4986826	60	4.087	6.67×10^{-5}
TAF2	p.I53V	rs112002462	36	−5.305	6.83×10^{-5}
NBEAL1	p.N1966S	rs143836127	32	−5.566	6.85×10^{-5}
ZNF516	p.C610R	rs117566743	320	1.744	6.86×10^{-5}
KRTAP5-2	p.C137R	rs138473551	150	2.601	6.92×10^{-5}
METTL11B	p.P112S	rs183687510	16	7.836	7.00×10^{-5}
PPRC1	p.A1192V	rs144188174	10	9.756	7.00×10^{-5}
PSG7	p.I83V		858	1.018	7.11×10^{-5}
P2RY2	p.R334C	rs1626154	201	2.111	7.55×10^{-5}
FAM111A	p.Q451E	rs116918730	36	5.793	7.89×10^{-5}
TTLL10	p.R623Q	rs540494380	29	6.376	7.90×10^{-5}
NPFFR2	p.V250A	rs61733659	141	2.398	8.10×10^{-5}
APOE	p.C130R	rs429358	2,587	−0.626	8.39×10^{-5}
HAUS3	p.Y402C	rs143360308	13	−9.915	8.69×10^{-5}
CHDH	p.N441S	rs34974961	16	−7.745	8.70×10^{-5}
UGT1A9	p.M33T	rs72551330	153	−2.666	9.01×10^{-5}
CTNND1	p.T113P	rs201815246	19	7.803	9.01×10^{-5}
SEMA3D	p.D186N	rs148351346	12	9.769	9.33×10^{-5}
MYO5A	p.R1320S	rs61731219	29	5.348	9.41×10^{-5}
RTN1	p.T124A	rs61736371	164	−2.657	9.45×10^{-5}
MRGPRX2	p.N62S	rs10833049	1,833	−0.806	9.61×10^{-5}
PDGFRA	p.R376Q	rs41279521	10	10.825	9.65×10^{-5}
SYNE1	p.L3050V	rs117360770	35	5.201	9.66×10^{-5}
TTLL2	p.G445S	rs9457304	1,110	−1.017	9.75×10^{-5}
ADAM19	p.G660D	rs2287749	1,600	−0.798	9.87×10^{-5}
Replication in UKB					
PNPLA3	p.I148M	rs738409	3,253	0.094	1.56×10^{-49}
*TM6SF2**	p.E167K	rs58542926	1,304	0.133	8.27×10^{-43}

(*Continued on next page*)

Table 2. *Continued*

Discovery in PMBB					
Gene	AA change	rs ID	N	Beta	p
*NCAN**	p.P92S	rs2228603	1,323	0.106	2.11×10^{-27}
SAMM50	p.D110G	rs3761472	2,546	0.072	1.57×10^{-23}
PARVB	p.W37R	rs1007863	4,923	0.049	8.17×10^{-18}
APOE	p.C130R	rs429358	2,512	−0.054	8.10×10^{-14}
*PNPLA3***	p.K434E	rs2294918	7,528	0.027	3.73×10^{-7}
*SAMM50***	p.I345V	rs8418	7,457	0.022	3.31×10^{-5}
PNPLA3	p.S453I	rs6006460	12	−0.268	5.78×10^{-3}
CITED2	p.S198_G199del	rs531316452	6	0.317	2.54×10^{-2}
FGD5	p.H600Y	rs144177006	197	0.056	2.60×10^{-2}

Top: list of significant ($p < 9.9 \times 10^{-5}$) single-variant associations with hepatic fat from the single-variant discovery analyses in PMBB. Each single variant is annotated with its gene name, amino acid change where appropriate, rs ID if available, and the number of individuals who carry at least one copy of the alternate allele who also have hepatic fat quantitated in PMBB. Each significant single variant is listed with its most significant beta and p value from linear regression analyses out of all analyses listed in Table S2A. Single variants are ranked by increasing p values. Pairs of variants marked with *, **, †, or †† are in linkage disequilibrium with $R^2 > 0.7$ according to 1000 Genomes.

Bottom: list of significantly replicated ($p < 0.05$) single-variant associations with hepatic fat in UKB. Each single variant is annotated with its gene name, amino acid change where appropriate, rs ID if available, and the number of individuals who carry at least one copy of the alternate allele who also have hepatic fat quantitated in UKB. Each significant single variant is listed with its most significant beta and p value from linear regression analyses out of all analyses listed in Table S2C. Single variants are ranked by increasing p values. Of note, betas are based on normalized MRI-based PDFF values and are thus different in magnitude compared with those in the PMBB discovery, where CT-based HU were used. Pairs of variants marked with *, **, †, or †† are in linkage disequilibrium with $R^2 > 0.7$ according to 1000 Genomes.

CADD[16] = 24.1), was associated with increased hepatic fat in PMBB and replicated in UKB. The H600Y variant was also associated with clinical diagnoses related to insulin resistance and both alcoholic and nonalcoholic liver diseases in PMBB. *FGD5* encodes a liver-expressed protein[17] that is expressed preferentially in the hepatic endothelium.[18] While *FGD5* has not previously been described in the liver, this expression pattern is consistent with its role in regulating vascular endothelial growth factor (VEGF) signaling during angiogenesis.[19] FGD5 also activates CDC42,[20] a member of the Rho GTPase family, and plays important roles in the regulation of the cytoskeleton as well as cell proliferation, polarity, and transport. Importantly, liver-specific knockout of *Cdc42* in mice has been shown to lead to excessive hepatic accumulation of lipids during liver regeneration after partial hepatectomy, likely due to impaired cytoskeletal organization and intracellular trafficking in hepatocytes.[21] Heterozygous knockout mice for *Fgd5* also had abnormal liver morphology and lower liver weight compared with wild type in unpublished data from the International Mouse Phenotyping Consortium.[22]

Additionally, we found that a 6-bp deletion (p.S198_G199del) with predicted deleteriousness (CADD = 20.1) in *CITED2*, a coactivator of HNF4α, was also associated with increased hepatic fat in PMBB and replicated in UKB. This variant was also associated with clinical diagnoses related to insulin resistance and liver transplantation in the UK Biobank, although underpowered. A burden of rare predicted deleterious coding variants in *CITED2* was also associated with increased hepatic fat (although underpowered). Importantly, HNF4α is essential for normal liver architecture and organization of the sinusoidal endothelium during development, and its expression is crucial for differentiation of hepatocytes, accumulation of hepatic glycogen stores, and generation of hepatic endothelium in adults.[23] Mice lacking *HNF4A* have high hepatic lipid accumulation, impaired gluconeogenesis during fasting, and defective lipid transport and metabolism.[24,25] Furthermore, *Cited2* has been shown to be essential for mouse fetal liver development, and knockout of *Cited2* in fetal liver leads to disrupted sinusoidal architecture and accumulation of lipid droplets in the sinusoidal space.[26]

By leveraging our whole-exome data and extending our analyses to gene burdens of rare deleterious variants in genes nominated by the single-variant discovery, we gained additional insights into genes not previously associated with hepatic fat in humans. For example, a relatively rare splicing variant in the gene *PTGR1* was associated with increased hepatic fat; we also found that a gene burden of rare pLOFs as well as predicted deleterious missense variants in *PTGR1* were also associated with increased hepatic fat in PMBB. *PTGR1* encodes an enzyme called Prostaglandin Reductase 1, which is involved in the inactivation of the chemotactic factor leukotriene B4 and has its highest expression in the liver.[27] Notably, leukotriene B4 has been shown to promote insulin resistance in mouse hepatocytes,[28] suggesting that haploinsufficiency of *PTGR1* could lead to increased activity of leukotriene B4 and adverse effects on liver metabolism. Additionally, the single variant G660D in *ADAM19*, a predicted deleterious missense variant (CADD = 23.9), was associated with decreased hepatic fat; we also found that a gene burden of pLOF variants in *ADAM19* was associated with decreased hepatic fat in PMBB (although underpowered) and UKB. Importantly, *ADAM19* has been suggested to be pro-obesogenic and enhance insulin resistance in mice,[29] consistent with our observation that reduced function is associated with decreased hepatic fat and suggesting that silencing could be a potential therapeutic approach to NAFLD. Furthermore, a single variant I232S in *QRFPR* (*GPR103*), a

predicted deleterious missense variant, was associated with increased hepatic fat; gene burdens of pLOFs as well as pLOFs combined with pDM variants in this gene were also associated with increased hepatic fat. *QRFPR* encodes the G-protein-coupled receptor for neuropeptide 26RFa (encoded by *QRFP*). 26RFa and *QRFPR* work both in the hypothalamic nuclei to control feeding behavior as well as in the gut and pancreatic islets.[30] Specifically, 26RFa increases insulin sensitivity and prevents pancreatic beta cell death and apoptosis, and disruption leads to dysregulation of glucose homeostasis and a deficit in insulin production by pancreatic islets.[31–33] The mechanisms by which haploinsufficiency of *QRFPR* increases risk for hepatic steatosis remain to be determined.

There is a substantial gap of knowledge regarding the clinical implications of genetic variants overrepresented among individuals of African ancestry.[34] In the PMBB discovery cohort, 27.4% of individuals with whole-exome sequencing linked to hepatic fat quantifications were of African ancestry, and, interestingly, we identified 27 African-ancestry-specific or -predominant single variants that were significantly associated with hepatic fat. These variants represented a challenge for replication in UKB given there were only 70 individuals who identified as having African, Caribbean, or any other black ethnic background in the subset of individuals with MRI-derived hepatic fat linked to whole-exome sequencing data in UKB. The previously described AFR-predominant S453I variant in *PNPLA3*, among the more common of the 27 AFR-predominant variants, was able to be replicated via cosmopolitan analyses in UKB. However, the rest of the AFR-predominant single variants were in very low numbers in UKB, with 20 of 27 variants having $N < 5$ carriers and thus not even being included in replication studies. Our findings suggest that larger experiments of this type in ethnically diverse cohorts are essential for improving our understanding of the contribution of ancestry-specific genetic variation to the regulation of intrahepatic fat.

Our single variant analysis identified a number of genes and variants previously associated with hepatic fat and/or NAFLD and in some cases extends previous observations. For example, *PNPLA3* I148M and K434E were both significantly associated with increased hepatic fat as previously reported,[6,35] whereas S453I, which is predominant in individuals of African ancestry, was significantly associated with *reduced* hepatic fat. Importantly, our gene burden analysis of pLOFs in *PNPLA3* also showed a significant association with reduced hepatic fat, suggesting that S453I is likely a reduced function allele, in contrast to I148M and K434E, which are likely to have a toxic gain of function. Our results support the concept that silencing of *PNPLA3* may be a potential strategy for therapeutic intervention for NAFLD.[36] We confirmed previous reports that *TM6SF2* E167K and L156P are associated with increased hepatic fat,[8,37] and that *GCKR* L446P[38] and *MTARC1* A165T[39] were associated with decreased hepatic fat. While we found no association between a burden of rare pLOF ± pDM variants in *TM6SF2* and *GCKR* with hepatic fat, we did see a nominal association between a gene burden of rare pLOF and pDM variants in *GCKR* with increased hepatic fat (beta = 1.82, p = 0.035, N = 54), which is consistent with previous observational data.[40]

We confirmed that D110G in *SAMM50* was associated with increased hepatic fat,[41] but noted that a gene burden of rare pLOFs and predicted deleterious missense variants in *SAMM50* was significantly associated with *decreased* hepatic fat in PMBB, suggesting that D110G could represent a gain-of-function allele. A similar situation was seen with *PARVB*, where the W37R variant was confirmed to be associated with increased hepatic fat,[42] but a gene burden of rare pLOFs alone as well as rare pLOFs plus pDM variants in *PARVB* was associated with decreased hepatic fat. While the *SAMM50* D110G and *PARVB* W37R variants are in relatively close proximity to *PNPLA3* I148M (about 43 and 71 kb away respectively), they are in weak or moderate linkage disequilibrium ($R^2 < 0.7$) with *PNPLA3* I148M. Thus, while additional functional work is needed to determine the mechanisms underlying the associations of *PARVB* and *SAMM50* gene burdens with hepatic fat, the signals seen from rare pLOFs and/or pDMs in each of these genes suggest a protective role for haploinsufficiency in these genes and nominates both *SAMM50* and *PARVB* as candidates for therapeutic silencing as an approach to NAFLD.

We also report rare variant gene burden discovery analysis for exome-wide gene-based associations with hepatic fat. We identified a burden of rare pLOF variants in *LMF2* as being associated with increased hepatic fat in PMBB and replicated this observation in UKB. Importantly, common coding variants in *LMF2* were not exome-wide significant in our single-variant discovery. *LMF2* is a paralog of *LMF1* and is the ancestral gene, and *Lmf1* emerged in echinoderms after losing an internal segment of the DUF1222 domain and gaining a C-terminal tail with lipase maturation activity.[43] While *LMF2* does not have this C terminus, it may share a common ancestral cellular function with *LMF1*, possibly the maintenance of ER homeostasis. Additionally, we found through a primary analysis of publicly available transcriptomic data in livers of histologically characterized human NAFLD cases that hepatic expression of *LMF2* is increased in NAFLD versus control. While additional investigation of the function of *LMF2* is needed, given the relevance of ER stress signaling in the pathogenesis of NAFLD and its progression to chronic liver disease,[44] our gene burden and transcriptomic analyses together suggest that *LMF2* plays a role in the ER stress response to accumulation of lipids in hepatocytes, and that *LMF2* haploinsufficiency may contribute to an increase in hepatic fat.

While several single variants successfully replicated in UKB following discovery analyses in PMBB, we noticed a relative lack of significant findings when interrogating rare variant gene burdens in UKB regarding replication of significant genes from the gene burden discovery as well as targeted analyses of genes nominated by the single variant discovery. Even after log-transformation of MRI-PDFF in UKB to account for MRI-specific differences in quantification of hepatic fat and allow for regression analyses for replication, the distribution of hepatic fat was still substantially skewed toward lower values in UKB compared with PMBB. This might be expected, given the UKB is a population-based biobank that is widely recognized to have a "healthy volunteer selection bias,"[45] and we have previously described the relative lack of replication for rare variant gene burdens in UKB compared with medical biobanks.[15] Thus, our study

suggests that additional experiments of this type linking exome sequencing to imaging-derived hepatic fat quantifications in medical biobanks are warranted to interrogate the impact of rare variation on differences in hepatic fat.

In 2021, the American Gastroenterological Association and the Lancet Global Commission both issued guidance on the urgent need to develop population-based approaches for NAFLD.[46,47] Studies suggest that 11% of patients with incidentally discovered hepatic steatosis are at high risk for advanced hepatic fibrosis, and experts recommend identification and further clinical evaluation of patients with suspected steatosis.[48] Identification of hepatic steatosis has clinical implications for monitoring of transaminases and interventions such as weight loss and bariatric surgery. Our automated approach to quantitating hepatic fat from chest and abdominal CT scans was validated by our PheWAS demonstrating many expected clinical associations and by the associations with genetic variants known to be associated with steatosis. This approach could be included as a standard aspect of chest and abdominal CT scans, leading to an automated reporting of liver fat as an incidental finding and potentially leading to improved monitoring and treatment of patients with hepatic steatosis. Further work is necessary to demonstrate the clinical impact of this approach. Additionally, broader application of automated machine-learning-based quantitation of hepatic fat from clinical CT scans could lead to greater opportunities for genomic discovery and for identification of individuals with hepatic steatosis who may be candidates for clinical trials.

Additionally, we recognize that while CT imaging provides the opportunity to quantify hepatic fat through rapid and scalable imaging, there are some limitations. In particular, the presence of iron, copper, glycogen, fibrosis, edema, and postsurgical hardware may confound attenuation values and lead to errors in hepatic fat quantification.[49] CT studies in the PMBB were obtained in routine clinical care. Given that these CT scans were performed for clinical indications, it was important for us to exclude potential confounding comorbidities. To address this concern, we excluded individuals with International Classification of Diseases (ICD) 9/10 diagnosis codes indicating chronic hepatitis B or C as well as alcohol-related conditions or dependence, such as alcoholic liver disease, alcoholic hepatitis, alcoholic fibrosis and sclerosis of the liver, alcoholic cirrhosis of liver and/or ascites, alcoholic hepatic failure, coma, unspecified alcoholic liver disease, and alcohol dependence. After exclusions (<10% of all scans analyzed), the remaining sample size for exome-wide association studies was 9,594 for analyses. For PheWAS, any indications for which there were <20 cases based on Phecodes were excluded.

Nevertheless, there are several rare distinct causes of steatosis that were not excluded, including Wilson disease, glycogen storage disease, and hemochromatosis.[50] However, exclusion of these diagnoses relies upon accurate recording of ICD billing codes, which are often incomplete. In the PMBB data, we believe the contribution from these rare cases has minimal impact on our findings given that these cases represent a small fraction of all CT studies analyzed based on available ICD codes. There are also additional causes of altered hepatic attenuation that are not captured by billing codes such as radiation,[51] IV hyperalimentation,[52] and certain drugs such as corticosteroids, amiodarone, and methotrexate.[53] In addition, deep learning methods applied to CT scans label the whole liver, including vasculature and intrahepatic lesions if present, and thus could incorrectly estimate hepatic attenuation for use in downstream association analyses. However, by conducting replication studies in UKB using MRI-PDFF, a more accurate measure of hepatic fat compared with CT-derived quantitation,[54] we suggest that replicated signals are less likely to be confounded by these factors. Furthermore, the substantially larger number of CT scans obtained in hospitals compared with MRI scans allows for considerably increased power for conducting these types of IDP-based association studies in healthcare-based populations. Thus, we suggest that there is value in interrogating multiple imaging modalities for increased specificity.

In conclusion, we leveraged machine learning of clinical CT imaging in a medical biobank to automate the quantitation of hepatic fat and integrated this clinically important phenotype with whole-exome sequence data in the same individuals for genomic discovery. Our study not only extends existing knowledge regarding genetic variation associated with hepatic fat but also demonstrates the feasibility and value of aggregating rare predicted deleterious coding variants into gene burdens on an exome-wide scale for association with hepatic fat quantifications for the discovery of genes that may regulate intrahepatic fat and confer risk for or protect from NAFLD. Furthermore, our study provides an example of the value of IDPs extracted from clinical imaging in the context of a medical biobank, making it evident that there is significant utility in extracting IDPs from images collected in routine clinical imaging. We suggest that much larger experiments of this type that link IDPs derived from clinical imaging to genetic sequencing will lead to insights into the genetic regulation of a range of important phenotypes that are highly relevant to human health and disease.

Limitations of the study

Our study replicates well-established associations between hepatic steatosis and genetic variants in genes such as *PNPLA3*, *TM6SF2*, and *APOE*, and also identifies additional variants, including those in *FGD5*, *CITED2*, and *LMF2*. While we support our findings with clinical characterization of these additional variants and bioinformatic analyses of relevant publicly available datasets, a limitation of this study is the lack of functional studies investigating the precise mechanism underlying these associations. Future directions of this work include studying the impact of these variants on histological features of liver damage, their biological processes, and pathways leading to NAFLD pathogenesis, and the exact cell types by which the variants exert their function. Another limitation of this study is that CT scans performed in a medical biobank may have clinical indications, which could potentially bias genetic associations. We address this concern through manual review of clinical indications of CT scans for carriers of our additional variants of interest in *FGD5*, *CITED2*, and *LMF2*. A future direction of this work includes a comprehensive analysis of the clinical indications for all CT scans included in this study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Setting and study participants
- METHOD DETAILS
  - Clinical data collection
  - Image analysis hardware and image pre-filtering
  - Detection of non-contrast CT scans
  - Segmentation
  - Quantification of hepatic fat
  - Phenome-wide association study of hepatic fat with EHR diagnoses and traits
  - Whole exome sequencing, variant annotation, and selection for association testing
  - Exome-wide association studies of hepatic fat
  - Replication analyses in the UK Biobank (UKB)
  - Analysis of publicly available expression datasets
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - PheWAS analyses in PMBB
  - Exome-wide association studies in PMBB and replication studies in UKB
  - Association analyses with laboratory measurements

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2022.100855.

ACKNOWLEDGMENTS

We thank JoEllen Weaver, Stephanie DerOhannessian, and Marjorie Risman of the PMBB. We also thank the Regeneron Genetics Center for performing whole-exome sequencing on PMBB participants:

- RGC management and leadership team: Goncalo Abecasis, Aris Baras, Michael Cantor, Giovanni Coppola, Andrew Deubler, Aris Economides, Luca A. Lotta, John D. Overton, Jeffrey G. Reid, Alan Shuldiner, Katia Karalis, and Katherine Siminovitch.
- Sequencing and lab operations: Christina Beechert; Caitlin Forsythe, M.S.; Erin D. Fuller; Zhenhua Gu, M.S.; Michael Lattari; Alexander Lopez, M.S.; John D. Overton; Thomas D. Schleicher, M.S.; Maria Sotiropoulos Padilla, M.S.; Louis Widom; Sarah E. Wolf, M.S.; Manasi Pradhan, M.S.; Kia Manoochehri; and Ricardo H. Ulloa.
- Genome informatics: Xiaodong Bai, Suganthi Balasubramanian, Boris Boutkov, Gisu Eom, Lukas Habegger, Alicia Hawes, Shareef Khalid, Olga Krasheninina, Rouel Lanche, Adam J. Mansfield, Evan K. Maxwell, and Mona Nafde, Sean O'Keeffe, Max Orelus, Razvan Panea, Tommy Polanco, Ayesha Rasool, Jeffrey G. Reid, William Salerno, and Jeffrey C. Staples.
- Clinical informatics: Michael Cantor, Dadong Li, Deepika Sharma, and Nilanjana Banerjee.
- Translational and analytical genetics: Jonas Bovijn, Adam Locke, Niek Verweij, Mary Haas, George Hindy, Tanima De, Parsa Akbari, Olukayode Sosina, and Manuel A.R. Ferreira.
- Research program management: Marcus B. Jones, Jason Mighty, Michelle G. LeBlanc, and Lyndon J. Mitnaul.

The PMBB is funded by the Perelman School of Medicine at the University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. Research reported in this paper was supported by grants from the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under award number F30HG010442 and the Blavatnik Family Foundation (to J.P.); the Sarnoff Cardiovascular Research Foundation (M.M.); the NIH under IK2-CX001780 (to S.M.D.; this publication does not represent the views of the Department of Veterans Affairs or the United States Government); the National Heart, Lung, and Blood Institute (NHLBI) of the NIH under award number 1R01HL137984, and the Center for Precision Medicine and the Institute for Translational Medicine and Therapeutics (ITMAT) at the University of Pennsylvania (to W.W.).

AUTHOR CONTRIBUTIONS

J.P., M.T.M., W.R.W., and D.J.R. conceived and performed experiments, acquired data, developed methods, wrote the manuscript, and secured funding. A.M.L., D.T., C.V.S., B.X., J.E.M., T.C., and Y.B. performed experiments. R.L.J., R.G.C., A.V., and S.M.D. were involved with data acquisition. M.D.R. conceived experiments, wrote the manuscript, and provided expertise and feedback.

DECLARATION OF INTERESTS

S.M.D. receives research support from RenalytixAI, in-kind research support from Novo Nordisk, and personal consulting fees from Calico Labs, outside the scope of the current research. The authors declare no other competing interests.

Received: January 26, 2022
Revised: August 22, 2022
Accepted: November 17, 2022
Published: December 12, 2022

REFERENCES

1. Nassir, F., Rector, R.S., Hammoud, G.M., and Ibdah, J.A. (2015). Pathogenesis and prevention of hepatic steatosis. Gastroenterol. Hepatol. *11*, 167–175.
2. Byrne, C.D., and Targher, G. (2015). NAFLD: a multisystem disease. J. Hepatol. *62*, S47–S64. https://doi.org/10.1016/j.jhep.2014.12.012.
3. Vujkovic, M., Ramdas, S., Lorenz, K.M., Guo, X., Darlay, R., Cordell, H.J., He, J., Gindin, Y., Chung, C., Myers, R.P., et al. (2022). A multiancestry genome-wide association study of unexplained chronic ALT elevation as a proxy for nonalcoholic fatty liver disease with histological and radiological validation. Nat. Genet. *54*, 761–771. https://doi.org/10.1038/s41588-022-01078-z.
4. Jamialahmadi, O., Mancina, R.M., Ciociola, E., Tavaglione, F., Luukkonen, P.K., Baselli, G., Malvestiti, F., Thuillier, D., Raverdy, V., Männistö, V., et al. (2021). Exome-wide association study on alanine aminotransferase identifies sequence variants in the GPAM and APOE associated with fatty liver disease. Gastroenterology *160*, 1634–1646.e7. https://doi.org/10.1053/j.gastro.2020.12.023.
5. Eslam, M., and George, J. (2020). Genetic contributions to NAFLD: leveraging shared genetics to uncover systems biology. Nat. Rev. Gastroenterol. Hepatol. *17*, 40–52. https://doi.org/10.1038/s41575-019-0212-0.
6. Romeo, S., Kozlitina, J., Xing, C., Pertsemlidis, A., Cox, D., Pennacchio, L.A., Boerwinkle, E., Cohen, J.C., and Hobbs, H.H. (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet. *40*, 1461–1465. https://doi.org/10.1038/ng.257.
7. Speliotes, E.K., Yerges-Armstrong, L.M., Wu, J., Hernaez, R., Kim, L.J., Palmer, C.D., Gudnason, V., Eiriksdottir, G., Garcia, M.E., Launer, L.J.,

et al. (2011). Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet. *7*, e1001324. https://doi.org/10.1371/journal.pgen.1001324.

8. Kozlitina, J., Smagris, E., Stender, S., Nordestgaard, B.G., Zhou, H.H., Tybjærg-Hansen, A., Vogt, T.F., Hobbs, H.H., and Cohen, J.C. (2014). Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet. *46*, 352–356. https://doi.org/10.1038/ng.2901.

9. Park, S.L., Li, Y., Sheng, X., Hom, V., Xia, L., Zhao, K., Pooler, L., Setiawan, V.W., Lim, U., Monroe, K.R., et al. (2020). Genome-wide association study of liver fat: the multiethnic cohort adiposity phenotype study. Hepatol. Commun. *4*, 1112–1123. https://doi.org/10.1002/hep4.1533.

10. Parisinos, C.A., Wilman, H.R., Thomas, E.L., Kelly, M., Nicholls, R.C., McGonigle, J., Neubauer, S., Hingorani, A.D., Patel, R.S., Hemingway, H., et al. (2020). Genome-wide and Mendelian randomisation studies of liver MRI yield insights into the pathogenesis of steatohepatitis. J. Hepatol. *73*, 241–251. https://doi.org/10.1016/j.jhep.2020.03.032.

11. Liu, Y., Basty, N., Whitcher, B., Bell, J.D., Sorokin, E.P., van Bruggen, N., Thomas, E.L., and Cule, M. (2021). Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning. Elife *10*, e65554. https://doi.org/10.7554/eLife.65554.

12. Haas, M.E., Pirruccello, J.P., Friedman, S.N., Wang, M., Emdin, C.A., Ajmera, V.H., Simon, T.G., Homburger, J.R., Guo, X., Budoff, M., et al. (2021). Machine learning enables new insights into genetic contributions to liver fat accumulation. Cell Genom. *1*, 100066. https://doi.org/10.1016/j.xgen.2021.100066.

13. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118. https://doi.org/10.1038/nature21056.

14. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA *316*, 2402–2410. https://doi.org/10.1001/jama.2016.17216.

15. Park, J., Lucas, A.M., Zhang, X., Chaudhary, K., Cho, J.H., Nadkarni, G., Dobbyn, A., Chittoor, G., Josyula, N.S., Katz, N., et al. (2021). Exome-wide evaluation of rare coding variants using electronic health records identifies new gene-phenotype associations. Nat. Med. *27*, 66–72. https://doi.org/10.1038/s41591-020-1133-8.

16. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894. https://doi.org/10.1093/nar/gky1016.

17. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419. https://doi.org/10.1126/science.1260419.

18. Brancale, J., and Vilarinho, S. (2021). A single cell gene expression atlas of 28 human livers. J. Hepatol. *75*, 219–220. https://doi.org/10.1016/j.jhep.2021.03.005.

19. Farhan, M.A., Azad, A.K., Touret, N., and Murray, A.G. (2017). FGD5 regulates VEGF receptor-2 coupling to PI3 kinase and receptor recycling. Arterioscler. Thromb. Vasc. Biol. *37*, 2301–2310. https://doi.org/10.1161/ATVBAHA.117.309978.

20. Kurogane, Y., Miyata, M., Kubo, Y., Nagamatsu, Y., Kundu, R.K., Uemura, A., Ishida, T., Quertermous, T., Hirata, K.i., and Rikitake, Y. (2012). FGD5 mediates proangiogenic action of vascular endothelial growth factor in human vascular endothelial cells. Arterioscler. Thromb. Vasc. Biol. *32*, 988–996. https://doi.org/10.1161/ATVBAHA.111.244004.

21. Yuan, H., Zhang, H., Wu, X., Zhang, Z., Du, D., Zhou, W., Zhou, S., Brakebusch, C., and Chen, Z. (2009). Hepatocyte-specific deletion of Cdc42 results in delayed liver regeneration after partial hepatectomy in mice. Hepatology *49*, 240–249. https://doi.org/10.1002/hep.22610.

22. Dickinson, M.E., Flenniken, A.M., Ji, X., Teboul, L., Wong, M.D., White, J.K., Meehan, T.F., Weninger, W.J., Westerberg, H., Adissu, H., et al. (2016). High-throughput discovery of novel developmental phenotypes. Nature *537*, 508–514. https://doi.org/10.1038/nature19356.

23. Parviz, F., Matullo, C., Garrison, W.D., Savatski, L., Adamson, J.W., Ning, G., Kaestner, K.H., Rossi, J.M., Zaret, K.S., and Duncan, S.A. (2003). Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. Nat. Genet. *34*, 292–296. https://doi.org/10.1038/ng1175.

24. Hayhurst, G.P., Lee, Y.H., Lambert, G., Ward, J.M., and Gonzalez, F.J. (2001). Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. Mol. Cell Biol. *21*, 1393–1403. https://doi.org/10.1128/MCB.21.4.1393-1403.2001.

25. Huang, K.W., Reebye, V., Czysz, K., Ciriello, S., Dorman, S., Reccia, I., Lai, H.S., Peng, L., Kostomitsopoulos, N., Nicholls, J., et al. (2020). Liver activation of hepatocellular nuclear factor-4alpha by small activating RNA rescues dyslipidemia and improves metabolic profile. Mol. Ther. Nucleic Acids *19*, 361–370. https://doi.org/10.1016/j.omtn.2019.10.044.

26. Qu, X., Lam, E., Doughman, Y.Q., Chen, Y., Chou, Y.T., Lam, M., Turakhia, M., Dunwoodie, S.L., Watanabe, M., Xu, B., et al. (2007). Cited2, a coactivator of HNF4alpha, is essential for liver development. EMBO J. *26*, 4445–4456. https://doi.org/10.1038/sj.emboj.7601883.

27. Schmidt, T., Samaras, P., Frejno, M., Gessulat, S., Barnert, M., Kienegger, H., Krcmar, H., Schlegl, J., Ehrlich, H.C., Aiche, S., et al. (2018). ProteomicsDB. Nucleic Acids Res. *46*, D1271–D1281. https://doi.org/10.1093/nar/gkx1029.

28. Li, P., Oh, D.Y., Bandyopadhyay, G., Lagakos, W.S., Talukdar, S., Osborn, O., Johnson, A., Chung, H., Maris, M., Ofrecio, J.M., et al. (2015). LTB4 promotes insulin resistance in obese mice by acting on macrophages, hepatocytes and myocytes. Nat. Med. *21*, 239–247. https://doi.org/10.1038/nm.3800.

29. Weerasekera, L., Rudnicka, C., Sang, Q.X., Curran, J.E., Johnson, M.P., Moses, E.K., Göring, H.H.H., Blangero, J., Hricova, J., Schlaich, M., and Matthews, V.B. (2017). ADAM19: a novel target for metabolic syndrome in humans and mice. Mediators Inflamm. *2017*, 7281986. https://doi.org/10.1155/2017/7281986.

30. Chartrel, N., Picot, M., El Medhi, M., Arabo, A., Berrahmoune, H., Alexandre, D., Maucotel, J., Anouar, Y., and Prévost, G. (2016). The neuropeptide 26RFa (QRFP) and its role in the regulation of energy homeostasis: a mini-review. Front. Neurosci. *10*, 549. https://doi.org/10.3389/fnins.2016.00549.

31. Granata, R., Settanni, F., Trovato, L., Gallo, D., Gesmundo, I., Nano, R., Gallo, M.P., Bergandi, L., Volante, M., Alloatti, G., et al. (2014). RFamide peptides 43RFa and 26RFa both promote survival of pancreatic beta-cells and human pancreatic islets but exert opposite effects on insulin secretion. Diabetes *63*, 2380–2393. https://doi.org/10.2337/db13-1522.

32. Prévost, G., Arabo, A., Le Solliec, M.A., Bons, J., Picot, M., Maucotel, J., Berrahmoune, H., El Mehdi, M., Cherifi, S., Benani, A., et al. (2019). Neuropeptide 26RFa (QRFP) is a key regulator of glucose homeostasis and its activity is markedly altered in obese/hyperglycemic mice. Am. J. Physiol. Endocrinol. Metab. *317*, E147–E157. https://doi.org/10.1152/ajpendo.00540.2018.

33. El-Mehdi, M., Takhlidjt, S., Khiar, F., Prévost, G., do Rego, J.L., do Rego, J.C., Benani, A., Nedelec, E., Godefroy, D., Arabo, A., et al. (2020). Glucose homeostasis is impaired in mice deficient in the neuropeptide 26RFa (QRFP). BMJ Open Diabetes Res. Care *8*, e000942. https://doi.org/10.1136/bmjdrc-2019-000942.

34. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The missing diversity in human genetic studies. Cell *177*, 26–31. https://doi.org/10.1016/j.cell.2019.02.048.

35. Donati, B., Motta, B.M., Pingitore, P., Meroni, M., Pietrelli, A., Alisi, A., Petta, S., Xing, C., Dongiovanni, P., del Menico, B., et al. (2016).

The rs2294918 E434K variant modulates patatin-like phospholipase domain-containing 3 expression and liver damage. Hepatology *63*, 787–798. https://doi.org/10.1002/hep.28370.

36. BasuRay, S., Wang, Y., Smagris, E., Cohen, J.C., and Hobbs, H.H. (2019). Accumulation of PNPLA3 on lipid droplets is the basis of associated hepatic steatosis. Proc. Natl. Acad. Sci. USA *116*, 9521–9526. https://doi.org/10.1073/pnas.1901974116.
37. Ehrhardt, N., Doche, M.E., Chen, S., Mao, H.Z., Walsh, M.T., Bedoya, C., Guindi, M., Xiong, W., Ignatius Irudayam, J., Iqbal, J., et al. (2017). Hepatic Tm6sf2 overexpression affects cellular ApoB-trafficking, plasma lipid levels, hepatic steatosis and atherosclerosis. Hum. Mol. Genet. *26*, 2719–2731. https://doi.org/10.1093/hmg/ddx159.
38. Santoro, N., Zhang, C.K., Zhao, H., Pakstis, A.J., Kim, G., Kursawe, R., Dykas, D.J., Bale, A.E., Giannini, C., Pierpont, B., et al. (2012). Variant in the glucokinase regulatory protein (GCKR) gene is associated with fatty liver in obese children and adolescents. Hepatology *55*, 781–789. https://doi.org/10.1002/hep.24806.
39. Emdin, C.A., Haas, M.E., Khera, A.V., Aragam, K., Chaffin, M., Klarin, D., Hindy, G., Jiang, L., Wei, W.Q., Feng, Q., et al. (2020). A missense variant in mitochondrial amidoxime reducing component 1 gene and protection against liver disease. PLoS Genet. *16*, e1008629. https://doi.org/10.1371/journal.pgen.1008629.
40. Pirola, C.J., Flichman, D., Dopazo, H., Fernández Gianotti, T., San Martino, J., Rohr, C., Garaycoechea, M., Gazzi, C., Castaño, G.O., and Sookoian, S. (2018). A rare nonsense mutation in the glucokinase regulator gene is associated with a rapidly progressive clinical form of nonalcoholic steatohepatitis. Hepatol. Commun. *2*, 1030–1036. https://doi.org/10.1002/hep4.1235.
41. Kitamoto, T., Kitamoto, A., Yoneda, M., Hyogo, H., Ochi, H., Nakamura, T., Teranishi, H., Mizusawa, S., Ueno, T., Chayama, K., et al. (2013). Genome-wide scan revealed that polymorphisms in the PNPLA3, SAMM50, and PARVB genes are associated with development and progression of nonalcoholic fatty liver disease in Japan. Hum. Genet. *132*, 783–792. https://doi.org/10.1007/s00439-013-1294-3.
42. Kleinstein, S.E., Rein, M., Abdelmalek, M.F., Guy, C.D., Goldstein, D.B., Mae Diehl, A., and Moylan, C.A. (2018). Whole-exome sequencing study of extreme phenotypes of NAFLD. Hepatol. Commun. *2*, 1021–1029. https://doi.org/10.1002/hep4.1227.
43. Péterfy, M., Ben-Zeev, O., Mao, H.Z., Weissglas-Volkov, D., Aouizerat, B.E., Pullinger, C.R., Frost, P.H., Kane, J.P., Malloy, M.J., Reue, K., et al. (2007). Mutations in LMF1 cause combined lipase deficiency and severe hypertriglyceridemia. Nat. Genet. *39*, 1483–1487. https://doi.org/10.1038/ng.2007.24.
44. Lebeaupin, C., Vallée, D., Hazari, Y., Hetz, C., Chevet, E., and Bailly-Maitre, B. (2018). Endoplasmic reticulum stress signalling and the pathogenesis of non-alcoholic fatty liver disease. J. Hepatol. *69*, 927–947. https://doi.org/10.1016/j.jhep.2018.06.008.
45. Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N.E. (2017). Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. Am. J. Epidemiol. *186*, 1026–1034. https://doi.org/10.1093/aje/kwx246.
46. Karlsen, T.H., Sheron, N., Zelber-Sagi, S., Carrieri, P., Dusheiko, G., Bugianesi, E., Pryke, R., Hutchinson, S.J., Sangro, B., Martin, N.K., et al. (2022). The EASL-Lancet Liver Commission: protecting the next generation of Europeans against liver disease complications and premature mortality. Lancet *399*, 61–116. https://doi.org/10.1016/S0140-6736(21)01701-3.
47. Kanwal, F., Shubrook, J.H., Adams, L.A., Pfotenhauer, K., Wai-Sun Wong, V., Wright, E., Abdelmalek, M.F., Harrison, S.A., Loomba, R., Mantzoros, C.S., et al. (2021). Clinical care pathway for the risk stratification and management of patients with nonalcoholic fatty liver disease. Gastroenterology *161*, 1657–1669. https://doi.org/10.1053/j.gastro.2021.07.049.
48. Wright, A.P., Desai, A.P., Bajpai, S., King, L.Y., Sahani, D.V., and Corey, K.E. (2015). Gaps in recognition and evaluation of incidentally identified hepatic steatosis. Dig. Dis. Sci. *60*, 333–338. https://doi.org/10.1007/s10620-014-3346-5.
49. Reeder, S.B., and Sirlin, C.B. (2010). Quantification of liver fat with magnetic resonance imaging. Magn. Reson. Imaging Clin. N. Am. *18*, 337–357, ix. https://doi.org/10.1016/j.mric.2010.08.013.
50. Idilman, I.S., Ozdeniz, I., and Karcaaltincaba, M. (2016). Hepatic steatosis: etiology, patterns, and quantification. Semin. Ultrasound CT MR *37*, 501–510. https://doi.org/10.1053/j.sult.2016.08.003.
51. Kim, J., and Jung, Y. (2017). Radiation-induced liver disease: current understanding and future perspectives. Exp. Mol. Med. *49*, e359. https://doi.org/10.1038/emm.2017.85.
52. Buchman, A.L., Dubin, M.D., Moukarzel, A.A., Jenden, D.J., Roch, M., Rice, K.M., Gornbein, J., and Ament, M.E. (1995). Choline deficiency: a cause of hepatic steatosis during parenteral nutrition that can be reversed with intravenous choline supplementation. Hepatology *22*, 1399–1403.
53. Rabinowich, L., and Shibolet, O. (2015). Drug induced steatohepatitis: an uncommon culprit of a common disease. BioMed Res. Int. *2015*, 168905. https://doi.org/10.1155/2015/168905.
54. Reeder, S.B., Hu, H.H., and Sirlin, C.B. (2012). Proton density fat-fraction: a standardized MR-based biomarker of tissue fat concentration. J. Magn. Reson. Imaging. *36*, 1011–1014. https://doi.org/10.1002/jmri.23741.
55. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.
56. MacLean, M.T., Jehangir, Q., Vujkovic, M., Ko, Y.-A., Litt, H., Borthakur, A., Sagraiya, H., Rosen, M., Mankoff, D.A., Schnall, M.D., et al. (2020). Linking abdominal imaging traits to electronic health record phenotypes. Preprint at medRxiv. https://doi.org/10.1101/2020.09.08.20190330.
57. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1505.04597.
58. Ma, X., Holalkere, N.S., Kambadakone R, A., Mino-Kenudson, M., Hahn, P.F., and Sahani, D.V. (2009). Imaging-based quantification of hepatic fat: methods and clinical applications. Radiographics *29*, 1253–1277. https://doi.org/10.1148/rg.295085186.
59. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. *31*, 1102–1110. https://doi.org/10.1038/nbt.2749.
60. Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics *30*, 2375–2376. https://doi.org/10.1093/bioinformatics/btu197.
61. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164. https://doi.org/10.1093/nar/gkq603.
62. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.
63. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: an Ensemble method for predicting the pathogenicity of rare missense variants. Am. J. Hum. Genet. *99*, 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016.
64. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141, 456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

65. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting splicing from primary sequence with deep learning. Cell *176*, 535–548.e24. https://doi.org/10.1016/j.cell.2018.12.015.

66. Graffy, P.M., and Pickhardt, P.J. (2016). Quantification of hepatic and visceral fat by CT and MR imaging: relevance to the obesity epidemic, metabolic syndrome and NAFLD. Br. J. Radiol. *89*, 20151024. https://doi.org/10.1259/bjr.20151024.

67. Kramer, H., Pickhardt, P.J., Kliewer, M.A., Hernando, D., Chen, G.H., Zagzebski, J.A., and Reeder, S.B. (2017). Accuracy of liver fat quantification with advanced CT, MRI, and ultrasound techniques: prospective comparison with MR spectroscopy. AJR Am. J. Roentgenol. *208*, 92–100. https://doi.org/10.2214/AJR.16.16565.

68. Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

69. Govaere, O., Cockell, S., Tiniakos, D., Queen, R., Younes, R., Vacca, M., Alexander, L., Ravaioli, F., Palmer, J., Petta, S., et al. (2020). Transcriptomic profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis and fibrosis. Sci. Transl. Med. *12*, eaba4448. https://doi.org/10.1126/scitranslmed.aba4448.

70. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Summary statistics for exome-wide association analyses of hepatic fat derived from clinical CT scans in PMBB	Mendeley	https://doi.org/10.17632/6tdp8phxx9.1
Deep learning model weights and relevant code for performing inferencing on computed tomography data	Zenodo	https://doi.org/10.5281/zenodo.7259578
Exome sequencing variant-level summary data for the Penn Medicine BioBank	PMBB Genome Browser	https://pmbb.med.upenn.edu/allele-frequency/
Human reference genome NCBI build 38, GRCh38	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Software and algorithms		
Python 3.5	Python	https://www.python.org/downloads/release/python-350/
Tensorflow R package version 1.12.0	Tensorflow	https://www.tensorflow.org/
R 3.5	R	https://cran.r-project.org/
PheWAS R package version 0.99.5-4	Github	https://github.com/PheWAS/PheWAS
ANNOVAR version 2019Oct24	ANNOVAR	https://annovar.openbioinformatics.org/en/latest/
SpliceAI	Github	https://github.com/Illumina/SpliceAI
Limma 3.52.2	Bioconductor	https://bioconductor.org/packages/release/bioc/html/limma.html
edgeR 3.38.4	Bioconductor	https://bioconductor.org/packages/release/bioc/html/edgeR.html
Other		
RNA-sequencing data of human livers from histologically characterized NAFLD samples	NCBI Gene Expression Omnibus (GEO)	GSE135251

RESOURCE AVAILABILITY

Lead contact

- Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Daniel J Rader (rader@pennmedicine.upenn.edu).

Materials availability

- This study did not generate new unique reagents.

Data and code availability

- All summary statistics for significant genomic findings from the exome-wide association studies of hepatic fat in the Penn Medicine BioBank (PMBB) discovery cohort are fully detailed in the supplemental information. These summary statistics are annotated with information regarding genomic location, variant effect, amino acid change, rs ID, minor allele frequency in gnomAD, and the number of carriers in the PMBB discovery cohort. Additionally, complete summary statistics have been deposited in Mendeley, with variants annotated by genomic location according to Genome Reference Consortium Human Build 38 (GRCh38) and gene name. DOIs are listed in the key resources table. Additionally, up-to-date summary data for genetic variants captured via whole exome sequencing in PMBB can be accessed via the Penn Medicine Biobank Genome Browser (pmbb.med.upenn.edu/allele-frequency/). Individual-level data, including sequencing, EHR phenotype, and original CT images analyzed in this study, are not made publicly available due to research participant privacy concerns; however, requests from accredited researchers for access to individual-level data relevant to this manuscript can be made by contacting the lead contact.

- The deep learning model weights and relevant code for performing inferencing on computed tomography data have been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Setting and study participants

All individuals recruited for the Penn Medicine BioBank (PMBB) are patients of clinical practice sites of the University of Pennsylvania Health System. Appropriate consent was obtained from each participant regarding storage of biological specimens, genetic sequencing, access to all available EHR data, and ability to recontact. These analyses focused on the subset of PMBB participants (N = 10,283) who had both CT-derived hepatic fat quantitation and whole exome sequence data (WES) available (Table 1). This study was approved by the Institutional Review Board of the University of Pennsylvania and complied with the principles set out in the Declaration of Helsinki.

METHOD DETAILS

Clinical data collection

All International Classification of Diseases Ninth Revision (ICD-9) and Tenth Revision (ICD-10) diagnosis codes, clinical imaging, and laboratory measurements were extracted from the patients' EHR. All ICD diagnosis codes and outpatient laboratory measurements available up to July 2020 were extracted for PMBB participants. Abdominal and chest CT images available up to March 2019 were extracted for PMBB participants regardless of the availability of whole exome sequences (N = 14,249). All laboratory values measured in the outpatient setting were extracted for participants from the time of enrollment in PMBB until July 2020; all units were converted to their respective clinical Traditional Units. Minimum, median, and maximum measurements of each laboratory measurement were recorded per individual for association analyses. Minimum, median, and maximum values for hemoglobin A1C, alkaline phosphatase, ALT, AST, and triglycerides were log-transformed to normalize their distributions.

We also queried the EHR image-server based on Current Procedural Terminology (CPT) codes for chest (CPT 71250 (N = 28,025), 71,270 (N = 1,149)), abdomen (CPT 74150 (N = 2,420), 74,170 (N = 1,422)), and abdomen/pelvis (CPT 74176 (N = 10,086), 74,178 (N = 1,923)) CT scans for liver fat analysis. Chest studies were included as these routinely include one-third to one-half of the liver, and an even larger fraction of the spleen, which is sufficient for evaluation of liver fat. All training data was manually generated by a trained technician under the supervision of a board-certified abdominal radiologist using 3D Slicer software. More detailed explanations about the exclusion criteria applied as well as the number of scans processed in each part of the study are shown in Figure S1.

Image analysis hardware and image pre-filtering

All convolution neural networks (CNN) were implemented in Python 3.5 using the Tensorflow package (version 1.12.0) in the cloud (Amazon Web Services).[55] Training was conducted using an NVIDIA P100 graphical processing unit (GPU) and inferences used parallel processing across 8 NVIDIA K80 GPUs. For both the classification and segmentation networks, the inputs were 2D axial slices with size 256 × 256. The input slices were transformed with a window width of 150 and level of 30, meaning that they were scaled such that voxels with Hounsfield Units (HU) between −45 and 105 occupied the 8-bit range between 0 and 255. Image analysis was performed consistent as described below and consistent with previously reported techniques.[56]

Detection of non-contrast CT scans

The first network (CNN_1) identified intravenous (IV) contrast CT scans and removed them from the analysis pipeline. This network consisted of convolutional layers which flattened into fully connected layers modeled after the VGG-16 classification network[13] (Figure S1). The network outputs a probability between 0 and 1 indicating the likelihood that a slice contains IV contrast. Scans were considered to have contrast if the average per-slice probability was greater than or equal to 0.5. This network was trained on 800 scans, 400 with IV contrast, and 400 without. Additionally, half these scans were of the abdomen/pelvis and half were thoracic. 320 scans (50,654 slices) were randomly placed in the training group, and 80 scans (12,867 slices) in a validation group. Training was conducted with a batch size of 32 and terminated when the model converged after 10 epochs. To evaluate performance of the classification network, model sensitivity and specificity was calculated on an additional 400 randomly selected scans, with 200 being thoracic and 200 abdomen/pelvis scans (Figure S1).

Segmentation

Additional networks were trained to segment the liver (CNN_2) and spleen (CNN_3) from axial 2D slices modeled after a U-Net architecture.[57] This model is composed of symmetric paths joined by skip connections where localized feature information from the contracting path is combined with contextual information from the expanding path. The complete architecture is shown in Figure S1. The networks output a probability for each voxel indicating the probability that it belongs to the organ of interest. For liver segmentation, the network was trained on a total of 106 abdomen/pelvis scans with 81 scans (7,999 slices) randomly selected for training and

25 scans (2,436 slices) for validation. For spleen segmentation, the network was trained on a total of 158 scans with 127 scans (12,399 slices) randomly selected for training and 31 scans (2,865 slices) for validation. Training data was selected iteratively when the model underperformed on a scan. Training was conducted with a batch-size of 32 and terminated when the model converged after 135 epochs for the liver, and 108 for the spleen.

To evaluate segmentation performance, a testing set of 20 abdomen/pelvis CT scans was randomly selected from PMBB and both manual as well as automated segmentations for liver and spleen were produced. Percent overlap was calculated to measure agreement between manual and automatic segmentations. Additionally, 50 scans were selected at random and mean attenuation was measured in the liver and spleen by the manual placement of ROIs. Eight spherical (20 mm diameter) ROIs were placed in the liver with four in the left and four in the right lobe. Two spherical (15 mm diameter) ROIs were placed in the spleen. Care was taken to avoid placement near edges or in regions of vasculature or lesions. The mean HU was computed between ROIs for the liver and spleen and compared to that obtained from the automated approach.

Quantification of hepatic fat

Consistent with the standard radiologic approach, hepatic fat was quantitated in PMBB by subtracting the mean attenuation of all voxels contained within the liver from the mean attenuation of all voxels contained in the spleen (spleen HU – liver HU) to create a measure that is directly proportional to intrahepatic fat.[58] Minimum, median, and maximum measurements of hepatic fat were recorded per individual given the multiple independent CT scans available per patient.

Phenome-wide association study of hepatic fat with EHR diagnoses and traits

A phenome-wide association study (PheWAS) approach was used to determine the phenotypes associated with the quantitative trait of median hepatic fat in PMBB for the 10,283 unrelated individuals in PMBB with both exome sequences and quantitated hepatic fat available.[59] ICD-10 encounter diagnoses were mapped to ICD-9 via the Center for Medicare and Medicaid Services 2017 General Equivalency Mappings (https://www.cms.gov/Medicare/Coding/ICD10/2017-ICD-10-CM-and-GEMs.html) and manual curation. Phenotypes for each individual were then determined by mapping ICD-9 codes to distinct disease entities (*i.e.* Phecodes) using the R package ''PheWAS'' version 0.99.5-4.[60] Patients were determined to have a certain disease phenotype if they had the corresponding ICD diagnosis on 2 or more dates, while phenotypic controls consisted of individuals who never had the ICD code. Individuals with an ICD diagnosis on only one date as well as individuals under control exclusion criteria based on PheWAS phenotype mapping protocols were not considered in statistical analyses. Each Phecode was tested for association with quantitated hepatic fat using a logistic regression model adjusted for age, genetically determined sex, and principal components (PC1-10) of genetic ancestry. Our association analyses considered only disease phenotypes with at least 20 cases based on power calculations in a prior simulation study.[15] This led to the interrogation of 1396 total Phecodes, and we used a Bonferroni correction to adjust for multiple testing ($p = 0.05/1396 = 3.58E\text{-}05$). These analyses were performed separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis.

Whole exome sequencing, variant annotation, and selection for association testing

A subset of 43,731 individuals in the PMBB had undergone whole exome sequencing (WES). We extracted DNA from stored buffy coats and then mapped exome sequences as generated by the Regeneron Genetics Center (Tarrytown, NY) to GRCh38 as previously described.[15] Samples with low exome sequencing coverage, high missingness (*i.e.* greater than 5% of targeted bases), dissimilar reported and genetically determined sex, and genetic evidence of sample duplication were not included in this subset. For subsequent phenotypic association analyses, we removed samples with evidence of 1st and 2nd-degree relatedness, leading to a total of sample size of 41,759 for analysis.

Genetic variants were annotated using ANNOVAR (version 2019Oct24)[61] for information regarding variant effect as determined by the NCBI Reference Sequence (RefSeq) database,[62] Rare Exonic Variant Ensemble Learner (REVEL) scores for missense variants,[63] and allele frequencies reported by the Genome Aggregation (gnomAD) v2.[64] Predicted loss-of-function (pLOF) variants were defined as frameshift insertions or deletions, gain of stop codon, and disruption of canonical splice site dinucleotides. For splicing variants, we removed those with SpliceAI scores <0.2 for loss or gain of acceptor or donor site.[65] For single variant association tests in the PMBB discovery, all nonsynonymous coding variants and splicing variants with minor allele frequency (MAF) > 0.1% in Africans or non-Finnish Europeans in gnomAD were selected for association testing. For gene burden association tests, rare (MAF $\leq$0.1% in gnomAD) pLOF variants were aggregated per gene with or without rare missense variants with REVEL score $\geq$0.5.

Exome-wide association studies of hepatic fat

This study focused on a subset of 10,283 unrelated individuals in PMBB with both WES and quantitated hepatic fat available. For exome-wide association studies of hepatic fat, individuals with ICD9/10 diagnosis codes indicating chronic hepatitis B or C (B18.0-B18.2, 070.32, 070.21, 070.22, 070.23, 070.31, 070.33, 070.54) or alcohol-related conditions or dependence, such as alcoholic liver disease (571.0, K70.0), alcoholic hepatitis (571.1, K70.1), alcoholic fibrosis and sclerosis of the liver (571.2, K70.3), alcoholic cirrhosis of liver and/or ascites (571.2, K70.2), alcoholic hepatic failure, coma, and unspecified alcoholic liver disease (571.3, K70.4, K70.40, K70.41, K70.9), and alcohol dependence (303.0, 303.9, F10.229, F10.20), were excluded (N = 689), leading to total sample size of 9,594 for analyses.

Exome-wide association studies of hepatic fat were conducted in two stages, namely single variant discovery and gene burden discovery. For the discovery analyses, single variants and gene burdens with at least 10 total carriers with hepatic fat quantifications available were associated with hepatic fat using a linear regression model adjusted for age, genetically determined sex, and principal components (PC) of ancestry (PC1-5 in Africans, PC1-10 in Europeans). For targeted gene burden analyses of genes nominated by the single variant discovery, gene burdens with at least 5 total carriers with hepatic fat quantifications available were associated with hepatic fat. For all gene burdens, we used an additive genetic model to aggregate variants as previously described.[15] These analyses were performed separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis. Additionally, trans-ancestral cosmopolitan analyses were also performed, adjusted for age, genetically determined sex, and cosmopolitan PC1-10.

We conducted a PheWAS for the gene burden of pLOF variants in *LMF2*, where we focused on a subset of 162 Phecodes in the ''digestive'' group, leading to a Bonferroni-corrected significance threshold of p = 0.05/162 = 3.09E-04. The gene burden PheWAS analysis was performed separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis. PheWAS was also performed for replicated single variants in PMBB and UKB, namely the *FGD5* H600Y variant and the 6-bp deletion (p.S198_G199del) in *CITED2*. Each Phecode was tested for association with genetic variation using a logistic regression model adjusted for age, genetically determined sex, and principal components (PC1-10) of genetic ancestry.

Replication analyses in the UK Biobank (UKB)

Replication analyses were conducted in the UK Biobank (UKB) looking for consistent directions of effect by linking exomes to hepatic fat based on liver proton density fat fraction (PDFF) extracted from abdominal MRI scans, given the strong linear correlation between hepatic fat extracted from non-contrast CT scans and PDFF quantifications derived from MRI scans.[66,67] We focused on 9,071 individuals with both exome sequences (after removing samples with evidence of 1st and 2nd-degree relatedness, high missingness, and dissimilar reported and genetically determined sex) and liver PDFF. Individuals with ICD10 diagnosis codes indicating chronic hepatitis B or C or alcohol-related conditions or dependence were excluded (N = 22) using the same exclusion criteria as in the PMBB discovery analyses, leading to a total sample size of 9,049 for analyses. Single variants and gene burdens with at least 5 total carriers with hepatic fat quantifications available selected based on discovery in PMBB were associated with hepatic fat using a linear regression model adjusted for age, genetically determined sex, and PC1-10 of ancestry. Similarly, for targeted gene burden analyses of genes nominated by the single variant discovery in PMBB, gene burdens with at least 5 total carriers with hepatic fat quantifications available were associated with hepatic fat in UKB. These analyses were performed in individuals of European ancestry, accompanied by trans-ancestral cosmopolitan analyses. Liver PDFF values from UKB were log-transformed to normalize their distribution for regression analyses. For replication studies in the UKB, International Classification of Diseases Tenth Revision (ICD-10) diagnosis codes and liver PDFF values derived from abdominal MRI scans were downloaded. Access to the UKB data for this project was from application 32133.

Analysis of publicly available expression datasets

We interrogated RNA-sequencing data publicly available on the NCBI GEO platform (https://www.ncbi.nlm.nih.gov/geo/).[68] We assessed hepatic expression levels for genes of interest informed by our exome-wide association studies in human livers from a cohort of histologically characterized NAFLD samples (GSE135251).[69] Among the 206 NAFLD cases were 168 early NAFLD cases and 38 moderate NAFLD cases, which were each compared to 10 control cases. Gene counts were normalized using the trimmed mean of M values method and transformed using limma's voom methodology, consistent with the original work. These normalized and transformed counts were analyzed for differential expression using linear models per implementation by limma.[70] Comparisons of gene expression in early or moderate NAFLD versus control were conducted for case-control ratios of 1:1, 2:1, and 5:1, with 10 random samplings per case-control ratio. For each gene, we calculated $\log_2$ fold-change of normalized gene expression in NAFLD vs. control, unadjusted p value, and Benjamini-Hochberg false discovery rate adjusted p value, and calculated mean and standard deviation for each statistic given the multiple samplings per analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

PheWAS analyses in PMBB

Each Phecode was tested for association with quantitated hepatic fat using a logistic regression model adjusted for age, genetically determined sex, and principal components (PC1-10) of genetic ancestry. Our association analyses considered only disease phenotypes with at least 20 cases based on power calculations in a prior simulation study.[15] This led to the interrogation of 1396 total Phecodes, and we used a Bonferroni correction to adjust for multiple testing (p = 0.05/1396 = 3.58E-05). These analyses were performed separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis. All phenome-wide association analyses were completed using R version 3.5 (Vienna, Austria). Further statistical details of these analyses can be found in the Results, figure legends, and supplemental figures/tables.

Exome-wide association studies in PMBB and replication studies in UKB

For the discovery analyses in PMBB, single variants and gene burdens with at least 10 total carriers with hepatic fat quantifications available were associated with hepatic fat using a linear regression model adjusted for age, genetically determined sex, and principal components (PC) of ancestry (PC1-5 in Africans, PC1-10 in Europeans). For targeted gene burden analyses of genes nominated by the single variant discovery, gene burdens with at least 5 total carriers with hepatic fat quantifications available were associated with hepatic fat. For all gene burdens, we used an additive genetic model to aggregate variants as previously described.[15] These analyses were performed separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis. Additionally, trans-ancestral cosmopolitan analyses were also performed, adjusted for age, genetically determined sex, and cosmopolitan PC1-10.

For replication analyses in UKB, single variants and gene burdens with at least 5 total carriers with hepatic fat quantifications available selected based on discovery in PMBB were associated with hepatic fat using a linear regression model adjusted for age, genetically determined sex, and PC1-10 of ancestry. These analyses were performed in individuals of European ancestry, accompanied by trans-ancestral cosmopolitan analyses. Liver PDFF values from UKB were log-transformed to normalize their distribution for regression analyses. All exome-wide discovery and replication association analyses were completed using R version 3.5 (Vienna, Austria). Further statistical details of these analyses can be found in the Results, figure legends, and supplemental figures/tables.

Association analyses with laboratory measurements

To associate hepatic fat phenotypes or genotypes with serum laboratory measurements in PMBB, we used a linear regression model adjusted for age, genetically determined sex, and PCs of genetic ancestry (PC1-5 in Africans, PC1-10 in Europeans). These analyses were performed across all ancestries (cosmopolitan, PC1-10) and/or separately by African and European genetic ancestry and combined with inverse variance weighted meta-analysis. All statistical association analyses were completed using R version 3.5 (Vienna, Austria). Minimum, median, and maximum values for hemoglobin A1C, alkaline phosphatase, ALT, AST, and triglycerides were log-normalized for regression analyses.